Linguistic- and Acoustic-based Automatic Dementia Detection using Deep Learning Methods

Yilin Pan
Supervisor: Prof Heidi Christensen and Dr Daniel Blackburn
Department of Computer Science
University of Sheffield

This dissertation is submitted for the degree of
Doctor of Philosophy

July 2022

I would like to dedicate this thesis to my loving family...

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Yilin Pan
July 2022

Acknowledgements

Through this three-and-a-half-year PhD journey, I have experienced many challenges and a lot of happiness. This will be an unforgettable experience. First and foremost, I would like to express a deep appreciation to my first supervisor, Prof Heidi Christensen. All of my studies in this thesis and the published papers can not be completed without her support, encouragement and guidance. I am fortunate to have had the chance to be one of her PhD students. Also, I would like to thank Dr Daniel Blackburn, my second supervisor, for his helpful comments and suggestions towards my research using his medical background. I am deeply grateful to the European Union's H2020 Marie Sklodowska-Curie programme TAPAS (Training Network for PAthological Speech processing) for having funded my PhD. Without this foundation, I would not have been able to finish my studies.

Many thanks to my PhD panels members: Prof Guy Brown and Dr Michael Mangan. Their insightful comments and suggestions regarding my studies at each panel meeting are invaluable. I would like to thank my thesis examiners Dr Saturnino Luz and Prof Thomas Hain for their suggestions and corrections to the thesis.

Special thanks to Dr Bahman Mirheidari for his endless help and guidance on the research areas in this thesis and for his collaborations in publishing the papers. Thanks to Ronan O'Malley, Prof Markus Reuber, Dr Traci Walker, and Prof Annalena Venneri for their collaboration in publishing the papers and providing the medical dataset to work with. I am thankful to the collaborators in TAPAS (Dr Nicholas Cummins, Dr Zhao Ren, Srikanth Nallanthighal and Julian Fritsch, Dr Mathew Magimai Doss) for their friendship, guidance and help. I also wish to thank Dr Aki Härmä for his supervision during my internship at Philips.

Thanks are also due to my colleagues in the SPandH lab (Dr Mashael AlSaleh, Dr Rabab Algadhy, Dr Lubna Alhinti, Bader Matar Alotaibi, Jack Deadman, Gerardo Roa Dabike, Dr Hector Romero, Jisi Zhang, Mingjie Chen, Dr Yanpei Shi, Zehai Tu, Zhengjun Yue, Wanli Sun, Dalia Attas, Fatimah Alzahrani, Nathan Pevy, Meg Thomas, Samuel Hollands, Dr Feifei Xiong, Dr Ning Ma) for their kindness during my PhD. They have enriched my research life and provided me with a vibrant experience in the UK.

Last, I would particularly like to thank my parents for their support and love throughout my life.

Abstract

Dementia can affect a person's speech and language abilities, even in the early stages. Dementia is incurable, but early detection can enable treatment that can slow down and maintain mental function. Therefore, early diagnosis of dementia is of great importance. However, current dementia detection procedures in clinical practice are expensive, invasive, and sometimes inaccurate. In comparison, computational tools based on the automatic analysis of spoken language have the potential to be applied as a cheap, easy-to-use, and objective clinical assistance tool for dementia detection.

In recent years, several studies have shown promise in this area. However, most studies focus heavily on the machine learning aspects and, as a consequence, often lack sufficient incorporation of clinical knowledge. Many studies also concentrate on clinically less relevant tasks such as the distinction between Healthy Control (HC) and people with Alzheimer's Disease (AD) which is relatively easy and therefore less interesting both in terms of the machine learning and the clinical application.

The studies in this thesis concentrate on automatically identifying signs of neurodegenerative dementia in the early stages and distinguishing them from other clinical, diagnostic categories related to memory problems: (Functional Memory Disorder (FMD), Mild Cognitive Impairment (MCI), and HC). A key focus, when designing the proposed systems has been to better consider (and incorporate) currently used clinical knowledge and also to bear in mind how these machine-learning based systems could be translated for use in real clinical settings.

Firstly, a state-of-the-art end-to-end system is constructed for extracting linguistic information from automatically transcribed spontaneous speech. The system's architecture is based on hierarchical principles thereby mimicking those used in clinical practice where information at both word-, sentence- and paragraph-level is used when extracting information to be used for diagnosis. Secondly, hand-crafted features are designed that are based on clinical knowledge of the importance of pausing and rhythm. These are successfully joined with features extracted from the end-to-end system. Thirdly, different classification tasks are explored, each set up so as to represent the types of diagnostic decision-making that is relevant in clinical practice. Finally, experiments are

conducted to explore how to better deal with the known problem of confounding and overlapping symptoms on speech and language from age and cognitive decline. A multi-task system is constructed that takes age into account while predicting cognitive decline. The studies use the publicly available DementiaBank dataset as well as the Intelligent Virtual Agent (IVA) dataset, which has been collected by our collaborators at the Royal Hallamshire Hospital, UK. In conclusion, this thesis proposes multiple methods of using speech and language information for dementia detection with state-of-the-art deep learning technologies, confirming the automatic system's potential for dementia detection.

List of Acronyms and Abbreviations

ReLU Rectified Linear Unit

PLDA Probabilistic Linear Discriminant Analysis

LPCC Linear Prediction Cepstral Coefficient

LLD Low Level Descriptor

CV cross validation

WER word error rate

ND Neurodegenerative disorders

GRU Gated recurrent unit

Glove Global Vectors embedding matrix for Word Representation

HBANN hierarchical bidirectional attention neural network

HBRNN hierarchical bidirectional recurrent neural network

HNR Harmonic to noise ratio

DT decision tree

SVR support vector regression

RMSE Root Mean Squared Error

eGeMAPS Geneva minimalistic acoustic parameter set

BERT Bidirectional Encoder Representations from Transformers

CFR Cumulative Frequency Response

ADReSS Alzheimer's Dementia Recognition through Spontaneous Speech

ADReSSo Alzheimer's Dementia Recognition through Spontaneous Speech *Only*

AD Alzheimer's Disease

ASR Automatic Speech Recognition

BLSTM Bidirectional Long Short-Term Memory

CA Conversation Analysis

CNN Convolutional Neural Network

CT Computed Tomography

DNN Deep Neural Network

FMD Functional Memory Disorder

FTD Fronto-Temporal Dementia

GloVe Global Vector

GMM Gaussian Mixture Model

HC Healthy Control

MAE Mean Average Error

RMSE Root Mean Square Error

IVA Intelligent Virtual Agent

BLSTM Bidirectional LSTM

LR Logistic Regression

LSTM Long Short-Term Memory

MCI Mild Cognitive Impairment

MFCC Mel Frequency Cepstral Coefficient

MMSE Mini Mental Status Examination

MRI Magnetic Resonance Imaging

MTL Multi-task learning

GP General Practitioner

SNR Signal to Noise Ratio

ND Neurodegenerative Disorder

NLP Natural Language Processing

PCA Principle Component Analysis

PET Positron Emission Tomography

PLDA Probabilistic Linear Discriminant Analysis

PLP Perceptual Linear Prediction

POS Part of Speech

ReLU Rectified Linear Unit

RNN Recurrent Neural Network

SVM Support Vector Machine

MLU mean length of utterance

TTR Type-token ratio

NLP natural language processing

WER Word Error Rate

Contents

Contents xi

List of Figures xv

List of Tables xvii

1 Introduction 1
- 1.1 Motivation . . . 3
- 1.2 Research Questions . . . 4
- 1.3 Thesis Contributions . . . 7
  - 1.3.1 Thesis Overview . . . 7
  - 1.3.2 Publications . . . 11

2 Dementia 15
- 2.1 What is Dementia? . . . 17
  - 2.1.1 Different Causes of Dementia . . . 18
  - 2.1.2 Stages of Dementia . . . 19
- 2.2 Effect of Dementia on Speech and Language . . . 21
  - 2.2.1 Effect on Speech . . . 21
  - 2.2.2 Effect on Language . . . 23
- 2.3 Current Diagnostic Procedures . . . 24
  - 2.3.1 Elements of the Diagnostic Procedures . . . 25
  - 2.3.2 Cognitive Assessment Tests . . . 26
  - 2.3.3 Speech- and Language-based Tasks for Dementia Detection . . . 30
  - 2.3.4 Problems with Existing Diagnostic Procedures . . . 32
- 2.4 Summary . . . 32

3 Automatic Dementia Detection Methods 35
- 3.1 Automatic Linguistic-based Dementia Detection . . . 37
  - 3.1.1 Knowledge-based Linguistic Features . . . 37
  - 3.1.2 Natural Language Processing Technologies . . . 39
- 3.2 Automatic Acoustic-based Dementia Detection . . . 40
  - 3.2.1 Knowledge-based Acoustic Features . . . 41
  - 3.2.2 End-to-end System based Acoustic Feature Extraction . . . 42
- 3.3 Advantages of Speech- and Language-based Automatic Dementia Detection 49
- 3.4 Drawbacks of Existing Automatic Methods . . . 50

3.5 Summary . . . 51

4 Overview of Available Datasets **53**
4.1 Publicly Available Datasets . . . 55
4.1.1 The DementiaBank Dataset . . . 55
4.1.2 The ADReSS Dataset . . . 59
4.1.3 The ADReSSo Dataset . . . 61
4.2 The IVA Datasets . . . 64
4.2.1 The IVA_{33} Dataset . . . 66
4.2.2 The IVA_{3class} Dataset . . . 67
4.2.3 The IVA_{60} Dataset . . . 67
4.2.4 The $\text{IVA}_{age\&MMSE}$ Dataset . . . 69
4.3 Summary . . . 69

5 Linguistic Information-based Dementia Detection **73**
5.1 Introduction . . . 75
5.2 Reasearch Background . . . 76
5.3 Dementia Detection System . . . 78
5.3.1 Word Embedding . . . 79
5.3.2 Word-level Structure . . . 79
5.3.3 Sentence-level Structure . . . 80
5.4 Experimental Setup . . . 81
5.4.1 Datasets . . . 81
5.4.2 Baseline Systems . . . 83
5.4.3 Evaluation Settings . . . 84
5.4.4 Model Configuration . . . 85
5.5 Results and Analysis . . . 86
5.5.1 Experimental Results . . . 86
5.5.2 Result Analysis . . . 89
5.6 Summary . . . 95

6 End-to-end Feature Extractor for Speech-based Dementia Detection **99**
6.1 Introduction . . . 101
6.2 End-to-end Feature Extractor . . . 102
6.3 Experimental Setup . . . 105
6.3.1 Dataset . . . 105
6.3.2 Evaluation Settings . . . 106
6.3.3 Model Configuration . . . 107
6.3.4 Baseline Feature Sets . . . 108
6.4 Results . . . 108
6.4.1 Classification Results on the IVA_{3class} Dataset . . . 108
6.4.2 Analysis of SincNet Filters from the IVA_{3class} Dataset . . . 110
6.4.3 Classification Results on the DementiaBank Dataset . . . 115
6.4.4 Dataset Comparison . . . 115
6.5 Summary . . . 116

7 High-performing Acoustic Feature Extraction 119
7.1 Introduction . . . 121
7.2 Background . . . 121
7.3 Methodology for Acoustic Feature Design . . . 123
7.3.1 Data Analysis . . . 123
7.3.2 Feature Construction . . . 125
7.3.3 Feature Classification . . . 127
7.4 Experimental Setup . . . 129
7.4.1 Pre-processing of the Audio Recordings and Transcripts . . . 129
7.4.2 Classifier Configuration . . . 130
7.5 Results . . . 131
7.5.1 Acoustic-based Results . . . 131
7.5.2 Linguistic-based Result . . . 132
7.5.3 Combined Feature Results . . . 133
7.6 Summary . . . 133

8 Multi-class Classification for Dementia Detection 135
8.1 Introduction . . . 137
8.2 Background . . . 138
8.3 System Construction . . . 139
8.3.1 Using Low- and High-quality Speech Segments . . . 140
8.3.2 Adding Traditional Features . . . 141
8.4 Experimental Setup . . . 144
8.4.1 Evaluation Setting . . . 144
8.4.2 Model Configuration . . . 145
8.5 Results . . . 146
8.5.1 Exploring the best speech input to the system . . . 146
8.5.2 Results with the TR-1 Feature . . . 148
8.5.3 Results with the TR-2 Feature . . . 149
8.5.4 Results Comparison . . . 150
8.6 Summary . . . 151

9 Multi-task Estimation of Age and Cognitive Decline from Speech 153
9.1 Introduction . . . 155
9.2 Background . . . 155
9.3 Data Analysis . . . 157
9.4 Multi-task System Construction . . . 160
9.4.1 End-to-end System . . . 160
9.4.2 Pipeline System . . . 162
9.5 Experimental Setup . . . 162
9.5.1 Datasets . . . 163
9.5.2 Evaluation Setting . . . 164
9.6 Results . . . 165
9.6.1 Baseline Results . . . 165
9.6.2 End-to-end System based Result . . . 166

9.6.3 Pipeline System based Result 168
9.7 Summary 169

10 Conclusions and Further Work **171**
10.1 Conclusions 172
10.2 Limitations 177
10.2.1 Limited Publicly Available Data 177
10.2.2 Inconsistent Settings and Performance 178
10.3 Future Work 179
10.3.1 Simplify the Automatic System Structure 179
10.3.2 Exploring Longitudinal Applications 180
10.3.3 Utilising Transfer Learning Technologies 180
10.4 Concluding Remarks 181

References **183**

A Traditional Features used in Chapter 8 **227**

List of Figures

1.1 Organisation of the thesis. The research questions addressed by each chapter are indicated. . . . 8

2.1 The Cookie Theft picture from the Boston Diagnostic Aphasia Examination. 31

4.1 The distribution of the MMSE scores and age of speakers in the DementiaBank dataset. . . . 56
4.2 The distribution of the MMSE and age in the $\text{IVA}_{age\&MMSE}$ dataset. . . . 68

5.1 The structure of the proposed hierarchical attention based system (HBANN) for dementia detection with the transcripts as the input. . . . 78
5.2 The structure of the baseline system: bi-LSTM. . . . 83
5.3 The structure of the baseline system: HBRNN. . . . 84
5.4 The effect of stop words on the manual and automatic transcripts for dementia detection. . . . 90
5.5 An example of visualising the word-level and sentence-level attention weights. 92
5.6 The extracted attention weights for the words from the HC and AD groups. 93
5.7 The word frequencies for the HC and AD groups. . . . 94

6.1 The structure of the Sinc-CLA feature extractor. . . . 103
6.2 Visualisation of the learned filters for the three classification tasks. . . . 111
6.3 Cumulative frequency response of SincNet filters on the three classification tasks; bold lines are the average response for the 10-fold CV and thin lines are the response for every fold trained system. . . . 112
6.4 The representation averaged across frames of the first five SincNet filters output; only the recordings from the first fold training set are shown. . . . 114
6.5 Recording samples from the two datasets used in this chapter. . . . 117

7.1 A piece of speech segment from the DementiaBank dataset, together with the manual transcript, automatic transcript, estimated confidence scores of each word and the time alignment information; the confidence threshold is set equal to 0.95 for classifying the speech segments into low confidence and high confidence. . . . 123
7.2 Visualisation of the designed three-dimension rhythm-related acoustic features from 30 selected recordings from HC and AD respectively. . . . 126

7.3 The neural network based classifier designed for classifying the extracted acoustic features for dementia detection. . . . 127
7.4 The combined system for utilising the acoustic information and linguistic information jointly for dementia detection. . . . 128
7.5 The relationship between the F-score and word confidence threshold on the evaluation and test set. . . . 131

8.1 The structure of the designed twin-CCLA feature extractor for extracting the TR-1 feature. . . . 141
8.2 Comparison of the extracted attention vectors from the low- and high-quality speech segments of the HC, FMD, MCI and ND. . . . 142
8.3 The Distribution of the Euclidean distance between the vectors' output of the attention layer from the four-way scenario. . . . 143
8.4 The structure of the feature fusion system designed for extracting the TR-2 feature. . . . 143
8.5 The confusion matrices of the classification results from the twin-CCLA extracted TR-1 features classified by the LR classifier. . . . 148
8.6 The confusion matrices of the classification results using TR-2 features classified by the LR classifier. . . . 149
8.7 The F-score for two-, three- and four-way systems using TR-1 and TR-2 features on the IVA_{60} data. . . . 151

9.1 The correlation between age and acoustic features (top left: F0_median, top right: speaking duration, bottom left: number of pauses, bottom right: number of syllables) with different cognitive status. . . . 159
9.2 The Euclidean distance between the anchor x-vectors and x-vectors extracted from people with different MMSE values. . . . 160
9.3 The structure of the multi-task Sinc-CLA system for the age and MMSE estimation. . . . 161
9.4 The structure of the multi-task pipeline system for the age and MMSE estimation. . . . 162
9.5 The learned normalised Cumulative Frequency Response from the three tasks. . . . 167

List of Tables

2.1 Summary of neuropsychological assessment tools for detecting cognitive decline. . . . 27

3.1 Acoustic-based speaker recognition development history. . . . 47
3.2 Speech emotion recognition development history. . . . 48

4.1 The participants and recording information of the DementiaBank dataset; #Rec is used to represent "the number of recordings". . . . 55
4.2 Previous research on the DementiaBank dataset using acoustic or linguistic information. . . . 58
4.3 The participant and recording information of the ADReSS dataset. . . . 60
4.4 The participant and recording information of the ADReSSo data to be used for the binary classification task and MMSE estimation tasks. . . . 62
4.5 The participant information and recording information of the ADReSSo data to be used for the disease progression detection task. . . . 63
4.6 The participant and recording information of the IVA dataset. M:F:U represents the number of speakers for male, female and un-known; #Rec represents the number of recordings; Rec. Dur. represents the recording duration. . . . 65
4.7 The recording information for the IVA dataset collected in different years. . 65
4.8 The information for the IVA_{3class} Dataset. . . . 67
4.9 The information for the IVA_{60} Dataset. . . . 68

5.1 Information about the DementiaBank dataset and the IVA_{33} dataset. . . . 82
5.2 The classification results of the HBANN system, bi-LSTM system and HBRNN system on the manual transcripts of the DementiaBank dataset. . 87
5.3 The detection results of the HBANN system and the baseline systems on the automatic transcripts of the DementiaBank dataset and the IVA_{33} dataset. 88
5.4 The results achieved using different word embedding initialisation and training methods on the manual and automatic transcripts. . . . 91

6.1 The speaker information, recording information and audio information for the DementiaBank Dataset. . . . 106
6.2 The F-score (%) for chunk-level classification on the IVA_{3class} dataset. . . . 109
6.3 The F-score (%) for the recording-level classification on the IVA_{3class} dataset. 109

6.4 The binary classification F-score (%) result for the recording-level classification on the DementiaBank dataset. 116

7.1 The average and variance of word duration, pause duration, number of words in the transcript and word confidence scores calculated for the HC and AD groups in the DementiaBank dataset. 125
7.2 The F-score (%) of the Linguistic-based system on the DementiaBank dataset. 133

8.1 The three scenarios designed for classifying the audio recordings from the HC, FMD, MCI and ND. 137
8.2 The parameter analysis for the recordings and automatic transcripts of the HC, MCI, FMD and ND groups in the IVA_{60} dataset. 139
8.3 Performance of the CCLA feature extractors designed with different types of waveform input on the four-way classification task. *full waveform*: the audio segments without pre-processing; *speech only*: low- and high-quality speech segments; *high quality*: the high-quality speech segments. 147
8.4 Performance of the twin-CCLA feature extractors designed with different types of waveform input on the four-way classification task. 147
8.5 The LR classification results with the TR-1 feature extracted by the Twin-CCLA system with low- and high-quality speech segments as the inputs. . 148
8.6 The LR classification results with the TR-2 feature extracted by the twin-CCLA system with the CA feature, WV-PCA feature, and low- and high-quality speech segments as the inputs. 149

9.1 The information for the datasets used for data analysis; CD: cognitive decline. 158
9.2 Detailed information for datasets used for the multi-task learning system training. 164
9.3 The results from the SVM based regression. 165
9.4 Results from the single-task and multi-task Sinc-CLA network. 166
9.5 The results with speaker embedding features on single-task/multi-task pipeline system estimation. 168
9.6 The regression results (RMSE) with speaker embedding features on single-task/multi-task pipeline system estimation. 169

A.1 *The traditional features used as the extra input of the designed twin-CCLA system for extracting TR-2 feature.* 228

Chapter 1

Introduction

Contents

1.1 Motivation **3**

1.2 Research Questions **4**

1.3 Thesis Contributions **7**

1.3.1 Thesis Overview 7

1.3.2 Publications 11

With an ageing society, the number of people with dementia is increasing rapidly all around the world. An estimated 55 million people worldwide are living with dementia in 2020, and this is predicted to almost double every 20 years, soaring to 139 million by 2050 [Alzheimer's Disease International, 2022]. The term *Dementia* is an umbrella term used to represent a set of symptoms arising from a range of progressive diseases. The term *Neurodegenerative Disorder (ND)* refers to slow progressive loss of neurons in the central nervous system that can lead to a recession in specific brain functions, which is irreversible. Dementia can be caused by different kinds of NDs, like Alzheimer's Disease (AD), Vascular Dementia and Parkinson's Disease. The most common cause of dementia is AD.

Before being diagnosed with dementia, people with early signs of cognitive decline often get diagnosed with Mild Cognitive Impairment (MCI). People living with MCI exhibit symptoms worse than those expected from normal ageing but not severe enough to be diagnosed as dementia [Elsey *et al.*, 2015]. About 10% to 15% of people living with MCI *convert* into living with AD per year, and in total 50% of people living MCI eventually get diagnosed with AD [Petersen *et al.*, 1999]. Dementia is incurable currently, but MCI is sometimes reversible. Early detection enables early treatment that can slow down and maintain mental function. As a result, early-stage diagnosis of dementia is meaningful and necessary.

To be diagnosed, a person with memory problems will typically first go to their General Practitioner (GP) and later be referred for more in-depth assessment at memory clinics. The diagnostic procedures include medical history taking, family interview, physical examination, cognitive assessment, laboratory testing and structural imaging. However, current manual assessment tests at the GPs are not consistently accurate, and too many people are referred to secondary care for further tests, which typically means long waiting times and increased anxiety, and often turns out to be unnecessary if they have memory problems not related to NDs. Also, frequently visiting a clinic can increase the physical and economic burden on people living with dementia and their families. Therefore, for convenience, an automatic dementia detection tool is expected to increase diagnostic accuracy and reduce the unnecessary waiting time before being sent to secondary care.

Even though memory impairment is the main symptom of dementia, language and speech are also affected, and the changes can often be seen many years before diagnosis [Berisha *et al.*, 2015; Pakhomov *et al.*, 2011; Ross *et al.*, 1990; Snowdon, 2003]. Specifically, for those people living at the early stage of dementia, although their speech remains largely informative, most of them experience some decline in their speaking ability, like having word-finding difficulties [Bird *et al.*, 2000], having their syntax/semantics being impoverished/simplified [Smith *et al.*, 1989] or exhibiting a more unstable fundamental frequency [Horii, 1979]. To diagnose, manual pen-and-paper style assessment tools are used by clinicians as a way to detect any degradation in a person's speech and language. However, as mentioned above, the accuracy (sensitivity and specificity) of such assessment tools is not satisfactory, and the further diagnostic process is time-consuming and costly. At the same time, consumer devices with good speech-recording abilities are becoming increasingly pervasive and affordable. Therefore, the investigation of automatic dementia detection methods using speech and language are of interest, and these methods are hoped to one day be part of the assessment in clinics and in people's homes.

1.1 Motivation

As mentioned above, studies have shown that acoustic and linguistic information embedded in a person's speech can be affected at the early stages of dementia [Pasquier, 1999]. Recent automatic approaches for detecting people living with dementia have shown a lot of promise [Fraser *et al.*, 2016; Khodabakhsh *et al.*, 2015; Mirheidari *et al.*, 2019b; Mueller *et al.*, 2018; Orimaye *et al.*, 2017; Warnita *et al.*, 2018; Weiner *et al.*, 2018]. In parallel, mainstream speech technology has seen huge benefits from developing speech and natural language processing based on deep learning technologies. This encourages us to explore how these state-of-the-art technologies may be applied to medical aids and help in clinical diagnosis. Therefore, this project aims to investigate the design and evaluation of automatic dementia detection systems using deep learning technologies for modeling both the acoustic and linguistic information embedded in a person's speech and language.

While constructing automatic systems for dementia detection, any current medical

knowledge used in the dementia diagnosis process should be considered carefully and could prove instructive for promising results [Covington *et al.*, 2006; Fraser *et al.*, 2016; Jarrold *et al.*, 2014; Khodabakhsh *et al.*, 2015; Mirheidari *et al.*, 2017; Orimaye *et al.*, 2017; Rentoumi *et al.*, 2017]. However, often research studies do not strive to combine such medical knowledge with the deep learning technologies. This observation encourages us to aim to introduce more medical knowledge while designing the systems or extracting the features, which are critical to the system's performance.

In current clinical practice, people who go to their GP with worries about their memory may live with Functional Memory Disorder (FMD), MCI or ND. Though they share similar symptoms, the treatments are different. However, current research is mostly based on the binary classification between Healthy Control (HC) and AD/ND. For clinical practice, evaluation frameworks and classification tasks should be designed based on the most relevant medical experience with a view to eventually apply the research in real-world settings. This thesis will address this limitation.

A final consideration, when taking into account the challenges of how research in this area might eventually be deployed in real systems, is to develop approaches that can deal with real-life challenges. One of these is that when detecting cognitive decline over time (using longitudinal data) both age and cognitive decline can result in acoustic changes. These are two independent processes that are highly correlated and exhibit overlapping systems and therefore should be considered jointly for better performance.

1.2 Research Questions

Based on the research context and clinical need as detailed above, the overall, high-level research aims of this research is to improve the detection and monitoring of cognitive impairment using automatic speech and language-based processing. Below, the more detailed research questions are laid out.

As described above, some recent work has focused on automatic speech and language-based dementia detection by analysing a person's speech and language. The speech-based analysis is normally based on the audio recordings [Beltrami *et al.*, 2018; Hoffmann *et al.*,

2010; Horley *et al.*, 2010; König *et al.*, 2015; Martínez-Sánchez *et al.*, 2012; Meilan *et al.*, 2018; Meilán *et al.*, 2020, 2014], whereas the language-based analysis is mostly carried out on either the manual or automatic transcripts generated from the audio recordings [Campbell *et al.*, 2020; Jarrold *et al.*, 2014; Khodabakhsh *et al.*, 2015; Mirheidari, 2018; Orimaye *et al.*, 2017; Rentoumi *et al.*, 2017; Roark *et al.*, 2007, 2011; Toledo *et al.*, 2018; Ujiro *et al.*, 2018; Vincze *et al.*, 2016; Yancheva & Rudzicz, 2016]. In previous research, the proposed dementia detection systems are mostly pipeline systems composed of front-end features and back-end classifiers. The performance of these systems is highly dependent on the quality of the extracted features as the back-end classifiers are mostly those similarly linear classifiers, such as Support Vector Machine (SVM), Logistic Regression (LR), and decision tree (DT). However, the front-end features are designed according to the classification tasks and different datasets.

The design and selection of features used for dementia detection are often time-consuming and highly dependent on the specific task and dataset. The most widespread features used for dementia detection are the commonly used feature sets selected depending on the specific task and dataset. This makes it hard to generalise between datasets and reuse the findings. In addition, there are very few publicly available datasets for investigating dementia detection studies, and most research is carried out on in-house, self-collected datasets, which introduces a considerable variation in accents, background noise and the collecting device. The variation in the quality and content of the datasets means there has not been much of a consensus concerning which features should be extracted. In other speech- and language-related research, deep learning technologies have recently been demonstrated to be efficient for learning data-driven features for specific dataset and classification tasks [Huang & Narayanan, 2016; Mirsamadi *et al.*, 2017; Rohdin *et al.*, 2018, 2020; Trigeorgis *et al.*, 2016; Xie *et al.*, 2019]. In addition, deep learning technologies have been shown to be powerful for modelling the information embedded in the speech [Fayek *et al.*, 2015; Huang & Narayanan, 2016; Mirsamadi *et al.*, 2017; Ravanelli & Bengio, 2018b; Snyder *et al.*, 2017, 2018; Stuhlsatz *et al.*, 2011; Xie *et al.*, 2019]. However, limited research has been proposed for speech- and language-based dementia detection due to data scarcity. In this thesis, the first research question is: **how can**

state-of-the-art deep neural networks be applied for speech- and language-based dementia detection? (RQ1)

When constructing an automatic dementia detection system, any medical expert knowledge that would usually be applied for the clinical diagnosis of people living with dementia could be instructive and is worth considering. However, a challenging part is how to model the linguistic and acoustic information embedded in the speech automatically. Early research for dementia detection was based on using standard machine learning techniques. Much recent research, based on making use of clinician's expert knowledge, has demonstrated promising results [Covington *et al.*, 2006; Fraser *et al.*, 2016; Jarrold *et al.*, 2014; Khodabakhsh *et al.*, 2015; Orimaye *et al.*, 2017; Rentoumi *et al.*, 2017; Rosenberg & Abbeduto, 1987]. However, modelling the medical knowledge into mathematical representation is challenging. For example, the collected data for dementia detection is mostly audio recordings. While diagnosing, the clinicians need to understand the speech and analyse the linguistic information embedded in the speech or analyse the manual transcripts generated from the audio recordings. The information embedded in the transcripts is hierarchical as the decline exists at both the word and sentence levels. While diagnosing, clinicians weigh the information and deficits exhibited in the words and the sentences carefully. However, no previous research has considered the hierarchical structure and weighted the words accordingly when constructing an automatic system for linguistic information extraction. For the acoustic part, the unclear pronunciation and long pauses in the speech are closely related to the symptoms and are used for diagnosis. Modelling this more explicitly could be beneficial. Therefore, the second research question in this thesis is: **how can the known clinical dementia detection knowledge help in constructing an automatic dementia detection systems and extracting useful features? (RQ2)**

The studies in this thesis aim to explore the potential of using an automatic system as clinical assistance for dementia detection. In general, people who go to their GP with worries about their memory often live with FMD, rather than MCI or ND. Though they share similar symptoms, people living with MCI or ND need referrals to secondary care, whereas people living with FMD do not need to. However, GPs lack the expert knowledge

to diagnose this and rule out dementia. They therefore refer these patients without distinction to secondary care resulting in expensive tests. Currently, for automatic dementia detection, most research has happened without taking into account the eventual use case in clinical practice as mentioned above. Specifically, the current research is mainly based on the binary classification for distinguishing people living with or without dementia. Under this situation, the third research question is: **how to design a framework for more clinically relevant diagnostic scenarios? (RQ3)**

The subsystems which make up the human speech production system undergo progressive physiological change affected by the decreasing rate and strength of muscle contraction [Kelly & Harte, 2011], resulting in acoustic changes. When dealing with longitudinally collected data, the changes in speech can result from both age and cognitive decline, which are two independent processes that are highly correlated. The age information has been used for improving the cognitive decline estimation [Fu *et al.*, 2020; Yancheva *et al.*, 2015], but without qualitative analysis of the relationship between the two factors before designing the system. In this thesis, the last research question is: **age and cognitive decline are confounding factors; how can the age information be used to improve the performance of the dementia detection system? (RQ4)**

1.3 Thesis Contributions

1.3.1 Thesis Overview

1. **A hierarchical attention based end-to-end system is proposed for linguistic-based information modelling:** A picture description task is used broadly for the detection of cognitive decline, like AD. When describing the provided picture, people with dementia show signs of cognitive decline at both the word and sentence levels. The clinical diagnosis also takes the hierarchical and sequential linguistic ability decline into consideration. Based on deep learning technologies, in our study, a hierarchical system is proposed that encompasses both the hierarchical and sequential structure of the description using the time sequence network and detects its informative units using the attention mechanism. The words and sentences are weighted with the attention mech-

anism at both the word and sentence levels. The designed system was the first to apply this hierarchical, medically-inspired architecture and it achieved a superior, state-of-the-art performance on the DementiaBank dataset [Becker *et al.*, 1994] when the study was published in the paper Pan *et al.* [2019]. The proposed system is also applied on the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge data, and the results are summarised in the paper Cummins *et al.* [2020] (see Chapter 5).

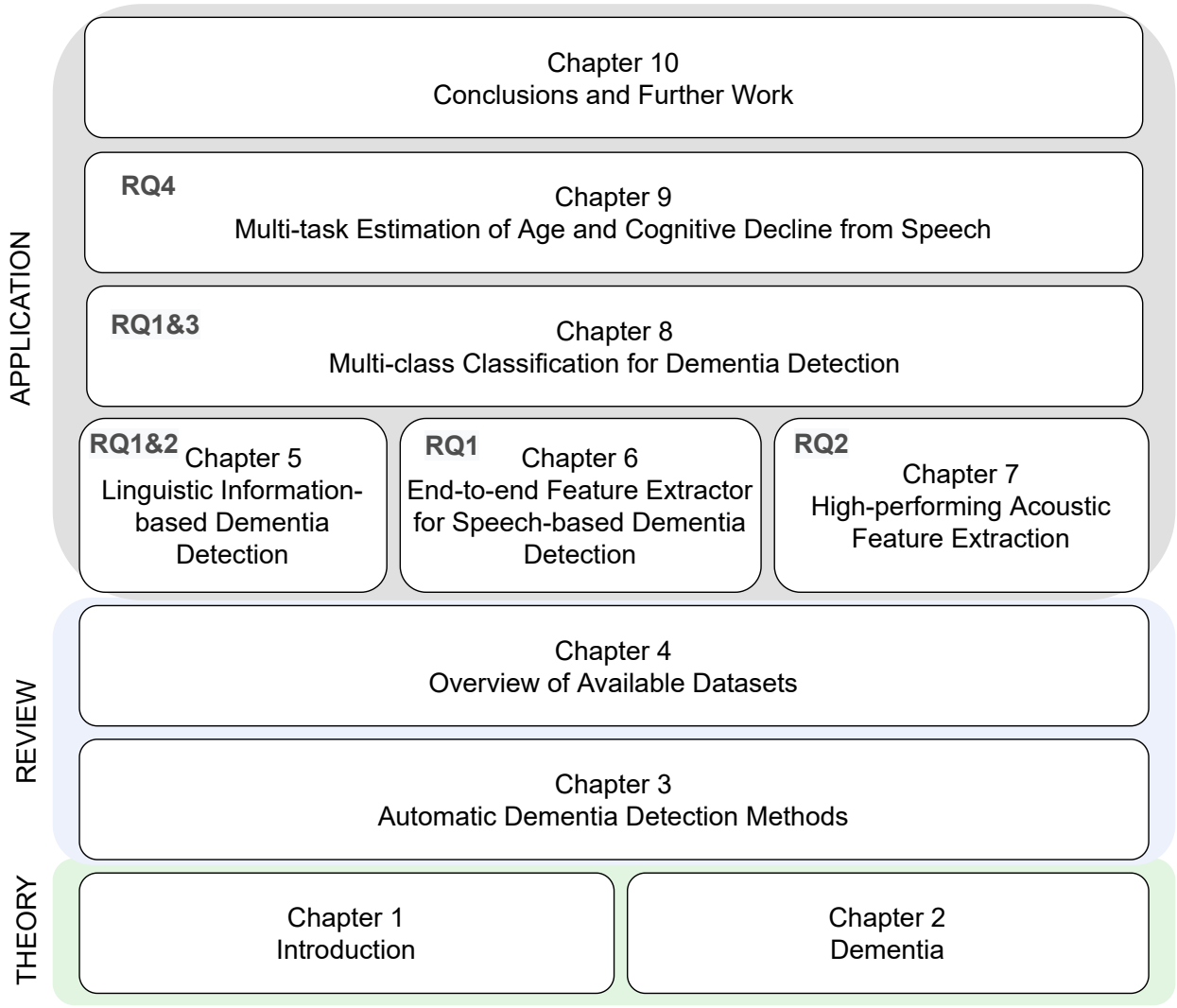


Figure 1.1: Organisation of the thesis. The research questions addressed by each chapter are indicated.

2. **Constructing an end-to-end system for extracting data-driven features from**

the raw waveform directly: To design and select features for different classification tasks, an end-to-end feature extractor is designed for extracting features directly from the raw waveform. Using a SincNet [Ravanelli & Bengio, 2018b] as the first layer of the designed feature extractor allows for some interpretable analysis of the learned features. This was the first system using an end-to-end system for acoustic-based dementia detection. Compared with the popular and commonly used feature sets, the trained features achieve a superior classification performance. The work is summarised in paper Pan *et al.* [2020b] (see Chapter 6).

3. **Using an Automatic Speech Recognition (ASR) system to help guide the extraction of high-performing acoustic features:** Unclear pronunciations and the presence of long and frequent pauses in a person's speech are all symptoms that currently feed into clinical diagnosis. To utilise this medical knowledge, an ASR system is designed to provide automatic transcription of the acoustic recordings and timing information for newly designing *rhythm-related features*. Novel *rhythm-related features* are extracted by using confidence scores and time alignment information estimated by the ASR system. In addition, improved *high-performing* acoustic features are extracted by categorising the audio recordings into speech segments with either *high* or *low* quality audio. This is the first time that ASR confidence scores have been used in this way for dementia detection. The results were published in the paper Pan *et al.* [2020a] (see Chapter 7).

4. **An end-to-end evaluation framework is designed for a broader, more clinically relevant, set of diagnostic classes:** Previous work has mostly concentrated on distinguishing AD from HCs. However, the distinction between MCI, ND and FMD and HC is much more clinically relevant. It is also more difficult due to the similar symptoms between people living with memory decline. Rather than only doing the four-class classification task, in terms of real-world clinical practice, regarding HC and FMD jointly as one class, and MCI and ND jointly as another class can be considered more appropriate. As a first, this clinical practice is taken into consideration when designing the systems and their evaluation in Chapter 8. The system (twin system

exploring the mismatch between the segments with high and low confidence scores estimated by the ASR system) follows on from work in Chapters 6 and 7[1].

5. **Making use of age information to improve the cognitive decline estimation accuracy:** The changes observed in a person's speech and language caused by ageing and cognitive decline arise from two independent processes but are highly correlated. The combined use of age and cognitive decline information is explored for the joint prediction of age and Mini Mental Status Examination (MMSE), which is a commonly used score for the assessment of cognitive decline. The results show that both age and MMSE prediction is improved by applying multi-task learning. The results are summarised in paper Pan *et al.* [2021b] (see Chapter 9).

A diagram of the organisation of this thesis is shown in Figure 1.1 with a division of chapters into 'theory', 'review' and 'application'. The remainder of the thesis is organised as follows:

Chapter 2 summarises the theory regarding dementia, different causes of dementia and different stages of dementia. Then, the impact of dementia on speech and language are presented, followed by the introduction of current diagnostic procedures for identifying people living with dementia.

Chapter 3 reviews the mainstream methods for automatic detection based on speech and language. Then, the advantages and drawbacks of the existing automatic dementia detection methods are summarised.

Chapter 4 introduces the available dementia-related datasets, including the three publicly available datasets: the DementiaBank dataset; the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) dataset [Luz *et al.*, 2020a]; and the Alzheimer's Dementia Recognition through Spontaneous Speech *Only* (ADReSSo) dataset [Luz *et al.*, 2021]. Then, the dataset collected by the Royal Hallamshire Hospital (Sheffield, UK) named as the Intelligent Virtual Agent (IVA) dataset is introduced, together with its several subsets that were defined throughout the duration of the data as

[1]The results has been written up into a journal paper and will likely be submitted to the *PLOS ONE* journal.

more data was collected. These are used in different experimental chapters as specified below.

Chapter 5 presents a hierarchical attention based end-to-end system for linguistic-based information modelling on the manual and automatic transcripts output by the ASR system. The designed system is tested with the DementiaBank dataset.

Chapter 6 presents a data-driven acoustic feature extractor for improving the performance of ND and MCI detection. The interpretability of the system is also considered when designing the feature extractor. The study is evaluated on both an IVA subset and the DementiaBank dataset.

Chapter 7 presents a proposed method that uses an ASR system for modelling symptoms of unclear pronunciation, and high-frequency pauses that exist in the speech of people living with AD. The study is evaluated on the DementiaBank dataset.

Chapter 8 presents an end-to-end feature extraction system based on the methods proposed in Chapter 6 and Chapter 7. The proposed system is evaluated with the classification tasks designed in accordance with clinical practice. The study is evaluated on an IVA subset.

Chapter 9 presents two multi-task systems designed to explore the intersection of age and cognitive decline estimation. The study is evaluated on an IVA subset and the ADReSS dataset.

Chapter 10 contains the conclusions, limitations and the further work of this thesis.

1.3.2 Publications

The publications refer to the works presented in this thesis are:

Pan, Y., Mirheidari, B., Reuber, M., Venneri, A., Blackburn, D., and Christensen, H. (2019). Automatic Hierarchical Attention Neural Network for Detecting AD. In Proceedings of Interspeech 2019 (pp. 4105-4109) [Pan *et al.*, 2019] (Chapter 5).

In this paper, Yilin Pan worked out the system structure, wrote the paper and did the experimental analysis. Dr Daniel Blackburn gave the knowledge support from the clinical perspective. Prof Markus Reuber and Prof Annalena Venneri contributed to

the collection and annotation of the dataset used in the paper. Dr Bahman Mirheidari and Prof Heidi Christensen discussed the related work with me in the weekly meeting.

Pan, Y., Mirheidari, B., Reuber, M., Venneri, A., Blackburn, D., and Christensen, H. (2020). Improving detection of Alzheimer's Disease using automatic speech recognition to identify high-quality segments for more robust feature extraction. In Proceedings of Interspeech 2020, 4961-4965 [Pan *et al.*, 2020a] (Chapter 7).

In this paper, Yilin Pan worked out the system structure, wrote the paper and did the experimental analysis. Prof Markus Reuber, Prof Annalena Venneri and Dr Daniel Blackburn gave the knowledge support from the clinical perspective. Dr Bahman Mirheidari and Prof Heidi Christensen discussed the related work with me in the weekly meeting. The ASR system used in the experimental part was provided by Dr Bahman Mirheidari.

Pan, Y., Mirheidari, B., Tu, Z., O'Malley, R., Walker, T., Venneri, A., and Christensen, H. (2020). Acoustic Feature Extraction with Interpretable Deep Neural Network for Neurodegenerative related Disorder Classification. In Proceedings of Interspeech 2020, 4806-4810 [Pan *et al.*, 2020b] (Chapter 6).

In this paper, Yilin Pan worked out the system structure, wrote the paper and did the experimental analysis. Zehai Tu provided insight on analysing the filters in the SincNet layer. Ronan O'Malley, Dr Traci Walker and Prof Annalena Venneri contributed to the collection and annotation of the dataset used in the paper. Dr Bahman Mirheidari and Prof Heidi Christensen discussed the related work with me in the weekly meeting.

Pan,Y., Nallanthighal, V., Blackburn, D., Christensen, H., and Härmä, A, (2021). Multi-task Estimation of Age and Cognitive Decline from Speech. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2021, 7258-7262 [Pan *et al.*, 2021b] (Chapter 9).

This paper was based on my work while I interned at Philips. Dr Venkata Srikanth Nallanthighal and Dr Aki H "arm "a was my colleague and supervisor there. We had meetings each week to discuss the related work. Dr Daniel Blackburn gave me clinical

knowledge support. This paper was revised by Prof Heidi Christensen.

Mirheidari, B.*, **Pan, Y.***, Blackburn, D., O'Malley, R., and Christensen, H. Towards clinically meaningful automatic assessment of cognitive health using spontaneous speech (* These authors contributed equally; To be submitted; Chapter 8).

I worked together with Dr Bahman Mirheidari, so we are the co-author of this paper. Only the part I did in Chapter 8 was summarised in the paper. Ronan O'Malley contributed to the collection and annotation of the dataset used in the paper.

Chapter 2

Dementia

Contents

2.1 What is Dementia? 17

2.1.1 Different Causes of Dementia 18

2.1.2 Stages of Dementia 19

2.2 Effect of Dementia on Speech and Language 21

2.2.1 Effect on Speech 21

2.2.2 Effect on Language 23

2.3 Current Diagnostic Procedures 24

2.3.1 Elements of the Diagnostic Procedures 25

2.3.2 Cognitive Assessment Tests 26

2.3.3 Speech- and Language-based Tasks for Dementia Detection . . 30

2.3.4 Problems with Existing Diagnostic Procedures 32

2.4 Summary 32

As outlined in Chapter 1, this research project aims to investigate methods for the early detection of signs of dementia by automatically analysing the acoustic and linguistic information in a person's speech. Under this motivation, this chapter will provide an introduction to dementia.

Dementia is an umbrella term used to define the loss of cognitive functioning leading to impairment in function. This chapter will focus on the symptoms of dementia and its impact on speech and language abilities. The existing neuropsychological assessment tools for dementia detection and their drawbacks are also introduced. This chapter is structured as follows:

Section 2.1 includes the definition of dementia and its different types together with the symptoms at different stages.

Section 2.2 describes how dementia affects speech and language abilities.

Section 2.3 briefly introduces current diagnostic approaches and their drawbacks.

Finally, **Section** 2.4 summarises the key information for this chapter.

2.1 What is Dementia?

The term *dementia* is used to define the loss of cognitive functioning, including memory, speech and language, visual perception, problem-solving, self-management, the ability to focus or pay attention, and behaviour abilities that may interfere with a person's daily life and activities [Hamilton, 2005; Wyss-Coray & Rogers, 2012]. Even though cognitive function loss also happens in healthy older people, they do not lose independence due to cognitive loss.

The diseases that cause dementia are incurable as the neurons in the brain cannot be replaced once they have died. Though incurable, early detection enables early treatment that can slow down and maintain mental function, meaning the early diagnosis of dementia is of great importance. As a progressive condition, the symptoms gradually get worse as the brain shrinks with the nerves dying [Bayles *et al.*, 1987].

Dementia can be caused by different diseases, including the Alzheimer's Disease (AD), Vascular Dementia, Dementia with Lewy Bodies, Parkinson's Disease, Fronto-Temporal Dementia and mixed dementia. Neurodegenerative Disorders (NDs) are caused by slow progressive loss of neurons in the central nervous system leading to an irreversible selective loss of brain functions, causing dementia. AD is the most common cause of dementia, accounting for approximately 60-70% of all dementia cases [Alzheimer's Society, 2018]. Mixed dementia is defined as a combination of more than one cause of dementia. Statistically, at least one in every ten people living with dementia is diagnosed with mixed dementia, and the percentage is even higher among the older age groups.

There is a large number of people living with dementia. In the UK, over $850,000$ people are living with dementia, and this situation is set to rise to over one million by 2025. Currently, one in fourteen people over 65 live with dementia, while for people over 80, one in six people are affected.

Symptoms of dementia depend on which part of the brain is damaged and what kind of diseases are causing dementia. Also, different stages of dementia have different symptoms. Moreover, the symptoms are often mild at the beginning and therefore harder to pick up, which makes getting a diagnosis early challenging. In the following, different types of

dementia are presented together with their related symptoms. The stages of dementia and the related symptoms are also summarised.

2.1.1 Different Causes of Dementia

As mentioned above, **Alzheimer's Disease**, is the most common cause of dementia. The exact cause of AD is still unknown. What is known is that 'plaques' and 'tangles' form in the brains of those that are living with AD (as shown in Figure 1 in Holston [2005]). In 2018, the US National Institute on Aging and the Alzheimer's Association proposed "A/T/N" ("A" refers to the value of a β-amyloid (Aβ); "T" refers to the value of tau bio-marker; "N" represents Neurodegeneration) to be used for defining and diagnosing AD [Jack Jr *et al.*, 2018]. As described in Gauthier [2001], the mood is first affected by AD, followed by the cognitive and functional abilities decline.

Vascular Dementia, as the second most common cause of dementia, is caused by problems with blood supply to the brain cells. As with AD, the symptoms tend to get worse over time. It can result in difficulties with planning, concentrating and understanding [Alzheimer's Society, 2022]. In addition, a person's mood, personality, and behaviour can also experience a change. People with Vascular Dementia tend to feel confused and disoriented.

Fronto-Temporal Dementia is a leading type of early-onset dementia [Vieira *et al.*, 2013]. Fronto-Temporal Dementia is caused by damage to cells in areas of the brain called the frontal and temporal lobes [Englund *et al.*, 1994]. The main symptoms can be classified into language, neuropsychiatric and other domains. Understanding language and factual knowledge are the areas most affected with respect to the language side, as the temporal lobe supports the functions. The symptoms include things such as word-finding difficulties, naming difficulties, and word repetition. Neuropsychiatric symptoms include changes in personality, loss of empathy, obsessive behaviours and delusions.

Parkinson's Disease is caused by a loss of brain nerve cells called substantia nigra that are responsible for producing dopamine. *Dopamine* is a neurotransmitter that acts as a messenger between parts of the brain and the neural system. As the amount of dopamine declines, the body's ability to coordinate declines [Fearnley & Lees, 1991]. People living

with Parkinson's Disease experience mobility problems (tremor and slowness to move) and can develop personality and behaviour changes, language problems (speaking slowly) [Politis *et al.*, 2010]. Generally speaking, the most common symptoms are obsessiveness, memory loss and a decline in controlling emotion.

Dementia with Lewy Bodies accounts for up to 20% of the cases of dementia that reach autopsy [McKeith, 2007]. There is some overlap between Dementia with Lewy Bodies and Parkinson's Disease in terms of medical features [Jellinger & Korczyn, 2018; McKeith *et al.*, 2017]. However, functional brain imaging with Positron Emission Tomography (PET) and postmortem studies have revealed some differences and overlappping biomarkers. For full details of the overlap and dissimilarities of two diseases, please see Table 1 in Jellinger & Korczyn [2018].

In conclusion, people living with different types of dementia have different symptoms, though they also share some symptoms. In general, the shared symptoms include perception, memory loss, mood changes, decline in speech and language abilities such as word-finding difficulties, confusion of time and place, and difficulties in carrying out daily tasks [Alzheimer's Society, 2020]. In Section 2.1.2, the symptoms at each stage are presented.

2.1.2 Stages of Dementia

Generally, the stages of dementia can be divided into three signs of degradation: early (mild), middle (moderate) and late (severe) stage [Cohan, 2013; Klimova *et al.*, 2015]. In addition, a pre-clinical stage and a Mild Cognitive Impairment (MCI) stage [Ringman *et al.*, 2008], that are also related to the development of dementia, are also introduced in this section.

In the pre-clinical stage, symptoms of cognitive function are normal, but pathological changes have started. As introduced by Jack Jr *et al.* [2013], at the beginning of this stage, *amyloid* may be the first positive marker, as proposed by Jack *et al.* [2013]. However, in the pre-clinical stage, there is no obvious chance of cognition. The term MCI is generally used to refer to a transitional state between the pre-clinical stage and AD. In Meilán *et al.* [2020] and Blackburn *et al.* [2014], MCI is classified into sub-categories. In our research,

no distinction is used between these sub-categories of MCI.

In the stage of MCI, compared with the pre-clinical stage, people experience a more pronounced cognitive decline. Approximately 12% to 18% of people age 60 or older are living with MCI [Alzheimer's Association, 2022]. About 10% to 15% of people living with MCI develop into living with AD per year (around 50% in 5 years) [Petersen *et al.*, 1999], while others may revert to normal cognition or remain stable. The diagnosis of the MCI is of great importance for delaying the development of dementia because the AD is irreversible and incurable currently, but it is hoped that starting treatment at the stage MCI might be able to stabilise and either slow down or halt progression to dementia.

People living with early-stage dementia (also known as mild stage dementia) usually experience short-term, and medium-term memory changes [Klimova *et al.*, 2015]. For different causes of dementia in the early stage, the common symptoms include memory problems, slow speed of thought, language and a decline in perception ability [Landspítalinn *et al.*, 1993]. People living with dementia may find it difficult to detect emotions (such as sarcasm or humour) embedded in speech [Ehernberger Hamilton, 1994]. Also, people living with the early stages of dementia tend to struggle to find suitable words and use limited words in their spoken language. In addition, they sometimes cannot concentrate on a specific topic while speaking. These symptoms affect language and speech ability at the same time. Due to this deterioration, their emotions are affected, and they more easily get depressed, anxious and feel frustrated [Klimova *et al.*, 2015].

In the middle (moderate) stage, compared with those living with early-stage dementia, people have more severe naming issues, more difficulties in maintaining meaningful, topic-relevant conversations and in speaking fluently in their speech [Ehernberger Hamilton, 1994]. In addition, they tend to repeat words and sentences more frequently. Compared with the early stage and the late stage, this stage lasts the longest, which is about 2 to 10 years [Central, 2020]. In this stage, people may also have symptoms of disorientation in their daily life, both in time and space [Lanza *et al.*, 2014].

In the late (severe) stage, symptoms such as disorientation relating to time, place and person may occur. At this stage, people are often not even aware of the presence of others and speak less [Ehernberger Hamilton, 1994]. In addition, social skills, reasoning ability,

and judgment ability may be lost in their life [Klimova *et al.*, 2015].

In conclusion, subtle language deficits exist even in the early stages of the disease and get more pronounced as the disease progresses [Forbes *et al.*, 2002, 2004; Forbes-McKay *et al.*, 2005; Vuorinen *et al.*, 2000]. Early diagnosis of dementia is of significant importance as it can ensure early treatment and access to support for their families. Our research concentrates on detecting dementia in the early stages, so speech and language ability are the two most informative biomarkers for dementia detection. Therefore, in the next section, the effect of dementia on speech and language in the early stages are summarised.

2.2 Effect of Dementia on Speech and Language

People with different types of dementia experience a decline in their speech and language abilities, even in the early stages. Our studies concentrate on diagnosing people living with or without dementia in the early stages without discriminating different causes of dementia. In this section, the effects of dementia on speech and language in the early stages are summarised.

2.2.1 Effect on Speech

As described in section 2.1 neural damage in the brain caused by dementia can lead to difficulties with people's spoken language, resulting in changes in speech characteristics. The acoustic information embedded in the speech that can be affected by dementia is summarised into three categories in this section: articulatory characteristics, vocal quality, and prosodic characteristics.

The movement of the articulators is known to be affected by the development of dementia [Baddeley, 1983; Morris, 1987; Östberg *et al.*, 2009]. The articulators include mandible, lip, tongue, velum, pharyngeal constrictors, and larynx [Wilhelms-Tricarico, 1995]. In Ackermann *et al.* [1997] and Ackermann & Ziegler [1991], some kinematic parameters are calculated for determining the movement and velocity trajectories of the articulators. The results show that slow and reduced articulatory movements of the lips and the jaw is shown in Parkinson's disease. In McRae *et al.* [2002], vowel duration, conso-

nant duration, articulation rate, vowel space area, and first-moment coefficient difference measures are each calculated based on the measurement of the articulators'movement. To capture the articulatory characteristics of the pathological speech, spectral features have been used in automatic dementia detection studies [Horwitz-Martin *et al.*, 2016; Sandoval *et al.*, 2013] (more details in Chapter 4).

In this thesis, *vocal quality* is referred to as the properties of speech affected by the organs inside the larynx. As in Sundberg & Sataloff [2005], the vocal quality relates to the shape of the vocal tract and vocal folds. The vocal tract and vocal folds are both partly genetic and determined at birth but can be affected by trauma or disease, such as Alzheimer's Disease [Gómez-Vilda *et al.*, 2015]. The measurement of vocal quality is difficult and always needs a combination of approaches [D'haeseleer *et al.*, 2017; Eskenazi *et al.*, 1990; Prosek *et al.*, 1987]. The commonly used measurements for describing the pathological voice quality include jitter (variation in frequencies) [Horii, 1979], shimmer (variation in amplitude) [Horii, 1980] and harmonic-to-noise ratio (ratio of format harmonics to inharmonic spectral peaks) [Yumoto *et al.*, 1982].

Prosody refers to the rhythm, stress and melody of speech [Nooteboom *et al.*, 1997]. It is concerned with the length and stress of syllables and units composed of several individual phonetic segments such as vowels and consonants. For measuring the rhythm of speech, the duration of the voice segments, duration of the silence segments, the variation of the fundamental frequency (F_0), and many other similar features can be used [Deterding, 2001]. Spontaneous speech from people living with dementia exhibits less fluency, more voiceless speech, and more repetitions compared with HCs [Meilán *et al.*, 2020]. In phonetics, stress is the degree of intensity given to the syllable in an utterance. It can be described by the loudness and energy of the syllable. The variation in stress patterns can result in different meanings and emotions embedded in the sentences of the exact same words. It has been found that the stress pattern mistakes in speech increase when the disease progresses [Colombo *et al.*, 2000, 2004]. The melody of the speech can be measured by "melodic line" (defined as the appropriate use of intonational contour, including alterations in F_0, volume, and duration [Lehiste, 1970]). Baum & Pell [1999] reviews the research relating to the analysis between prosody change and neurological damage.

The results in Ross *et al.* [1988] reveal that mean F_0 and F_0 standard deviation reduce when the right hemisphere of the brain deactivates.

In summary, speech characteristics are affected by dementia in different ways. This section summarised the symptoms into three aspects: the effect on articulatory movements, vocal quality, and prosody. In the following, the effects of dementia on the language are summarised.

2.2.2 Effect on Language

The memory deficits, attentional deficits, and visual perceptual problems resulting from dementia can all affect language ability. As dementia progresses, almost all aspects of language can be affected [Groves-Wright *et al.*, 2004]. For connected speech, the effects can be summarised into two categories: word-level and sentence-level.

The word-level analysis refers to the analysis of the separate words in the speech, including vocabulary size, informative units, the number of words and unique words in speech. For example, Burke & Shafto [2011] and Kemper *et al.* [2001] found that vocabulary size increases with ageing before it declines slightly at a certain age. However, for people living with dementia, vocabulary size declines much more rapidly, especially for those low-frequency and more specific words [Bird *et al.*, 2000; Burke & Shafto, 2011; Maxim & Bryan, 1994]. Compared with the HC group, the Dementia group tends to produce fewer informative units [Croisile *et al.*, 1996] and the language is less rich. For example, a longitudinal study on U.S. President Ronald Reagan (who was diagnosed with AD in his late-life) concerning his free-style speech showed that the percentage of unique words declined over the last several years of the presidency [Berisha *et al.*, 2015]. Whether the quantity of words in the speech from people living with dementia and HCs is informative for dementia detection is controversial. Giles *et al.* [1996] and Croisile *et al.* [1996] reported that people living with dementia are less talkative, whereas other researchers [Bschor *et al.*, 2001; Smith *et al.*, 1989; Tomoeda & Bayles, 1993] found there is no difference between the number of words in the speech from the two groups. The word-level information has been manually analysed by some specifically designed measurements, such as the *anomia index* [Hier *et al.*, 1985].

The sentence-level or semantic-level analysis refers to the amount of meaningful information covered in sentences (conciseness) [Smith *et al.*, 1989], the contextual relevance between sentences (local coherence) and whether a sentence relates to the current central topic (global coherence) [Laine *et al.*, 1998]. In Laine *et al.* [1998], *local coherence*, *global coherence*, *use of non-referential lexical items* and *informativeness scale* are used as the criteria for evaluating the severity of people living with AD, vascular dementia and HCs. The results show that global coherence is more informative than local coherence in the early stages of dementia, and the informativeness scale is more informative than non-referential lexical items. In Hier *et al.* [1985], *conciseness index* is proposed for evaluating the conciseness of expression. In Glosser & Deser [1991], multiple evaluation criteria for thematic coherence, cohesion, and thematic coherence are proposed to evaluate the development of dementia.

In summary, the linguistic ability in the word-level and sentence-level is affected, even in the early stages of dementia. Therefore, for detection, some criteria are proposed to evaluate the linguistic ability of speech from people who might have dementia.

2.3 Current Diagnostic Procedures

There is a wide range of neuropsychological diagnostic tests for determining if a person is living with dementia. Doctors diagnose the people who might live with dementia by a series of procedures [Central, 2020].

The diagnostic procedure includes clinical evaluation, cognitive assessment, basic laboratory tests and structural imaging. The clinical evaluation is based on patient history, family interview, and physical examination [Feldman *et al.*, 2008].

For dementia detection, the most common diagnostic procedure is as below. At first, people or their families notice their memory problems, and they suspect they may have dementia, like AD. Then, they visit their General Practitioner (GP) who may refer to a psychiatrist or neurologist for diagnosis. The doctor will take a "history". In this process, families or friends are usually required to accompany them for the consultant to take a collateral history. Whether people have dementia can be diagnosed by a series of

cognitive tests. They may also have laboratory testing and structural imaging to exclude the possibility of other diseases.

The following section will introduce the current elements of the diagnostic procedures used for dementia detection, followed by a more detailed introduction of popular cognitive assessment tools. Then, the speech- and language-based tasks designed for dementia detection are introduced. Finally, the problems of the current diagnostic procedure are summarised. Dementia screening tools are designed according to the symptoms for screening for dementia in clinical or research settings. Compared with the other diagnostic procedures summarised in Section 2.3.1, most of the screening tests take less than 30 minutes. In addition, the process is easy to carry out. Detailed information about the screening tests is introduced below.

2.3.1 Elements of the Diagnostic Procedures

The diagnostic procedure aims to determine whether a person's symptoms can be classified as dementia and what might be causing dementia. Knowing these can help the doctor carry out a targeted treatment plan. The procedures are shown as below:

1. Patient history: The clinician carries out a medical history inquiry about the psychiatric history and history of cognitive and behavioural changes of the person. Whether their families were living with dementia previously will also be asked, depending on the genetic influence. Also, the history taking should consider the disease that can indirectly cause dementia, such as stroke.

2. Family interview: The interview is between the family members or caregiver of the patient and clinicians. It can help the doctor find out about functional impairments such as the attention to hobbies [Feldman *et al.*, 2008], that the family members or caregiver might have noticed. Sometimes, the people living with dementia in the middle or late stages cannot answer the questions asked by the doctor independently, and their families can provide the answer to the doctor as the assistant.

3. Physical examination: The examination is targeted at examining whether a person living with cognitive decline has dementia or other diseases that have a high risk of

developing into or overlapping with dementia, such as stroke and heart disease. The examination includes gait, balance, sensory deficits, and other neurological changes [Finkel & Woodson, 1997].

4. Cognitive assessment: Cognitive assessment tests can be marketed as screening tests for evaluating the executive function, judgement, memory, attention and language. The most commonly used assessment test is the Mini Mental Status Examination (MMSE) test. More details about the Dementia assessment tools are introduced in Section 2.3.2.

5. Laboratory testing: Laboratory tests are designed to rule out the diseases that can cause chronic confusion and memory loss [Feldman *et al.*, 2008]. For testing, the blood sample can be collected. As mentioned in Section 2.1.1, Aβ and tau are the biomarkers of the pathogenesis of AD [Selkoe & Hardy, 2016]. Already in 2016, it was found that the secondary structure distribution of $\alpha\beta$ in blood plasma can reflect the $\alpha\beta$ burden in the brain, which can be used for dementia detection [Nabers *et al.*, 2016a,b].

6. Structural imaging: It can also be referred to as a 'brain scan'. In general, the brain scan is based on the Magnetic Resonance Imaging (MRI) or Computed Tomography (CT) for ruling out other causes that have similar symptoms as AD. In addition, the Positron Emission Tomography (PET) is a functional scan that can be helpful in neurodegenerative disorders detection.

Though the full neuropsychological evaluation is sensitive and specific, the process is time-consuming and expensive. In comparison, cognitive assessment tools can be used alone as screening tools because they are faster and more accessible than the full neuropsychological evaluation. In Section 2.3.2, the popular dementia assessment tools are introduced. The main information of these dementia assessment tools is summarised in Table 2.3.2.

2.3.2 Cognitive Assessment Tests

The commonly used assessment tests are summarised in Table 2.1. More detailed information about each screening tool is introduced below.

Table 2.1: Summary of neuropsychological assessment tools for detecting cognitive decline.

Assessment Tool	Usage	Assessment Skills
Mini Mental Status Examination (MMSE)	Measure cognitive impairment	Orientation to time and place, repeating/remembering words, calculation, naming objects
Montreal Cognitive Assessment (MOCA)	Specially designed for MCI	Visuo-constructional, (drawing cubes and clocks), naming, repeating, short-term memory, verbal fluency, attention and abstraction
Addenbrooke's Cognitive Examination (ACE)	Evaluate cognitive skills	Attention/orientation, memory, language, verbal fluency, and visuospatial ability
Boston Naming Test (BNT)	Assess language difficulties	Naming items, picture description ability
Wechsler Adult Intelligence Scale (WAIS)	Intelligence quotient (IQ) test	Verbal Comprehension Index, Perceptual Reasoning Index, Working Memory Index, and Processing Speed Index
Wechsler Memory Scale (WMS)	Assess different memory functioning	Logical memory, verbal paired associates, visual reproduction, brief cognitive decline exam, designs, spatial addition and symbol span
Patient Health Questionnaire-9 (PHQ-9)	Detect depression and its severity	Assess features like having little interest/pleasure doing things, feeling down/hopeless, sleeping trouble, tiredness, poor appetite or overeating, feeling bad about yourself, trouble concentrating, moving/speaking slowly or feeling restless, thought of better being dead or self-harm [Inoue *et al.*, 2012]
Generalised Anxiety Assessment-7 (GAD-7)	Assess generalised anxiety and its severity	Seven questions designed for assessing anxiety

The Mini Mental Status Examination (MMSE) [Folstein *et al.*, 1975] is the most commonly used cognitive evaluation test. The measurements cover five cognitive domains: orientation, registration, attention and calculation, recall, and language. The maximum achievable score in the test is 30, and the cut-off point for being considered healthy is 25. Those who achieve a score between 10 and 20 show a moderate impairment, and a score less than 10 represents a person living with severe cognitive impairment. It takes about 10 minutes for testing.

The Montreal Cognitive Assessment (MoCA) is also widely used for cognitive decline stratification, and it is more sensitive to MCI [Nasreddine *et al.*, 2005]. In addition, it can also recognise different causes of cognitive decline, such as Parkinson's disease [Zadikoff *et al.*, 2008], cerebrovascular disease [Pendlebury *et al.*, 2010], Huntington's disease [Videnovic *et al.*, 2010], and brain metastases [Olson *et al.*, 2008]. The test is available in multiple languages. The test lasts about 10 minutes, and it is designed for evaluating the abilities of visual-constructional (drawing cubes and clocks), naming, repeating, short-term memory, verbal fluency, attention and abstraction. For the MoCA test, the cut-off score is 26, and anyone who gets a score lower than 26 indicates cognitive impairment.

The Addenbrooke's Cognitive Examination (ACE), introduced by Mathuranath *et al.* [2000], examines five cognitive skill and its highest point is 100. The scores of the five parts are 18 points for attention, 26 points for memory, 14 points for fluency, 26 for language and 16 for visuospatial ability. Compared with the MMSE, which a general practitioner can administer, the ACE requires more specific knowledge. This test can not only detect dementia, but also differentiate AD from Fronto-Temporal Dementia (FTD) with its subscore: the VLOM ratio [Mathuranath *et al.*, 2000]. The test is about 15 minutes. It has been reported that the sensitivity and specificity of the ACE are high and it has good patient acceptability. The test has been translated into other languages. [Bier *et al.*, 2004; García-Caballero *et al.*, 2006; Mathuranath *et al.*, 2004].

The Boston Naming Test (BNT), introduced by Edith Kaplan, Harold Goodglass and Sandra Weintraub, was designed for evaluating language difficulties [Goodglass *et al.*, 1983]. Sixty pictures are provided, and patients are required to name objects in the

pictures (line-drawing) within 20 seconds. The objects include both the commonly used objects, like 'tree', to rare objects, like 'abacus'. Previous research has mentioned that the BNT is efficient in dementia detection, which can detect dementia even in the very early stages [MacKay *et al.*, 2005]. On the other side, previous review shows that the BNT is sensitive to age [Albert *et al.*, 1988; Borod *et al.*, 1980; Goodglass, 1980], and education is also related to the performance [King *et al.*, 1993; Nicholas *et al.*, 1985a]. Also, it is copyrighted and needs to be purchased for detection.

The Wechsler Adult Intelligence Scale (WAIS) is the most commonly used test for evaluating the intelligence quotient (IQ). The fourth version (Wechsler Adult Intelligence Scale-IV), includes four main index scores: Verbal Comprehension Index, Perceptual Reasoning Index, Working Memory Index, and Processing Speed Index [Wechsler, 2008]. Different versions of the WAIS have been used for dementia detection [Donnell *et al.*, 2007; Izawa *et al.*, 2009; Shimomura *et al.*, 1998].

The Wechsler Memory Scale is designed for memory functioning evaluation [Wechsler, 1945]. It has been revised several times, and the latest version (Wechsler Memory Scale-IV) contains seven sub-tests, as shown in Table 2.1. There are five index scores for evaluating the person's performance: auditory memory, visual memory, visual working memory, immediate memory, and delayed memory [Carlozzi *et al.*, 2013].

The Patient Health Questionnaire-9 [Kroenke & Spitzer, 2002] consists of 9 questions, and the highest point for each question is 3, making 27 in total. Different scores refer to the different stages of depression. The higher, the more severe. A score below 4 refers to normal. The questionnaire can be found in Curriculum [2022].

Similarly, the Generalized Anxiety Disorder assessment-7 (GAD-7) includes seven questions, and the highest value for each question is 3 [Spitzer *et al.*, 2006]. 0 represents not at all and 3 represents nearly every day. The cut-off points are 5, 10, 15 for mild, moderate and severe stages of anxiety [Arthurs *et al.*, 2012]. These cognitive assessment tests are faster and more convenient than the other doctor diagnostic procedures summarised in Section 2.3.1. For testing, people need to go to the clinic and get the test from clinicians or neurologists. The tests can be used alone or together with other tests for dementia diagnosis and severity evaluation. The GAD-7 has been used in the studies

from different countries, including English, German, and French studies [Oechsle *et al.*, 2013; Spitzer *et al.*, 2006; Wild *et al.*, 2014].

2.3.3 Speech- and Language-based Tasks for Dementia Detection

As mentioned in Section 2.2, speech and language abilities are affected at the early stages of dementia. In addition to the diagnostic procedures introduced above, specific tests have been designed to test specifically speech and language abilities for dementia detection [Cerhan *et al.*, 2002; Chatwin, 2014; Mueller *et al.*, 2018; Sherod *et al.*, 2009]

Speech-based dementia detection is based on both isolated and connected speech. Isolated speech tasks require the participants to say isolated words rather than a sentence or a whole paragraph. It can check the participant's semantic ability, like 'animal naming', or phonemic fluency, like 'saying words beginning with **p**'. Connected language analysis, including picture description, conversation analysis and story recall test, is also widely used for testing episodic memory, semantic processing skills, grammatical constituents, and syntactic complexity [Mueller *et al.*, 2018]. In this section, the tasks designed for dementia detection based on isolated speech and connected speech are introduced.

The verbal fluency test is a task designed for testing verbal functioning [Muriel Lezak *et al.*, 2012]. It can be categorised into a semantic subtest and a category subtest. The two tests are designed to ask the participants to verbally list as many things as possible according to the requirements within 60 seconds. For example: " Please name as many animals as you can in one minute, you can name any animal, you may now begin". A common finding reported in Caccappolo-Van Vliet *et al.* [2003]; Monsch *et al.* [1992]; Ober *et al.* [1986] is that people living with AD produce fewer words compared with the HC group when carrying out the fluency test. In addition, semantic fluency test is demonstrated to be more powerful than phonemic fluency test in discriminating between people living with AD and HCs [Cerhan *et al.*, 2002].

The picture description task is a constrained task that relies less on episodic memory but requires more semantic knowledge and retrieval ability [Mueller *et al.*, 2018]. A

picture is presented as a prompt in the test, and the person is asked to describe what they have seen in the picture. During this process, the answer is often recorded, and this is subsequently used when the neuropsychologist scores the test. The most commonly used picture is a line drawing called the "Cookie Theft" (shown in Figure 2.1). This picture originates from a test for aphasia [Goodglass & Kaplan, 1972].

Figure 2.1: The Cookie Theft picture from the Boston Diagnostic Aphasia Examination.

Three pioneers (Emanuel Schegloff, Harvey Sacks and Gail Jefferson) initially introduced conversation analysis around 1967/1968. It is designed to investigate the structural organisation carrying out everyday social interaction. The conversations took place between two or more people. It has been used for dementia detection in a home-based environment [Chatwin, 2014].

The story recall test [Green & Kramar, 1983] is designed for evaluating the verbal memory function, which requires the participants to recall details of a story that is told or read to them. It is one of the most reliable neuropsychological assessments for distinguishing between normal ageing, MCI, and AD [Baek *et al.*, 2011].

For these tests, the speech is recorded for subsequent analysis. The speech- and language-based symptoms caused by dementia have been summarised in Section 2.2. For detecting dementia, the acoustic information and linguistic information can be extracted

from the audio recordings and the manual transcripts for analysis.

2.3.4 Problems with Existing Diagnostic Procedures

Currently, the diagnostic procedures can usually be categorised into invasive and non-invasive diagnostic procedures. The invasive diagnostic procedures, like blood tests, can cause increased anxiety and stress before diagnosing people. Also, even though these tools can detect the disease before it is noticed at the early stage, they are costly and not suitable for large-scale stratification purposes. For the other category, the non-invasive diagnostic procedures, like cognitive assessment tests and family interviews, are not as expensive as the invasive procedures. However, it is both time-consuming, and energy-consuming as people who are worried about their health need to go to secondary care for diagnosis as GPs are not consistently accurate. However, the wait for the secondary care diagnosis can be long.

The speech- and language-based tasks that can be used for spoken language collection are shown in Section 2.3.3. The collected spoken language is transcribed manually into text and analysed manually by clinicians or neurologists in the clinic. The analysis process is most time-consuming. To ease the burden of attending a clinic for diagnosis, the automatic speech- and language-based dementia detection methods are of great importance. One promising way to do this is by developing tools that assess people's cognitive impairment based on their speech and language and diagnosis automatically, which can be used for assisting the GPs when deciding whether to refer to secondary care. Such a test could potentially be done in a home-based environment. Previous automatic dementia detection research based on speech and language will be reviewed in Chapter 3.

2.4 Summary

Dementia is an umbrella term for defining the decline in cognitive abilities that may interfere with a person's daily life and activities. The cause of dementia can be categorised into AD, Vascular Dementia, Dementia with Lewy Bodies, Parkinson's Disease, Fronto-Temporal Dementia and mixed dementia. Among different causes of dementia,

AD accounts for 60%-70% of cases, which is the most common cause of dementia. The symptoms of people living with dementia depend on the brain damage and what types of dementia they are living with. The diseases that cause dementia are incurable. Early detection is of great importance as it can enable early treatment, which can slow down and maintain mental functions.

The development of dementia is divided into three stages: early-stage, middle-stage and late-stage. The symptoms of different stages were summarised in this section. Both speech and language can be affected by dementia, even in the early-stage. The speech ability decline includes more frequent pauses and longer pauses caused by word-finding difficulties, vagueness in speech and a change in the F_0. The language ability decline includes being 'wordy', 'imprecise' and 'off-topic' in the speech, shown in the word-level and sentence-level.

The existing diagnostic procedures can be categorised into invasive and non-invasive procedures. The invasive diagnostic tests include laboratory testing, blood tests and structural imaging, while the most common non-invasive diagnostic procedure is based in the clinic. Though the diagnostic procedures are widely used, each procedure has its limitations with respect to sensitivity, specificity, convenience, and price.

The non-invasive cognitive assessment tests are designed according to the symptoms for screening dementia in clinical or research settings. Detailed information about these cognitive assessment tests was introduced in this section. The most commonly used cognitive evaluation test is the MMSE, which takes about 10 minutes to administer. These tests can be used alone or together with other tests for dementia diagnosis and severity evaluation.

With the increasing proportion of older people, our elderly society needs a non-invasive and reliable assessment tool for conveniently diagnosing people at risk of developing dementia. Therefore, automatic dementia detection methods should be explored. In the next chapter, recent research on automatic methods for speech- and language-based dementia detection methods are reviewed.

Chapter 3

Automatic Dementia Detection Methods

Contents

3.1 Automatic Linguistic-based Dementia Detection 37

3.1.1 Knowledge-based Linguistic Features 37

3.1.2 Natural Language Processing Technologies 39

3.2 Automatic Acoustic-based Dementia Detection 40

3.2.1 Knowledge-based Acoustic Features 41

3.2.2 End-to-end System based Acoustic Feature Extraction 42

3.3 Advantages of Speech- and Language-based Automatic Dementia Detection 49

3.4 Drawbacks of Existing Automatic Methods 50

3.5 Summary 51

Though automatic speech- and language-based dementia detection is a relatively new area of research, some significant approaches have been proposed by utilising machine learning methods for modelling the dementia symptoms automatically. Compared with the existing diagnostic procedures summarised in Chapter 2, the automatic dementia detection methods have many advantages, like low cost and convenience, which can be used for clinical stratification assistance and remote assessment. In this chapter, the automatic dementia detection related research is summarised and categorised into speech-based and language-based methods. The advantages and drawbacks of the related research are also presented. This chapter is structured as follows:

Section 3.1 summarises the methods proposed for language based dementia detection.

Section 3.2 reviews the methods proposed for speech based dementia detection.

Section 3.3 and **Section** 3.4 summarises the advantages of speech- and language-based automatic dementia detection, and lists the drawbacks of the current automatic methods respectively.

3.1 Automatic Linguistic-based Dementia Detection

As seen in Chapter 2, almost all aspects of language can be affected as dementia progresses. To represent the changes caused by dementia in language, researchers have proposed a bank of linguistic features to capture these changes. Some related research about speech- and language-based dementia detection has been reviewed in the recent papers [de la Fuente Garcia *et al.*, 2020; Mueller *et al.*, 2018; Petti *et al.*, 2020]. In this section, a review of both commonly features and state-of-the-art natural language processing (NLP) technologies used for linguistic-based automatic dementia detection is provided. *Traditional feature* in this thesis is used to refer to the features extracted according to a specific predefined algorithm based on expert knowledge, different from the features extracted from the deep neural networks.

3.1.1 Knowledge-based Linguistic Features

The effects of dementia on language ability have been summarised in Section 2.2.2. For detecting dementia automatically, knowledge-based features are used for representing the symptoms of dementia on language. The commonly used linguistic features like Part of Speech (POS) (reports the changes in the number of pronouns and verbs) [Jarrold *et al.*, 2014], Type-token ratio (TTR) (measures the vocabulary richness), hesitations related features, vocabulary variation and syntactic complexity evaluation features.

Jarrold *et al.* [2014] proposed to use both the word frequencies related feature [Pennebaker *et al.*, 2001] and POS tagging for extracting linguistic information from manual and automatic transcripts, respectively. In Khodabakhsh *et al.* [2015], the hesitation and puzzlement features (like incomplete sentence ratio), POS tagging, intelligibility and complexity features (like unintelligible word ratio) were extracted for AD detection. Rentoumi *et al.* [2017] represented the linguistic information embedded in the text with vocabulary variation and syntactic complexity. Toledo *et al.* [2018] analysed the sentence complexity and semantic richness for designing the linguistic feature. In Campbell *et al.* [2020], a 13-dimension feature related to the vocabulary richness, key informational concepts and features designed with the POS tagging was designed for the AD detection. Similarly, for

the MCI detection, Vincze *et al.* [2016] used features designed with the POS tagging as the linguistic features.

In addition, some statistical features, like the number of information units (a metric for quantifying the amount of information comprehended and reproduced in the context), the total number of utterances per patient, mean length of utterance (MLU), mean length of the clause and idea density (the number of expressed propositions divided by the number of words [Roark *et al.*, 2011]) have also been demonstrated to be effective for describing the linguistic ability decline in previous research. For the MCI detection, Roark *et al.* [2011] used idea density as one of the proposed linguistic features. The linguistic features used in Orimaye *et al.* [2017] included the number of utterances, MLU, sentence dependencies for the AD detection.

More complicated methods for the syntactic analysis of natural language can also assess cognitive decline. A commonly used method is *parsing* the language based on grammatical rules and language-dependent syntactical (also known as *syntax analysis*) [Lin, 1996; Magerman, 1995; Rentoumi *et al.*, 2017; Roark *et al.*, 2011; Yngve, 1960], which can measure the lexical and syntactical complexity of a sentence. The motivation of the parsing analysis is that either the working or semantic memory limitations caused by dementia can result in a decrease in complex grammatical construction ability. For parsing analysis, a parse tree needs to be built. In Roark *et al.* [2007], the syntactic complexity was measured with both the manual and automatic parse trees for the MCI detection.

In conclusion, most of the research is based on relatively small amounts of data, such as in Jarrold *et al.* [2014]; Khodabakhsh *et al.* [2015]. In addition, the variation in data quality and speech accent across the different datasets makes it harder to generalise the learned features. In addition, while diagnosing, a long list of knowledge-based features are required for describing the dementia symptoms embedded in the linguistic abilities [Khodabakhsh *et al.*, 2015; Roark *et al.*, 2011; Vincze *et al.*, 2016], making the feature design and selection very time-consuming. Under this situation, when faced with different classification tasks and datasets, deep learning technologies are expected to be used for extracting the data-driven features across different classification tasks and datasets.

3.1.2 Natural Language Processing Technologies

Various NLP technologies have been used for the automatic dementia detection and has gained considerable success, like the application of word vector embedding [Landauer & Dumais, 1997; Mikolov *et al.*, 2013; Pennington *et al.*, 2014], probabilistic language models [Bengio *et al.*, 2003] and the Bidirectional Encoder Representations from Transformers (BERT) [Devlin *et al.*, 2018].

The semantic similarity between different sentences can be calculated by embedding the text into a high dimensional vector space and estimating the distance between the word vectors. For mapping words into word vectors, word embedding methods are proposed. The commonly used un-supervised word embedding methods are the *word2vec* [Mikolov *et al.*, 2013] and *Glove* [Pennington *et al.*, 2014]. The supervised word embedding method is based on an embedding layer in an end-to-end neural network [Goldberg, 2017]. In previous research, the word vector embedding was used for representing the words in spoken transcripts into vectors for dementia detection, and promising results have been reported in Mirheidari [2018]; Yancheva & Rudzicz [2016]. In Yancheva & Rudzicz [2016], the Glove model was trained to convert the words into word vectors for training cluster models. In Mirheidari [2018], glove and word2vec were used for generating word vectors. Linear classifiers or neural networks can be used for classification, or the cosine distance can be used for calculating the word vector similarity.

Recently, BERT, which is based on the attention mechanism, has been proposed for generating word embeddings. In comparison with word2vec, which only processes a single word each time and outputs a single word vector. BERT takes a sequence of words (such as a sentence) as the input and outputs a fixed-length vector as the representation of that word sequence. Since being proposed, the end-to-end structure BERT has achieved excellent performance on multiple NLP research tasks. Also, BERT [Vaswani *et al.*, 2017] has been demonstrated to be effective in the linguistic-based dementia detection research since 2020 [Farzana & Parde, 2020; Koo *et al.*, 2020; Pompili *et al.*, 2020b; Searle *et al.*, 2020; Syed *et al.*, 2020; Yuan *et al.*, 2020]. More detailed information can be found in Section 3.2.2.4.

A language model can be trained for learning the joint probability of a sequence of

words in a language [Bengio *et al.*, 2003], and this has been used for exploring the long-term information embedded in the language for dementia detection. The language model technologies can be summarised into the statistical language model and neural network language model. A representative method of the statistical language model technology is the n-gram language model [Kneser & Ney, 1995], which estimates the possibility of the sequence with the Markov chain hypothesis [Gilks *et al.*, 1995]. In Fritsch *et al.* [2018], n-gram and neural network language models were both used for modelling the linguistic mismatch between the AD and HC groups, respectively. Specifically, the frequency of every single word or word sequence is modelled by the language models. The difference between the perplexity evaluated on the AD and HC language models are used for dementia detection. In Orimaye *et al.* [2017], to capture the difference of sequences of words between the HC and AD groups, both bigrams and trigrams were used for characterising the linguistic information. It was shown that top 20 n-gram (bigrams and trigrams) features extracted from the transcripts from the AD and HC show significant p-values at the 95% Confidence Interval.

Compared with the knowledge-based features mentioned in Section 3.1.1, deep learning based research has recently shown a lot of promise. However, unlike designing knowledge-based features, most of the deep learning based research did not consider much clinical expertise. For example, the systems were designed on full speech segments [Warnita *et al.*, 2018] without considering the decline in linguistic abilities that is known to exist at the word and sentence levels. Exploring how to incorporate such clinical expertise while using deep learning technologies is therefore of interest.

3.2 Automatic Acoustic-based Dementia Detection

The effects of dementia on speech have been summarised in Section 2.2.1. There are two main streams for speech-based automatic dementia detection methods used in previous studies: the knowledge-based features and features learned from the end-to-end system. In this section, the previous automatic speech-based dementia detection methods are summarised into the two categories and introduced in Section 3.2.1 and Section 3.2.2

respectively.

3.2.1 Knowledge-based Acoustic Features

The most popular speech-based knowledge-based features include the Mel Frequency Cepstral Coefficient (MFCC) [Davis & Mermelstein, 1980], Perceptual Linear Prediction (PLP) [Hermansky, 1990] and F_0 [De Cheveigné & Kawahara, 2002]. Furthermore, some statistical features, like the articulation rate, speech tempo, hesitation ratio, speech rate, voice duration and pause (silence) duration related features, are used for automatic dementia detection [Luz, 2009; Luz *et al.*, 2018]. Some acoustic feature sets, like the Interspeech 2010 Paralinguistic Challenge feature set (IS10) [Schuller *et al.*, 2010] as well as the ComParE 2013 feature set (ComParE) [Eyben *et al.*, 2013], are also adopted for acoustic-based dementia detection [Haider *et al.*, 2019]. The ComParE feature set was proposed for the Computational Paralinguistics Challenge which includes 6373-dimension features (the functionals applied energy, spectral, MFCC, and voicing related 65-dimension low-level descriptors (LLDs)) [Eyben *et al.*, 2013]. The IS10 feature set was proposed for the Interspeech 2010 Paralinguistic Challenge, which includes 1582-dimension features [Schuller *et al.*, 2010]. In addition, some features like the *conversation analysis features* and *acoustic-only features* proposed in [Mirheidari *et al.*, 2019a] are designed according to the knowledge of the clinical diagnosis.

A very early study for automatic dementia detection based on speech analysis was published by Hoffmann *et al.* [2010]. For analysis, four features were extracted from the speakers (articulation rate, speech tempo, hesitation ratio, and grammatical errors) using the PRAAT software [Boersma & Weenink, 1996] from spontaneous speech samples collected from three stages of AD (mild, moderate and severe) and HCs. In the same year, Horley *et al.* [2010] proposed to analyse the emotional prosody features (like the F_0) from speech samples collected from the sentence repetition task. In the next several years, the speech and pause duration, F_0, fluctuation in the frequency (Jitter) and amplitude (Shimmer), and the Harmonic to noise ratio (HNR) were proposed for dementia detection [König *et al.*, 2015; López-de Ipiña *et al.*, 2013; Martínez-Sánchez *et al.*, 2012; Meilán *et al.*, 2014; Roark *et al.*, 2011].

In 2016, the speech rate variability and higher-order spectra were analysed respectively in Nasrolahzadeh *et al.* [2016a] and Nasrolahzadeh *et al.* [2016b] on the speech samples collected from stories telling and conversations for the understanding of the speech pattern differences of the HC and AD groups. Similarly, in Haider *et al.* [2020], the expression of emotion reflected by voice volume, speech rate and fundamental frequency among AD and non-AD are different. In Beltrami *et al.* [2018], the combination of rhythm and acoustic features were extracted for classifying the recordings from the HC, MCI and AD. In Martínez-Sánchez *et al.* [2012]; Meilan *et al.* [2018]; Meilán *et al.* [2020], the analysis was based on the recordings collected from the participants while they were reading. The speech rhythm information like change in speech chunking and speech timing, voiceless segment variance, duration and phonation time were estimated for classifying [Mirheidari, 2018; Mirheidari *et al.*, 2016].

3.2.2 End-to-end System based Acoustic Feature Extraction

Deep learning has been demonstrated to be efficient in various research fields for extracting high-level representation from input data, including for various speech pathology tasks [Chung *et al.*, 2014]. However, until 2019 (after the studies in this thesis started), limited research was conducted using deep learning technologies for automatic dementia detection. It quickly became clear though that deep learning technology is efficient and should be explored for dementia detection. Therefore, this section reviews the most popular deep learning technologies at first, and in addition, the domains of research fields related to speech are reviewed.

3.2.2.1 Conventional Deep Neural Network

Deep neural network is a bio-inspired model [McCulloch & Pitts, 1943] which can also be referred to as feed-forward neural networks (FFN), or multi-layer perceptrons (MLP). A variant of feed-forward artificial neural networks has achieved great success across a range of speech processing tasks, such as speech recognition [Hinton *et al.*, 2012] and speaker recognition [Snyder *et al.*, 2018; Variani *et al.*, 2014]. The basic building block or unit of such networks is designed based on the neural node analogous to a biological neuron.

The layers between the input and output layer are named *hidden layers*. Deep Neural Network (DNN)s are typically comprised of multiple hidden layers, which can extract high-level information from the input data. For each hidden layer, a non-linear function is used, like the Rectified Linear Unit (ReLU) [Nair & Hinton, 2010], Logistic, and Tanh function [Karlik & Olgac, 2011], to introduce non-linearities into the network. The chain of layers and the non-linearities introduced by the activation functions enable the network to learn complicated relationships between the inputs and outputs. Back-propagation is used in the DNN training to minimise the discrepancy between the target outputs and actual outputs [Rumelhart *et al.*, 1988], which is measured by using a loss function. DNNs have been used for dementia detection [Orimaye *et al.*, 2016].

3.2.2.2 Recurrent Neural Network

An Recurrent Neural Network (RNN) is a feed-forward network that can capture dynamic information from time series data, like speech, via cycles in the network of nodes. With a bi-directional RNN, it can learn both the previous and the future context in the neighbouring sequence [Schuster & Paliwal, 1997]. However, RNNs lose the gradient value in long time sequences and cannot learn long-term temporal contextual information properly [Hochreiter, 1998]. To overcome this drawback, Long Short-Term Memory (LSTM) [Hochreiter & Schmidhuber, 1997] was proposed, which can extract the long-term information in speech more efficiently. Specifically, unlike the traditional recurrent unit in the RNN, which overwrites its hidden state at each time step, the forget gate in the LSTM can decide whether to keep the existing memory in a hidden state. The LSTM can capture potentially longer-distance dependencies with the gate mechanism. Similarly, the Gated recurrent unit (GRU) is also proposed for making each recurrent unit capture long-term dependency information, but without a separate memory cell compared with the LSTM unit [Cho *et al.*, 2014]. In the previous research, a system composed by the LSTM and Convolutional Neural Network (CNN) was proposed for depression detection [Ma *et al.*, 2016b].

3.2.2.3 Convolutional Neural Networks

The Convolutional Neural Network (CNN) is also a type of deep neural network, and the ideas about convolution and weight sharing have a very long history [Hubel & Wiesel, 1959, 1962]. CNNs became popular in the late 1990s [LeCun *et al.*, 1989]. It then achieved success in image processing, speech processing and natural language processing. A CNN block is generally composed of three types of layers: convolutional layers, pooling layers, and fully connected layers [Shen *et al.*, 2018]. The convolutional layers contain a set of convolutional kernels that can perform a set of filtering (kernel) operations. Each neuron of the filter is connected with a local receptive field of previous layers and extracts a local feature map. Non-linearities can be introduced to the feature extraction processes by using the non-linear activation function, such as ReLU, at the output of the convolutional layers. The role of the pooling layer is to reduce the dimensionality of each feature but retain the essential information, which allows the CNN to improve the robustness of the learned feature. The fully connected layers are required to achieve the prediction output. The final fully connected layer can be obtained by flattening or global pooling, followed by an activation function for prediction. In Warnita *et al.* [2018], a Gated Convolutional Neural Network (GCNN) was used to capture the temporal information in audio paralinguistic features based on the features extracted by the Opensmile [Eyben *et al.*, 2010].

3.2.2.4 Attention Mechanism

The attention mechanism is designed to encourage the neural networks to focus on the parts that should be paid attention to for the training target. It was introduced into the NLP field in 2014 [Cho *et al.*, 2014; Sutskever *et al.*, 2014]. In the beginning, in Bahdanau *et al.* [2014], an attention mechanism was designed between the decoder and encoder for deciding which parts of the source sentence should be paid attention to when doing neural machine translation. As described in Bahdanau *et al.* [2014], facing with a sequence of annotations output by the encoder $(h_1, ..., h_{T_x})$, the context vector c_i is generated by an attention mechanism:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_i \tag{3.1}$$

where α_{ij} is computed by:

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})} \tag{3.2}$$

where e_{ij} is output by an alignment model which can measure how well the inputs h_j around position j (in the encoder) and the output s_{i-1} at position i (in the decoder) match.

In Xu *et al.* [2015], two kinds of attention mechanisms (hard attention and soft attention) were used for object detection in images. *Soft attention* is the same as the attention mechanism proposed in Bahdanau *et al.* [2014], which was trained in the end-to-end system by using standard back-propagation based on all outputs of the encoder. In contrast, *hard attention* only depends on selected outputs of the encoder, so the Monte Carlo method [Metropolis & Ulam, 1949] is utilised for back-propagation. In Yang *et al.* [2016], a hierarchical based system was proposed for the classification task, rather than the sequence generation tasks in previous research [Bahdanau *et al.*, 2014; Cho *et al.*, 2014; Sutskever *et al.*, 2014; Xu *et al.*, 2015]. More detailed description about the hierarchical attention is presented in Chapter 5.

In 2017, the multi-head self-attention mechanism was proposed for constructing the transformer [Vaswani *et al.*, 2017] and BERT [Devlin *et al.*, 2018], which is marked as one of the significant breakthroughs of the decade in the NLP field. The soft attention and hard attention mechanisms are designed between the encoder and the decoder. However, the self-attention mechanism is the mechanism inside the encoder or the decoder. Compared with RNNs in the neural network, self-attention can learn longer-range term dependencies [Vaswani *et al.*, 2017]. The multi-head attention mechanism is composed of multiple self-attention mechanisms, which allows the proposed model to learn the information embedded in the input from different representation subspaces.

Currently, the attention mechanism has demonstrated its efficiency in speech recognition [Chorowski *et al.*, 2015], machine translation [Luong *et al.*, 2015] and image caption

[Xu *et al.*, 2015] related fields. It should be explored whether the attention mechanism can be used in speech and language-based automatic dementia detection.

3.2.2.5 Development of Speech-related Research Fields

Compared with the knowledge-based features, more information is presented in the raw waveform. However, the feature extraction process is more complex and challenging due to the high dimension and noise that exists in the raw waveform. Before 2018, very little research was done on extracting acoustic features from the raw waveform using deep learning methods for dementia detection. However, this is very popular in other speech-related fields, like speaker recognition, speech emotion recognition and speech recognition. The features used in speaker recognition and speech emotion recognition are also sometimes useful for dementia detection [Egas López *et al.*, 2019; Lugger & Yang, 2007; Zhang, 2008]. Therefore, the development of these fields inspires us to explore how to use the deep learning technologies for dementia detection. The research development histories of speaker recognition and speech emotion recognition are summarised in Table 3.1 and Table 3.2 respectively.

As shown in Table 3.1, before 1980, only the very general traditional acoustic features, like the Cepstral Measurements, the MFCC and LPCC were used as the acoustic representation for speaker recognition. However, in 2010, the i-vector was proposed and started to be used broadly as the speaker identity information representation [Dehak *et al.*, 2010].

Since 2016, deep learning technologies have been adopted for identifying the extracted speaker features [Ghahabi *et al.*, 2016] or extracting the speaker identity features [Snyder *et al.*, 2017, 2018; Variani *et al.*, 2014]. In comparison, the end-to-end system for speaker recognition is more difficult than deep learning-based speaker features extraction or identification, though they both rely on deep neural networks.

The speaker embedding vectors, like the i-vector or x-vector, are used as the input of Probabilistic Linear Discriminant Analysis (PLDA) for determining whether a pair of segments belong to the same speaker [Ghahabi *et al.*, 2016; Snyder *et al.*, 2017, 2018; Variani *et al.*, 2014]. The end-to-end system construction started from mimicking the i-vector + PLDA structures [Rohdin *et al.*, 2018, 2020] which use NN module for extracting

Table 3.1: Acoustic-based speaker recognition development history.

System Type	Features or Methods	Related Publications
knowledge-based features	Cepstral measurements	[Luck, 1969]
	Spectrogram	[Bolt *et al.*, 1970]
	Linear prediction model	[Crichton & Fallside, 1974]
		[Atal, 1976]
	Linear Prediction Cepstral Coefficient (LPCC) feature	
	MFCC feature	[Davis & Mermelstein, 1980]
	I-vector	[Dehak *et al.*, 2010; Pan *et al.*, 2017]
Deep learning based	I-vector + DNN	[Ghahabi *et al.*, 2016]
	D-vector	[Variani *et al.*, 2014]
	X-vector	[Snyder *et al.*, 2017, 2018]
End-to-end system	Mimic the i-vector + Probabilistic Linear Discriminant Analysis (PLDA) system	[Rohdin *et al.*, 2018, 2020]
	Raw waveform with SincNet	[Ravanelli & Bengio, 2018b,c]

i-vectors, and followed by proposing a new version of a CNN: the SincNet for the speaker recognition system construction [Ravanelli & Bengio, 2018b,c].

Similarly, the development of speech emotion recognition also started from using knowledge-based features. At first, the frame-wise features (also named as Low Level Descriptor (LLD)s in the emotion recognition research field), like the LPCC, MFCC, F_0, energy, zero-crossing rate and energy slope were used alone or jointly for emotion recognition [Huang & Ma, 2006; Lee *et al.*, 2004; Nakatsu *et al.*, 1999]. In addition to using these frame-wise features, the sentence level features, like the mean, standard deviation and covariance matrices of the frame-wise features, were also calculated for emotion recognition [Kwon *et al.*, 2003; Schuller *et al.*, 2005; Ye *et al.*, 2008].

With the development of deep learning, the DNN based system was designed for learning features [Han *et al.*, 2014; Stuhlsatz *et al.*, 2011]. In 2015, an end-to-end system was proposed for emotion recognition [Fayek *et al.*, 2015] based on the one-second spectrograms. Since 2015, raw waveform started to be used as the input of the CNN [Trigeorgis *et al.*, 2016] or RNN based systems [Huang & Narayanan, 2016; Mirsamadi *et al.*, 2017;

Table 3.2: Speech emotion recognition development history.

Systems	Features or Methods	Related Publications
knowledge-based features	Combination of LPCCs and pitch related features	[Nakatsu *et al.*, 1999]
	MFCCs	[Lee *et al.*, 2004]
	Combination of pitch, energy, zero crossing rate and energy slope	[Huang & Ma, 2006]
	Prosodic features and vocal quality features	[Lugger & Yang, 2007; Zhang, 2008]
	sentence level feature calculation	[Kwon *et al.*, 2003; Schuller *et al.*, 2005; Ye *et al.*, 2008]
Deep learning based	Learn discriminative features with DNN	[Stuhlsatz *et al.*, 2011]
	Extract sentence level features with DNN	[Han *et al.*, 2014]
End-to-end system	DNN on spectrograms	[Fayek *et al.*, 2015]
	CNN on raw signal	[Trigeorgis *et al.*, 2016]
	RNNs with attention mechanism	[Huang & Narayanan, 2016; Mirsamadi *et al.*, 2017; Xie *et al.*, 2019]
	CNN with attention mechanism	[Zhang *et al.*, 2018]

Xie *et al.*, 2019]. Also, the attention mechanism has been demonstrated to be useful in the end-to-end systems for emotion recognition [Huang & Narayanan, 2016; Mirsamadi *et al.*, 2017; Xie *et al.*, 2019; Zhang *et al.*, 2018].

Speech and language-based dementia detection is a relatively new research field. Before 2018, most research was based on pipeline systems with knowledge-based features as the front-end features. Inspired by the developing histories reviewed above, both the pipeline system and deep learning technologies based systems are explored for dementia detection in this thesis.

3.3 Advantages of Speech- and Language-based Automatic Dementia Detection

The clinical diagnostic procedures have been reviewed in Section 2.3. Generally speaking, frequently visiting a clinic can increase both the physical and economic burden on people living with dementia and their families. According to the statistics [Mundt & King, 2003], drop-out rates are estimated at 40% over for 4-year duration research. Compared with the clinical diagnosis, the automatic dementia detection system can be used for home-based assessment testing, which may release the burden of patients on the clinic visiting and ensure more regular test attendance. As a result, regular test attendance can track the development of the disease and ensure early treatment, which that can slow down and maintain mental function.

As reviewed in 2.2, both acoustic and linguistic information embedded in the speech can be affected by dementia. Compared with the existing diagnostic procedures reviewed in Section 2.3, the speech- and language-based automatic dementia detection methods are of interest because it is non-invasive, low-cost and potentially able to aid diagnostic accuracy. Furthermore, as the review in Section 3.1 and Section 3.2, the previous research based on speech and language has been demonstrated to be efficient for automatic dementia detection.

Another advantage of constructing the speech- and language-based automatic dementia detection system is the convenience of data collection. No specific equipment is re-

quired except smartphones or laptops, which makes the automatic system installation possible for most people. In addition, speech and language can also be used as the health sensing modalities for fusing with other modalities like videos in the application. Also, it is possible to collect longitudinal data in an extended care relationship, possibly over tens of years, for the longitudinal disease development analysis. The collected longitudinal data can be used for disease tracking for increasing diagnostic accuracy.

3.4 Drawbacks of Existing Automatic Methods

As mentioned in Section 3.2, the acoustic features used for automatic dementia detection are mostly the traditional acoustic features, like the MFCC and PLP. However, these features cannot describe the symptoms of dementia embedded in the speech when being used alone. For dementia detection, a long list of features is usually combined to describe the symptoms in speech and language. The feature sets proposed in the speech-related fields, like the ComParE and IS10, have been used for speech-based dementia detection. The performance of the combined features depends on the dataset and the task. The feature sets are optimal for one dataset, or the classification task cannot display a stable performance on other datasets or tasks. For home-based and real-world dementia detection, the data collection environment and the accents may vary, so the current knowledge-based features or the designed feature sets cannot ensure high-performing and task-specific across these different conditions and contexts. As shown in the previous research reviewed in Section 3.1 and Section 3.2, there was no unified/standard feature set that could be used for dementia detection across different datasets and classification tasks. To ensure desirable performance, feature designing and selecting for the specific dataset and task is quite time-consuming.

In clinics, doctors will use both the word-level and sentence-level information used simultaneously for the analysis and subsequent diagnosis. The extraction of the word-level and sentence-level features is based on the traditional design for existing methods. As reviewed in Section 3.2.2, the deep learning technologies and end-to-end system have been used for speaker recognition and emotion recognition, but the state-of-the-art deep

learning technologies have not been used for dementia detection. Furthermore, including known clinical knowledge for constructing the systems has not been explored much previously.

Furthermore, though challenging, more diagnostic classes, like FMD and MCI should be considered for making the research more aligned with real clinical practice. The current research has mostly focused on binary classifiers. However, the treatment of FMD, MCI and ND are different though they share similar symptoms, so more diagnostic classes should be considered for clinical practice while designing the automatic system.

3.5 Summary

This chapter reviewed previous research on acoustic-based and linguistic-based automatic dementia detection by utilising machine learning methods. The commonly linguistic-based dementia detection methods were summarised into traditional linguistic features and NLP technologies. When diagnosing, a long list of the traditional linguistic features was extracted from the transcripts for being used as the input of the back-end classifier. With the development of NLP research fields, some NLP technologies, such as word embedding methods, started to be used for dementia detection and gained considerable success.

The acoustic-based dementia detection systems are mostly pipeline systems using knowledge-based features. Though deep learning technologies have not been used for acoustic-based dementia detection before 2018 (when the project in this thesis started), they have shown promising results in related research fields, like emotion recognition and speaker recognition. The development history of speech-based emotion recognition and speaker recognition encouraged us to explore using deep neural networks for constructing speech-based dementia detection systems.

Finally, the advantages of speech- and language-based automatic dementia detection were summarised. Compared with the clinical diagnosis, automatic dementia detection are of interest because it is non-invasive, low-cost and potentially able to aid diagnostic accuracy. Also, it can release the physical and economic burden on people living with

dementia and their families. On the other hand, the existing automatic methods have some drawbacks. The data collection environment and accents may vary for home-based and real-world dementia detection. However, there was no unified/standard feature set that could be used for dementia detection across different datasets and classification tasks. This encourages us to use deep neural networks for training the task-/data-specific system. The next chapter summarises the available dementia-related datasets and their related research.

Chapter 4

Overview of Available Datasets

Contents

4.1 Publicly Available Datasets 55

4.1.1 The DementiaBank Dataset 55

4.1.2 The ADReSS Dataset 59

4.1.3 The ADReSSo Dataset 61

4.2 The IVA Datasets 64

4.2.1 The IVA_{33} Dataset 66

4.2.2 The IVA_{3class} Dataset 67

4.2.3 The IVA_{60} Dataset 67

4.2.4 The $\text{IVA}_{age\&MMSE}$ Dataset 69

4.3 Summary 69

This chapter introduces the available dementia-related datasets and their corresponding research, including the publicly available datasets and the datasets collected by our collaborators in Royal Hallamshire Hospital (Sheffield, UK). Specifically, the publicly available datasets includes the DementiaBank dataset [Becker *et al.*, 1994], the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) dataset [Luz *et al.*, 2020b] and the Alzheimer's Dementia Recognition through Spontaneous Speech *Only* (ADReSSo) dataset [Luz *et al.*, 2021]; The dataset collected by Royal Hallamshire Hospital (Sheffield, UK) is named as the Intelligent Virtual Agent (IVA) dataset, which is a confidential dataset collected by using an animated talking head installed on a laptop [Mirheidari, 2018]. The DementiaBank dataset is the largest publicly available dataset of speech for assessing cognitive impairments. The ADReSS and the ADReSSo are the two datasets shared by the Interspeech Challenge Organisers. Some of the datasets are used in the thesis for speech- and language-based dementia-related studies. The availability information of the dataset, including the information about the recordings, access to manual transcripts, the age and MMSE scores are summarised. In Chapter 3, the literature review relating to automatic speech- and language-based dementia detection was presented; here the research studies using the respective datasets are summarised in order to understand the progress of the research on each dataset. This chapter is structured as follows:

Section 4.1 introduces the information of the publicly available datasets and their related research.

Section 4.2 summarises the information about the IVA dataset, which is collected by the Royal Hallamshire Hospital (Sheffield, UK), in a real clinical setting.

Section 4.3 contains the summary of the chapter.

4.1 Publicly Available Datasets

In this section, three datasets: DementiaBank, ADReSS and ADReSSo are introduced, including the available information about the dataset, like the number of recordings, recording duration, the number of speakers, their age and gender. Also, the related main research studies using the datasets are summaries briefly.

4.1.1 The DementiaBank Dataset

The DementiaBank dataset is the largest publicly available dataset of speech for assessing cognitive impairments. It was collected from 319 individuals between March 1983 and March 1988 [Becker *et al.*, 1994]. The participants were asked to describe a cookie theft picture shown in Figure 2.1. The audio recordings were collected while the participants were describing the picture, and the transcript was also generated manually by transcribing the audio into text. More information about the cookie theft picture description task can be found in Section 2.3.3. In this section, both the dataset information and the existing research for the dataset are presented.

4.1.1.1 Dataset Information

Table 4.1: The participants and recording information of the DementiaBank dataset; #Rec is used to represent "the number of recordings".

Patient Group	**Gender (M:F)**	**Age at Testing; Means[#Rec]**	**Duration (Sec.); Means[#Rec]**	**MMSE; Means[#Rec]**
AD	85:170	71.60±(8.41)[234]	56.14±(23.77)[255]	18.59±(5.10)[234]
HC	79:143	64.17±(7.99)[166]	59.28±(31.40)[222]	29.10±(1.19)[163]
Others	44:30	68.29±(9.31)[62]	54.82±(22.97)[74]	26.97±(3.87)[62]
Sum	208:343	68.46±(9.08)[462]	57.82±(28.15)[551]	23.45±(6.39)[459]

The DementiaBank dataset provides both the audio recordings and the corresponding manual transcripts. The transcripts for the recordings were automatically morphosyntactic analysed, and the Part of Speech (POS) tagging, description of tense, repetition markers and the speaker's identity of the recording segments processed by the CHAT

[MacWhinney, 2014] were also provided. There are 551 samples in total provided by the DementiaBank dataset. Among them, there are 222 samples from 89 HCs and 255 samples from 168 people living with AD. The rest of the samples are collected from people living with other causes of dementia and those converted from the MCI to AD during the five years of data collection.

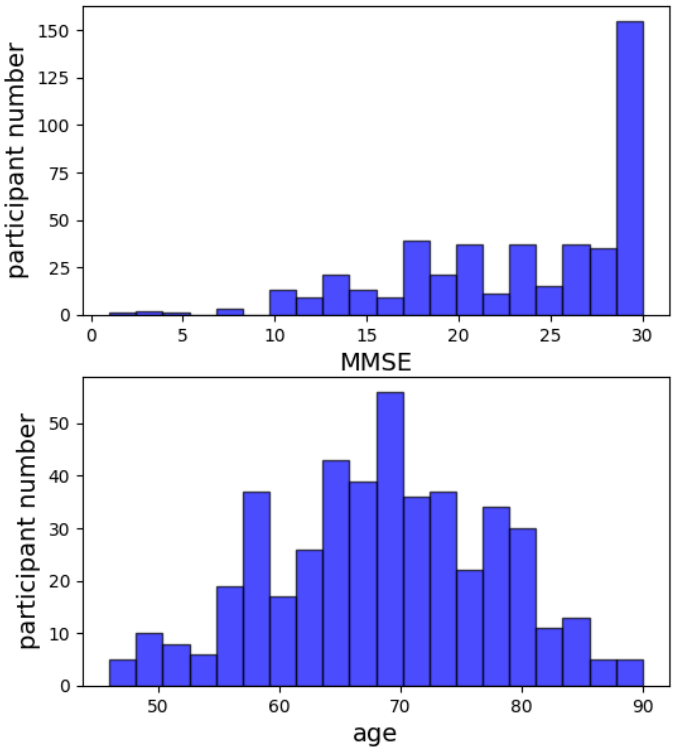


Figure 4.1: The distribution of the MMSE scores and age of speakers in the DementiaBank dataset.

The information for the dataset is summarised in Table 4.1. As shown, the average recording duration for the HC group is 59.28 seconds, and for the AD group is 56.14 seconds. For the 551 recordings, not every audio recording has the age and MMSE information. In total, 462 recordings are tagged with the age information, and 459 recordings include both the age and MMSE information. Specifically, for the AD group, there are 85 males and 170 females that have the gender information, and in total, 234 recordings have both the age and MMSE information. The average age is 71.6, and the average MMSE

is 18.50. For the HC group, there are 79 males and 143 females. 163 out of the 222 recordings have the MMSE information, and the averaged MMSE is 29.10. The average age for the 166 HCs is 64.17.

The distribution of the age and MMSE is shown in Figure 4.1. Only the information from the HC and AD groups are calculated for plotting the Figure 4.1. As shown, the age range is from 46 to 90, and the MMSE value is between 1 and 30. The age distribution is close to a Gaussian distribution.

4.1.1.2 Previous Research on the DementiaBank Dataset

The effect of dementia on the speech and language abilities has been reviewed in Section 2.2. The research studies using the DementiaBank dataset is summarised into two categories according to the information used for disease detection: linguistic-based [Fraser *et al.*, 2016, 2019; Fritsch *et al.*, 2018; Mittal *et al.*, 2020; Orimaye *et al.*, 2017, 2018; Pompili *et al.*, 2020a; Sarawgi *et al.*, 2020; Wang *et al.*, 2020; Yancheva *et al.*, 2015] and acoustic-based [Fraser *et al.*, 2016; Haider *et al.*, 2019; Hernández-Domínguez *et al.*, 2018; Li *et al.*, 2021; Luz, 2017; Sarawgi *et al.*, 2020; Triapthi *et al.*, 2021; Yancheva *et al.*, 2015]. Of these, [Sarawgi *et al.*, 2020; Yancheva *et al.*, 2015] have used both acoustic and linguistic information.

Before 2018, most of the research is based on a large number of features [Fraser *et al.*, 2016; Hernández-Domínguez *et al.*, 2018; Luz, 2017; Orimaye *et al.*, 2017; Yancheva *et al.*, 2015]. For example, in Yancheva *et al.* [2015], a set of 477-dimension features is extracted for the MMSE estimation. The feature set includes two major types of features: linguistic features like the POS tagging and word frequency, and acoustic features like the MFCC. Fraser *et al.* [2016] proposed to use 370-dimension features for describing the symptoms embedded in the speech and language. Similarly, Orimaye *et al.* [2017] proposed to use a 16,926 dimension linguistic feature set. In addition, [Yancheva & Rudzicz, 2016] word vector started to be used for representing the linguistic information embedded in the word.

After 2018, the technologies or features proposed for other related research fields, like the language model [Fritsch *et al.*, 2018] and the ComParE set [Haider *et al.*, 2019; Sarawgi *et al.*, 2020] are used for dementia detection. Also, neural networks, like the LSTM with

Table 4.2: Previous research on the DementiaBank dataset using acoustic or linguistic information.

Year	Study	Modality	Accuracy(%)/MAE	Fully Automatic
2015	[Yancheva *et al.*, 2015]	Fusion	3.83 (MAE)	Yes
2016	[Fraser *et al.*, 2016]	Fusion	81.9	Manual transcript
2016	[Yancheva & Rudzicz, 2016]	Fusion	80.0	Manual transcript
2017	[Orimaye *et al.*, 2017]	Linguistic	Unknown	Statistical analysis
2017	[Luz, 2017]	Acoustic	68.0	Yes
2018	[Hernández-Domínguez *et al.*, 2018]	Acoustic	62.0	Yes
2018	[Masrani, 2018]	Fusion	82.2	Manual transcript
2018	[Mirheidari, 2018]	Linguistic	75.6 62.3	Manual transcript Yes
2018	[Fritsch *et al.*, 2018]	Linguistic	85.6	Manual transcript
2019	[Haider *et al.*, 2019]	Acoustic	78.7	Yes
2020	[Triapthi *et al.*, 2021]	Acoustic	87.9	Yes
2020	[Sarawgi *et al.*, 2020]	Fusion	88.0	Manual transcript
2020	[Pompili *et al.*, 2020a]	Fusion Fusion	85.5±(2.9) 79.7±(3.5)	Manual transcript Automatic transcript
2020	[Mittal *et al.*, 2020]	Acoustic Linguistic Fusion	68.6 83.4 75.7 85.3 78.8	Yes Manual transcript Automatic transcript Manual transcript Automatic transcript

attention mechanism classifier [Li *et al.*, 2021] and the BERT embedding [Mittal *et al.*, 2020] are demonstrated to be useful for the information classification and modelling.

The performance and information of the research mentioned above are selected and shown in Table 4.2. For the proposed linguistic-based methods, both the manual transcripts [Fritsch *et al.*, 2018; Kong *et al.*, 2019; Masrani, 2018; Sarawgi *et al.*, 2020; Yancheva & Rudzicz, 2016] and automatic transcripts generated by the ASR system [Li *et al.*, 2021; Mittal *et al.*, 2020; Pompili *et al.*, 2020a] are used for the linguistic feature extraction. However, the results are not comparable to other research due to the mismatch of the experimental setting but are comparable in the same paper. The selected research before and after the studies reported in this thesis on the DementiaBank dataset are presented in Table 4.2. The conclusions are summarised as follow:

- As shown, before 2018, Masrani [2018] achieved the best result (82.2%) by using features extracted from manual transcripts and audio recordings. In this period, most of the research is based on manual transcripts when extracting the linguistic information, thereby avoiding the issues arising when needing to extract linguistic information from erroneous ASR transcripts.

- It is clear that a gap exists between the performance when using a manual versus an automatic transcript. For example, in Mittal *et al.* [2020] a 75.7% accuracy was achieved on the automatic transcripts, and a 83.4% accuracy was achieved on the manual transcripts. Similarly, the accuracy is 85.5±(2.9) and 79.7.±(3.5) respectively on the manual and automatic transcripts in Pompili *et al.* [2020a].

- Compared with the performance of the acoustic-based dementia detection systems, the linguistic-based dementia detection systems often perform better, such as the results shown in Li *et al.* [2021] and Mittal *et al.* [2020].

4.1.2 The ADReSS Dataset

The ADReSS dataset [Luz *et al.*, 2020b] was constructed and shared as part of the Interspeech-2020 Challenge. The aim of this challenge was to provide researchers with

a benchmark dataset for linguistic- and acoustic-based dementia detection tasks. In this section, both the dataset information and the papers published for the challenge at the Interspeech-2020 conference are reviewed.

4.1.2.1 Dataset Information

The shared dataset provides both the acoustic recordings and the corresponding manual transcripts. The data is a subset of the DementiaBank dataset and is constructed to consider the balance of gender and age. The audio recordings from the DementiaBank dataset had noise reduction applied first and were then normalised across all the speech segments chunked by the signal energy. Thus, the shared dataset includes both the pre-processed complete recordings and the processed segmented recordings. In total, the recordings were segmented into 1955 speech segments from 78 people living with AD and 2122 speech segments from 78 living without AD. The ADReSS Challenge defines two tasks: the binary classification task, which aims to classify the AD and non-AD, and the MMSE prediction task, aiming to predict the MMSE score by analysing the acoustic recordings and manual transcripts. The detailed information of the training set and test set is shown in Table 4.3.

Table 4.3: The participant and recording information of the ADReSS dataset.

Subset	**Patient Group**	**Gender (M:F)**	**Age at Testing; Means[#Rec]**	**Duration; Means[#Rec]**	**MMSE; Means[#Rec]**
Training	AD	24:30	66.91±(6.52)[54]	82.24±(43.21)[54]	17.17±(5.40)[54]
	HC	24:30	66.21±(6.41)[54]	61.46±(20.76)[54]	29.11±(0.98)[54]
Test	AD	11:13	66.13±(7.28)[24]	90.47±(51.75)[24]	19.46±(5.27)[24]
	HC	11:13	66.13±(6.94)[24]	74.55±(31.51)[24]	28.79±(1.47)[24]

4.1.2.2 Previous Research on the ADReSS Dataset

The best acoustic-only classification results reported by the organiser of the challenge [Luz *et al.*, 2020a] is 62.00% accuracy on the classification task and 7.28 Root Mean Square Error (RMSE) on the regression task, respectively, by using the acoustic feature sets extracted by the OpenSMILE v2.1 toolkit [Eyben *et al.*, 2010]. The best linguistic-only

results reported by [Luz *et al.*, 2020a] is 75.00% accuracy on the classification task and 4.38 RMSE on the regression task, respectively. These results are the baseline results of the Challenge.

In the 13 papers published in Interspeech-2020, seven papers used the BERT for modelling the linguistic information [Balagopalan *et al.*, 2020; Farzana & Parde, 2020; Koo *et al.*, 2020; Pompili *et al.*, 2020b; Searle *et al.*, 2020; Syed *et al.*, 2020; Yuan *et al.*, 2020]. The best linguistic-only accuracy achieved was 89.60% on the test set [Yuan *et al.*, 2020], which used the pause and disfluency annotation as the extra information of the manual transcripts while using the BERT for linguistic information modelling.

In comparison, the performances of the acoustic-only based systems are not as good as the linguistic-only based systems, which is consistent with the research reviewed in Section 4.1.1 on the DementiaBank dataset. The best acoustic-only classification accuracy (72.92%) and regression RMSE (5.08) is achieved by using the pre-trained VGGish system [Hershey *et al.*, 2017] for acoustic information extraction [Koo *et al.*, 2020]. Compared with the other proposed acoustic-only methods, the pre-trained network can provide more powerful features than extracting features from scratch, like in Cummins *et al.* [2020]; Edwards *et al.* [2020]. As the acoustic features, speaker embeddings like the i-vector [Dehak *et al.*, 2010] and x-vector [Snyder *et al.*, 2018] have been used for dementia detection in previous research on other datasets and demonstrated to be efficient [López *et al.*, 2019; Weiner & Schultz, 2018; Zargarbashi & Babaali, 2019], but the result in Pappagari *et al.* [2020]; Pompili *et al.* [2020b] did not outperform the Challenge's baseline performance [Luz *et al.*, 2020a].

4.1.3 The ADReSSo Dataset

The ADReSSo dataset [Luz *et al.*, 2021] is associated with the Interspeech-2021 Challenge, but unlike the ADReSS dataset only provides the acoustic recordings without the corresponding manual transcripts. Both the dataset information and the papers accepted by the Interspeech-2021 conference are reviewed below.

4.1.3.1 Dataset Information

The Interspeech-2021 ADReSSo Challenge defines three tasks: the binary classification task (HC vs. AD), the MMSE estimation task, and the disease progression detection task, which aims at predicting the change of cognitive decline longitudinally. For the disease progression detection task, the criteria of *decline* or *no-decline* is defined by the difference of the MMSE score collected in the two-year period. Any recordings with the MMSE difference of no smaller than five are classified as decline. The recordings shared for the binary classification task and the MMSE estimation task were collected from the individuals when they were describing the cookie theft picture as shown in Figure 2.1.

All the collected audio recordings had noise removal applied first and were then normalised across all the speech segments to control for volume variation resulting from recording conditions such as microphone placement. The dataset also provides the segmentation information with speaker's identifications (participant or interviewer). For the three tasks, the Geneva minimalistic acoustic parameter set (eGeMAPS) feature set [Eyben *et al.*, 2015] was used as the acoustic feature set reported by the organisers of the challenge [Luz *et al.*, 2020a]. The Google Cloud-based Speech Recogniser was first used to transcribe the audio recordings into transcripts to extract the linguistic features. Then the CLAN [MacWhinney, 2017] was used for the transcript processing and linguistic-based feature extraction.

Table 4.4: The participant and recording information of the ADReSSo data to be used for the binary classification task and MMSE estimation tasks.

Subset	**Patient Group**	**Duration; Means[#Rec]**	**MMSE; Means[#Rec]**
Training	AD	87.61±(46.31)[87]	17.44±(5.30)[87]
	HC	68.76±(25.04)[79]	28.99±(1.14)[79]
Test	AD	79.42±(36.26)[35]	18.86±(5.72)[35]
	HC	66.35±(28.18)[36]	28.86±(1.25)[36]

Table 4.4 summarises the information of the audio recordings provided for the classification task and the MMSE disease progression detection task. As shown, the training set includes 87 recordings from AD and 79 recordings from HC. The average length of

the audio recording is 87.61 seconds from AD , and 68.76 seconds from HC. The test set includes 35 AD recordings and 36 HC recordings.

Table 4.5: The participant information and recording information of the ADReSSo data to be used for the disease progression detection task.

Subset	Patient Group	Duration; Means [#Rec]
Training	Decline	147.15±(49.65) [15]
	No-decline	142.94±(18.02) [58]
Test	Decline	146.83±24.43 [10]
	No-decline	141.05±29.43 [22]

Table 4.5 summarises the information of the audio recordings provided for the disease progression detection task. The training set shows that 15 recordings correspond to the declined, and 58 recordings correspond to the no-declined. The test set includes 10 recordings from the decline group and 22 recordings from the no-decline group.

4.1.3.2 Previous Research on the ADReSSo dataset

For the classification task, the baseline classification accuracy reported by Luz *et al.* [2021] on the test set is 64.79% and 77.46%, respectively, by using acoustic and linguistic features. For the MMSE regression task, the best baseline result is 6.09 and 5.28 RMSE on acoustic and linguistic features achieved by the support vector regression (SVR). The F-score (F-measure) of the progression prediction task is 53.62% and 66.67%, respectively, based on acoustic and linguistic information on the test set.

In the Interspeech 2021 ADReSSo special session, in addition to the paper proposed by the challenge organiser, eleven papers were accepted. The best result for the classification task is 84.51% accuracy reported by Pan *et al.* [2021a]; Pappagari *et al.* [2021]; Syed *et al.* [2021] (joint winners). Among the Interspeech accepted papers, five out of eleven evaluated their methods on the MMSE regression task and two out of eleven evaluated their methods on the disease progression detection task. Pappagari *et al.* [2021] achieved the best performance (3.85 RMSE) on the MMSE estimation. The best result on the disease progression detection task is 70.91% accuracy achieved by [Zhu *et al.*, 2021].

Among all the accepted papers, nine out of eleven papers used the pre-trained ASR system for generating the automatic transcripts, which were then used for extracting the linguistic features [Chen *et al.*, 2021; Pan *et al.*, 2021a; Pappagari *et al.*, 2021; Pérez-Toro *et al.*, 2021; Qiao *et al.*, 2021; Rohanian *et al.*, 2021; Syed *et al.*, 2021; Wang *et al.*, 2021; Zhu *et al.*, 2021]. Among the nine papers, only Wang *et al.* [2021] extracted the linguistic features without using the BERT. For the other eight papers, only Zhu *et al.* [2021] integrated the acoustic feature extraction and linguistic feature extraction with one end-to-end system, rather than using two separate systems.

For extracting the acoustic features, some popular traditional features or feature sets were used, like the IS10 [Eyben *et al.*, 2013], VGGish [Hershey *et al.*, 2017] and x-vector [Snyder *et al.*, 2017]), MFCC, ComParE [Eyben *et al.*, 2013], The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [Eyben *et al.*, 2015], eGeMAPS [Eyben *et al.*, 2015], prosody features, disfluency features and COVAREP [Degottex *et al.*, 2014]. In addition, the pre-trained wav2vec 2.0 system [Baevski *et al.*, 2020] was used by four papers [Balagopalan & Novikova, 2021; Gauder *et al.*, 2021; Pan *et al.*, 2021a; Zhu *et al.*, 2021] for extracting the acoustic embedding.

No experimental work using the ADReSS nor the ADReSSo dataset was reported in this thesis, but challenge work was published in Cummins *et al.* [2020] (ADReSS) and Pan *et al.* [2021a] (joint winner in ADReSSo).

4.2 The IVA Datasets

The IVA dataset has been collected by the Royal Hallamshire Hospital (Sheffield, UK) in a real clinical setting during the summers since 2016 [Mirheidari *et al.*, 2019b]. A *Digital Doctor* (or *Intelligent Virtual Agent*), which is an animated talking head displayed on a laptop screen, asks a series of conversational questions and administers a series of verbal fluency tests designed to match the questions asked by the neurologists in a clinical environment. The conversational questions asked by the IVA varied slightly in the different years, but all for the same target: examining the participant's description ability, short-term and long-term memory. The verbal fluency tests are the standard screening tests

for the AD diagnosis known as "fluency semantic" (naming from a category e.g. animal or fruit) and "fluency phonemic" tests (naming words beginning with a letter e.g. "P"). More details were given in Section 3.1.

Table 4.6: The participant and recording information of the IVA dataset. M:F:U represents the number of speakers for male, female and un-known; #Rec represents the number of recordings; Rec. Dur. represents the recording duration.

Patient Group	**Gender (M:F:U)**	**Age; [#Rec]**	**Full Rec. Dur.; [#Rec]**	**Cookie Theft Rec. Dur.; [#Rec]**	**MMSE; [#Rec]**
FMD	7:10:0	55.1±(6.39)[17]	617.30±(305.24)[17]	41.17±(18.03)[10]	27.50±(0.50)[5]
ND	17:12:4	69.2±(6.57)[28]	749.12±(525.77)[33]	68.17±(40.09)[27]	22.91±(3.90)[12]
MCI	20:12:4	62.9±(8.35)[27]	548.90±(268.00)[36]	53.33±(37.41)[29]	27.57±(1.18)[18]
HC	6:10:46	70.6±(8.44)[16]	536.62±(161.84)[62]	75.12±(33.61)[62]	unknown [0]
In total	50:44:54	64.5±(9.63)[88]	606.09±(322.96)[148]	66.06±(36.81)[128]	25.54±(4.02)[35]

Table 4.7: The recording information for the IVA dataset collected in different years.

	FMD	ND	MCI	HC	Other	Total
IVA2016	7	6	6	0	5	24
IVA2017	3	7	10	0	1	21
IVA2018	6	14	11	28	2	61
IVA2019	1	6	9	34	14	64
Total number of recordings	17	33	36	62	22	170
Total number of speakers	16	27	25	40	22	130

From 2016 to 2019, a total number of 170 recordings have been collected. Among all the recordings, only the recordings from the HC, ND, MCI and FMD with diagnostic label and manual transcripts are considered in our studies. Therefore, only the information of the collected data from the four classes is summarised in Table 4.6. As shown, there are 148 recordings in total for the four classes. Not all the recordings include the age, MMSE and gender information. Specifically, 35 out of 148 recordings include the estimated MMSE scores, and 128 out of 148 recordings include the cookie theft picture description recordings. The age range is between 41 and 86, and 88 out of 148 recordings include the age information.

These recordings were part of a larger study, initially focusing on the manual analysis of conversations between neurologists and patients [Elsey *et al.*, 2015] and later in automating this analysis and comparing neurologist-led and IVA-led conversations Mirheidari [2018]; Walker *et al.* [2020]. The statistical test result shown in Chapter 7 of Mirheidari [2018] shows that no significant difference exists between the accuracy of the classifier obtained from the neurologist-led and the IVA-led conversations. Furthermore, the research carried out in Walker *et al.* [2020] shows the computer's questions are invariant and more suitable for comparative diagnostic purposes.

The IVA dataset collected from 2016 to 2019 for each year is shown in Table 4.7. There are three versions of the question list. The question lists are designed according to those questions asked in a real assessment situation by the clinicians. More details about the question list can be found in Mirheidari *et al.* [2019b]. As shown in the table, 170 recordings were collected from 130 speakers in total. In addition to the 148 recordings collected from the HC, MCI, FMD and ND groups, 22 collected recordings were collected from other types of disease, like psychiatric or depression or anxiety related (named as "Other" in the table). In 2016 and 2017, no recordings were collected from the HCs.

The availability of the IVA dataset is restricted, so the research on the IVA dataset is mostly carried out by researchers at the University of Sheffield. In previous research, both the word vector based linguistic information representation [Mirheidari, 2018] and acoustic features designed based on the medical knowledge [Mirheidari *et al.*, 2019a,b] were utilised for analysing the conversation for dementia detection. In this thesis, different splits of the IVA dataset are used in the experimental chapters, which is a result of the data collection progress that was ongoing during the project. In the following, the information about these subsets will be introduced.

4.2.1 The IVA_{33} Dataset

The IVA_{33} dataset is used for the study presented in Chapter 5. It is being used as the extra dataset while using the DementiaBank dataset for training the designed system. For collecting the audio recordings, three versions of the question list have been designed since 2016. The first question list version did not include the picture description task.

The study in Chapter 5 was started in October 2018, and the IVA_{33} dataset was used for the study presented in Chapter 5. For this subset, only the picture description recordings from the HC and AD groups collected between 2017 and October 2018 are selected for the study. For class balancing, the dataset is composed of 17 recordings from the AD and 16 recordings the HC group, which is 33 recordings in total.

4.2.2 The IVA_{3class} Dataset

The IVA_{3class} dataset is composed of the selected recordings collected in 2017, 2018 and the first several months of 2019 from the HC, MCI and ND groups. It is used for testing the efficiency of the designed system in Chapter 6. In Chapter 6 where the research aim is to understand the acoustic difference while doing different classification tasks among HC, MCI and ND. When choosing the recordings from the IVA dataset, the recordings collected in 2016 are not included in the IVA_{3class} dataset considering the vast difference exists between the question lists used before and after 2016. The information of the recordings is shown in Table 4.8. In total, there are 88 recordings collected from 70 speakers. The IVA_{3class} dataset includes 12 hours and 31 minutes in total. This dataset is used for evaluating the study proposed in Chapter 6.

Table 4.8: The information for the IVA_{3class} Dataset.

Diagnostic Category	Number of Speakers	Number of Recordings	Audio Duration
MCI	24	29	3h13min
ND	21	24	4h35min
HC	25	35	4h43min
Total	70	88	12h31min

4.2.3 The IVA_{60} Dataset

The IVA_{60} dataset, composed of the recordings collected in 2016, 2017 and 2018, is used for the research in Chapter 7. The research aim of this chapter is to explore how to design a broader and more clinical relevant set of diagnostic classes. To this end, a balanced set

Table 4.9: The information for the IVA_{60} Dataset.

Diagnostic Category	**Average Age**	**Age Range**	**Audio Duration**
HC	69.5	[55, 86]	2h30min
FMD	54.9	[41, 69]	2h54min
MCI	63.0	[50, 78]	2h39min
ND	67.9	[52, 79]	3h46min
In total	63.8	[41, 86]	11h39min

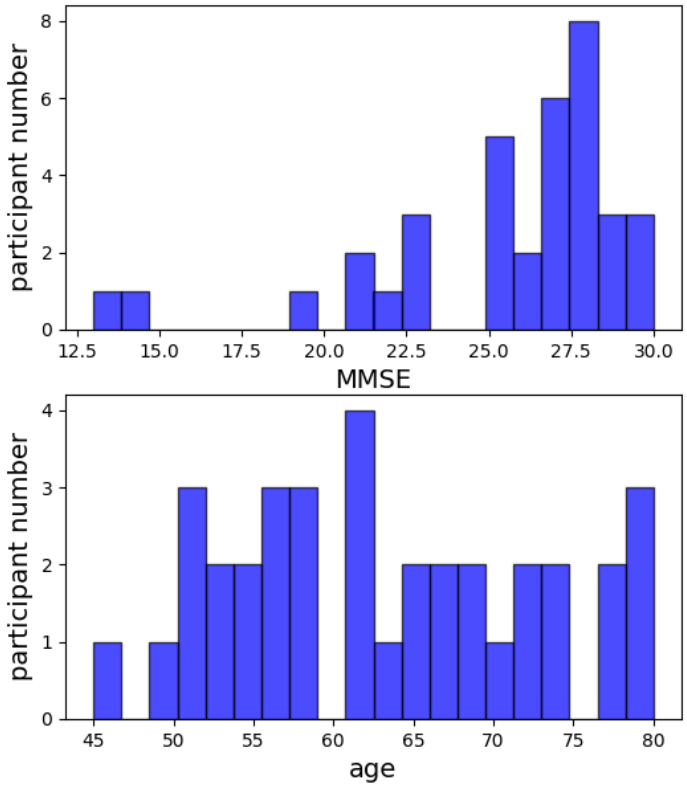


Figure 4.2: The distribution of the MMSE and age in the $IVA_{age\&MMSE}$ dataset.

of 60 interactions (15 FMD, 15 ND, 15 MCI and 15 ND) is chosen for the study in Chapter 8. This dataset has also been used in previous research [Mirheidari *et al.*, 2019b; O'Malley *et al.*, 2021]. The information of the IVA_{60} dataset is summarised in Table 4.9. One out of all the 60 recordings collected, the MCI group lacks the age information. Therefore, the average age and the age range information for the MCI group are calculated with the age information from the rest of the 14 recordings. As shown in the table, the age range is between 41 and 86. The recordings include 11 hours 39 minutes in total.

4.2.4 The $\text{IVA}_{age\&MMSE}$ Dataset

The $\text{IVA}_{age\&MMSE}$ dataset is used to test the confounding factors between age and cognitive decline, and how to make use of age information to improve the cognitive decline estimation accuracy. To this end, the audio recordings corresponding to the cookie theft picture description tagged with the age and MMSE information in the IVA dataset are selected for composing the $\text{IVA}_{age\&MMSE}$ dataset. In total, 34 recordings were selected. The distribution of the age and MMSE information from the 34 recordings is shown in Figure 4.2. As shown in the figure, the age range is between 45 and 80, and the MMSE value is between 13 and 30. This subset is used in Chapter 9.

4.3 Summary

This chapter presented the available dementia-related datasets. Three publicly available datasets were presented, including the DementiaBank dataset, the ADReSS dataset, and the ADReSSo dataset. In addition, the information about the IVA dataset collected by Royal Hallamshire Hospital (Sheffield, UK) is also summarised. These datasets include both the audio recordings and corresponding manual transcripts. For each dataset, the dataset information and the related proposed research were summarised in this section.

DementiaBank dataset is the largest publicly available dataset of speech for assessing cognitive impairments, which was collected from 319 individuals between March 1983 and March 1988. Among the 551 recording samples in the dataset, there are 222 samples from 89 HCs and 255 samples from 168 people living with AD. For classifying the samples

in the DementiaBank dataset, the research studies are summarised into two categories according to the information used for disease detection: linguistic-based and acoustic-based. As summarised, the research before 2018 was mostly based on the pipeline systems that extract a large number of features from the audio recordings or manual transcripts. Since 2018, the automatic transcripts generated by the ASR system started to be used for constructing the automatic linguistic-based system. Though the automatic transcripts include some wrongly transcribed words that can result in ambiguity, the performance of the linguistic-based system with the automatic transcripts as the input performed better than the acoustic-based system with the audio recordings as the input.

The ADReSS dataset, constructed and shared as a part of the Interspeech-2020 Challenge, is a subset of the DementiaBank dataset, which is constructed considering the balance of gender and age. The ADReSS dataset includes 78 recordings tagged with label and MMSE value from HC and AD respectively. Luz *et al.* [2020b] reported two baseline systems proposed by the challenge organiser for the classification and regression tasks. Thirteen papers were accepted by Interspeech-2020 ADReSS special session. Seven out of thirteen papers used BERT for linguistic information modelling. Similarly to the research proposed for the DementiaBank dataset, the linguistic-only system performed better than the acoustic-only system on the ADReSS dataset.

The ADReSSo dataset provided by the Interspeech-2021 Challenge defined three tasks: the binary classification task, the MMSE estimation task and the disease progression detection task. The binary classification task and MMSE estimation task share the same audio recordings without manual transcripts, including 166 recordings from the training set and 71 from the test set. The disease progression detection task includes 73 recordings from the training set and 32 recordings from the test set. For the eleven papers accepted by the Interspeech-2021 special session, nine papers used the pre-trained ASR system for generating the automatic transcripts for extracting linguistic features.

The IVA dataset is a confidential dataset collected by the Royal Hallamshire Hospital (Sheffield, UK) in a real clinical setting during the summers since 2016 using an intelligent virtual agent displayed on a laptop screen. The conversational questions asked by the IVA varied slightly in the different years. The information of IVA and its different subsets used

in this thesis, was introduced. A different subset was used considering the data collection process and research target.

To ensure the research's reliability, publicly available and confidential datasets are used in this thesis. The studies based on the datasets introduced in this chapter will be presented in the next several chapters

Chapter 5

Linguistic Information-based Dementia Detection

Contents

5.1 Introduction 75

5.2 Reasearch Background 76

5.3 Dementia Detection System 78

5.3.1 Word Embedding 79

5.3.2 Word-level Structure 79

5.3.3 Sentence-level Structure 80

5.4 Experimental Setup 81

5.4.1 Datasets 81

5.4.2 Baseline Systems 83

5.4.3 Evaluation Settings 84

5.4.4 Model Configuration 85

5.5 Results and Analysis 86

5.5.1 Experimental Results 86

5.5.2 Result Analysis 89

5.6 Summary 95

In Chapter 2, the effects of dementia on language have been summarised. As mentioned, the effects on language exist at both the word and the sentence levels. For clinical diagnosis, the hierarchical information embedded in the speech has been used for dementia detection. The research in this chapter aims at exploring the answer to the first and second research questions:" how can state-of-the-art deep neural networks be applied for speech- and language-based dementia detection? (**RQ1**)" and "how can the known clinical dementia detection knowledge help in constructing an automatic dementia detection systems and extracting useful features? (**RQ2**)". Picture description is a commonly used assessment method for collecting speech recordings from people living with dementia. Considering the transcripts have a hierarchical structure, this chapter proposes a hierarchical system for modelling both the word and sentence level linguistic information for detecting dementia based on the picture description transcripts. The structure of this chapter is as follow:

Section 5.1 is the introduction to the process of how clinicians diagnosis people who are living with dementia with the transcripts of picture description and how to benefit from the process when constructing an automatic system.

Section 5.2 is the background about the related technologies that can be used for constructing the linguistic-based dementia detection system.

Section 5.3 describes the system proposed for extracting the linguistic information for dementia detection.

Section 5.4 contains the information about the experimental setup of the proposed system.

Section 5.5 summarises the classification results and the corresponding analysis of the proposed system.

Section 5.6 contains the summary of this chapter.

5.1 Introduction

As mentioned in Section 2.3.3, the picture description task is used broadly for the detection of cognitive impairment associated with dementia. The most commonly used picture is the "Cookie Theft" picture, as shown in Figure 2.1. To detect people living with or without dementia using the recordings and transcripts collected from the picture description task, the clinicians manually analyse the audio recordings and the corresponding transcripts, which is relatively time-consuming. Selected automatic dementia detection methods based on the picture description task have been summarised in Section 4.1.1. In this chapter, the dataset collected from the participants that were describing the cookie theft picture is used to test the designed dementia detection system.

As in Section 2.2.2, from a language point of view, people living with dementia show signs of language ability decline at both the word and sentence levels. The word-level signs include a decline in the number of semantic elements [Forbes-McKay & Venneri, 2005], more frequently repairing errors (lemma repairs and reformulation repairs) [McNamara *et al.*, 1992] and decline of vocabulary richness [Le *et al.*, 2011]. At the sentence-level, the signs include a decline in sentence coherence (local coherence and global coherence) and the amount of meaningful information covered in sentences (conciseness) [Mueller *et al.*, 2018]. However, in previous research, no study has managed to extract both the word-level and sentence-level information for dementia detection with deep learning technologies.

The histories of speech-related research fields are reviewed in 3.2.2. With the outstanding performance of deep learning for many domains like speech processing and NLP, researchers have started to introduce this technology into automatic dementia detection. As shown in Chapter 3, for the same database, deep neural networks have the potential to achieve a better result than the pipeline systems [Fritsch *et al.*, 2018; Warnita *et al.*, 2018]. A clinician carries out language-based dementia detection by considering both the word and sentence levels information. In addition, while diagnosing, clinicians take the words and the sentences into account separately. However, so far, automatic dementia detection methods have not taken the *hierarchical structure* (word-level and sentence-level) of the transcript into account when using deep neural networks. Also, the practice of weighting

the words and sentence has not been incorporated in previously proposed approaches. This chapter aims to construct a system for hierarchically modelling and weighting of the transcripts at both the word and sentence levels to mimic the clinicians' diagnostic procedures with the automatic end-to-end system.

5.2 Reasearch Background

ticalIn the previous research, hierarchical systems were proposed for written essay classification and scoring by modelling the word and sentence levels information simultaneously [Yang *et al.*, 2016]. Hierarchical systems are generally composed of two sub-systems: the word and sentence level systems. The word-level system is designed to extract the sentence representation from the words, and the sentence-level system is designed to extract the transcript representation from the sentences. Though the hierarchical system was proposed for written essays originally, it has also been used successfully to automatic transcripts [Tseng *et al.*, 2016]. However, compared with the written essays, transcripts generated from people living with dementia include more grammatical errors and repairing errors (lemma repairs and reformulation repairs) [McNamara *et al.*, 1992], making the information modelling more challenging than the written essays. Furthermore, the unclear pronunciation in the audio collected from the AD group increase the difficulties of the ASR transcribing, resulting in more word errors, which can further increase the feature extraction difficulties.

The DementiaBank dataset is the largest publicly available dataset of speech for assessing cognitive impairments. The audio recordings were collected from people when they described the cookie theft picture (as shown in Figure 2.1). More details about the dataset can be found in Section 4.1.1. Two transcript examples from the HC and AD are shown in the next two paragraphs.

A transcript from HC: *The scene is in the in the kitchen . The mother is wiping dishes and the water is running on the floor . A child is trying to get a boy is trying to get cookies out of a jar and he's about to tip over on a stool . Uh the little girl is reacting to his falling . Uh it seems to be summer out . The window is open . The*

curtains are blowing . It must be a gentle breeze . There's grass outside in the garden . Uh mother's finished certain of the the dishes . Kitchen's very tidy . The mother seems to have nothing in the house to eat except cookies in the cookie jar . Uh the children look to be almost about the same size . Perhaps they're twins . They're dressed for summer warm weather . Um you want more . The mother's in a short sleeve dress . I'll have to say it's warm .

***A transcript from AD**: There's a little girl reaching for the cookie jar and she can't reach it apparently . And the the young man is helping her . He's on a stool and he's reaching for the cookie jar . And uh and the lady is drying dishes . And the water is pouring out of the sink for some reason . There's uh some plates on the on the counter . And she's drying a dish . I may have said that . And the young man is going to fall off the stool . I guess maybe I said that too . And they're reaching for the cookie jar . And the and the sink is overflowing . I guess I might have said that too . And I guess that's about uh all the salient things of it . I can see .*

The decline is shown at the word level, as the nouns are not as specific and accurate in the AD transcript as in the HC transcript. For example, people living with AD may use "the woman" or "the lady" rather than "the mother" while describing the picture. At the sentence-level, there are more vague sentences, like "I may have said that" in the AD transcripts than in the HC transcripts. Also, the coherence between the sentences is less tight. As shown, the description from the HC group starts with a summary sentence: *The scene is in the in the kitchen*, and then describes the details methodically. In comparison, the description from the AD group starts with a girl (which actually is a boy) who is stealing cookies, followed by describing the mother but comes back to describing the boy and girl again. The logic is confusing, and the description is wordy.

For dementia detection related research, except for the DementiaBank dataset, the other databases are mostly self-collected and not shareable due to ethical constraints. However, training a good deep learning system often requires large amounts of data. Therefore, in this chapter, the IVA_{33} dataset (more information about the dataset can be found in Chapter 4.2.1) is also used as the extra data for the system training. The goal is to explore whether enlarging the training set can benefit the system's performance on

the test set, even though an apparent mismatch exists between the two datasets.

5.3 Dementia Detection System

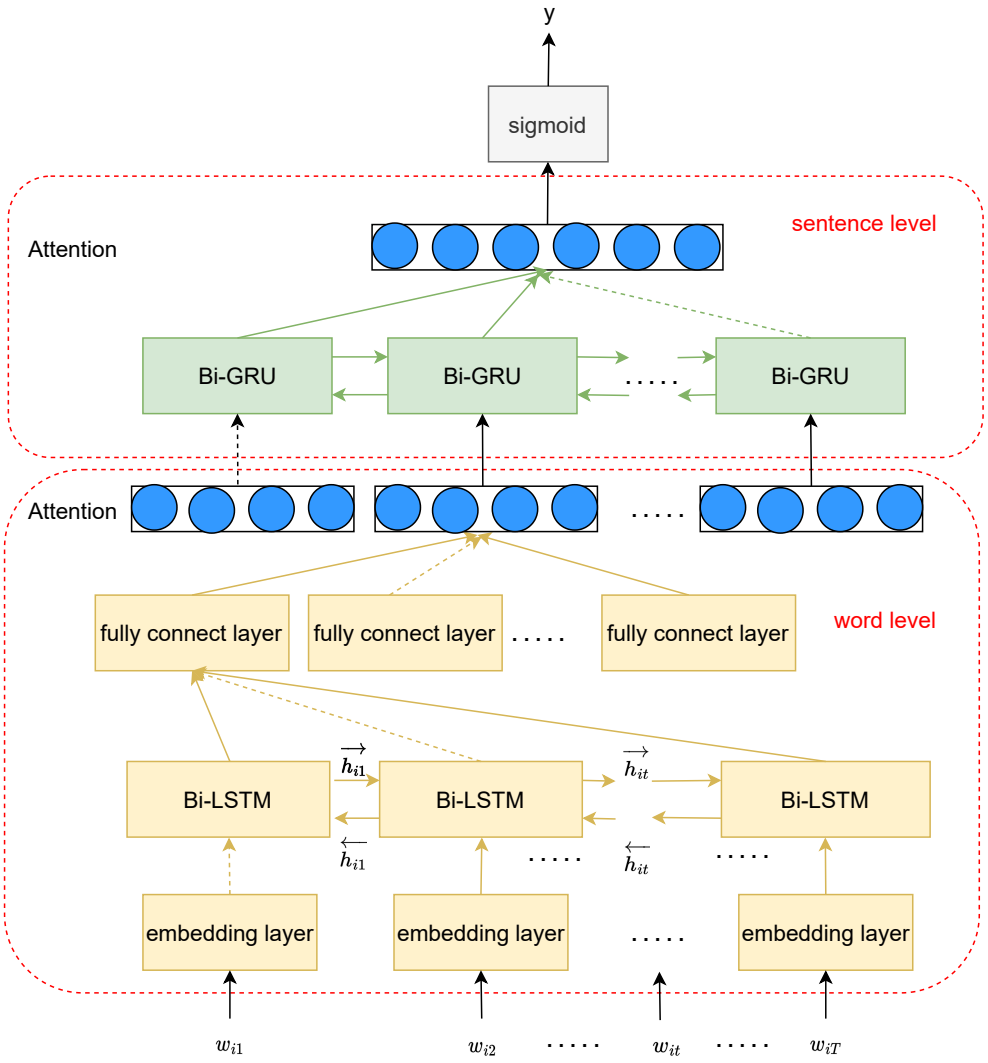


Figure 5.1: The structure of the proposed hierarchical attention based system (HBANN) for dementia detection with the transcripts as the input.

In this section, a hierarchical attention system designed for linguistic-based dementia detection is introduced. The system structure is shown in Figure 5.1 and named as the hierarchical bidirectional attention neural network (HBANN). The dashed line with an arrow is used to represent the dropout. The ellipsis "..." is used to represent the units that have not been drawn. This section describes how to represent a transcript in a

vector and then estimate its diagnostic class with an end-to-end system. The system is introduced in three parts: word embedding, word-level and sentence-level structures.

5.3.1 Word Embedding

First of all, the words need to be transformed into high-dimension vectors. This process is called word embedding, and the output is called word vectors. The benefit of this operation is to capture the semantic information about the words as represented in the sentences.

For dementia detection, word embedding has been proposed to be used for converting transcripts into vectors for dementia detection [Mirheidari, 2018]. As reviewed in Section 3.1.2, the word2vec [Mikolov *et al.*, 2013] and Global Vectors embedding matrix for Word Representation (Glove) [Pennington *et al.*, 2014] were proposed as the unsupervised methods, while the embedding layer used in the neural network is the supervised methods for mapping the words into real-valued vector representations [Goldberg, 2017]. In the proposed system, a trainable embedding layer is used for word embedding. Considering the limitation of the available dataset, a pre-trained word embedding matrix is used for initialising the parameters in the embedding layer. More details can be found in 5.4.4.

5.3.2 Word-level Structure

It has been shown in Section 2.2.2 that people living with dementia tend to phrase things using 'vague' or ineffective information, repeat words and phrases more frequently and decline the vocabulary richness. Therefore, in our proposed system, a Bidirectional LSTM (BLSTM) is applied to the word vectors to extract word-level information from the variable-length sequence.

As in 3.2.2, the BLSTM can get the representation of words and their surrounding information from both the forward and backward directions. In the system, the representation h_{it} is the t_{th} output of the LSTM layer, which is achieved by adding the vector from the forward LSTM $\overleftarrow{h_{it}}$ and the backward LSTM $\overrightarrow{h_{it}}$. W_e is used for representing the weights in the hidden layer and w_{it} is used for representing the tth word ($t \in [0, T]$) of

the ith sentence in the transcript.

$$\begin{aligned} \overleftarrow{h_{it}} &= \overleftarrow{\text{LSTM}}(W_e w_{it}, h_{it+1}) \\ \overrightarrow{h_{it}} &= \overrightarrow{\text{LSTM}}(W_e w_{it}, h_{it-1}) \end{aligned} \tag{5.1}$$

For obtaining a vector representation h_{it} for each word w_{it}, the backward hidden state $\overleftarrow{h_{it}}$ and forward hidden state $\overrightarrow{h_{it}}$ is added together: $h_{it} = \overrightarrow{h_{it}} + \overleftarrow{h_{it}}$. Then, a fully-connected layer with the ReLU activation function [Agarap, 2018] is applied in the following:

$$d_{it} = \text{ReLU}(W_d h_{it} + b_d) \tag{5.2}$$

where W_d and b_d are used to represent the weights and bias of the fully-connected layer. h_{it} is the output of the BLSTM layer. To model each word's importance, an attention mechanism is used, followed by a fully connected layer as described in Yang *et al.* [2016]. The attention mechanism is defined as in Equation 5.3.

Specifically,

$$\begin{aligned} u_{it} &= tanh(W_w d_{it} + b_w) \\ \alpha_{it} &= \frac{exp(u_{it}^T u_w)}{\sum_t exp(u_{it}^T u_w)} \\ s_i &= \sum_t \alpha_{it} h_{it} \end{aligned} \tag{5.3}$$

where u_{it} is the representation of the hidden layer. W_w and b_w are the embedding matrix and bias of the fully-connected layer. The word importance is measured by calculating the similarity of u_{it} with a word-level vector u_w, which is initialised randomly and used as a high-level representation of a fixed query over the words like in a memory network [Sukhbaatar *et al.*, 2015]. Finally, a sentence representation s_i is calculated by the weighted sum of the words in the ith sentence.

5.3.3 Sentence-level Structure

After obtaining the sentence representations (vectors) from the word-level structure, a bidirectional GRU layer is applied to each sentence vector. According to Chung *et al.*

[2014], whether to use a LSTM or a RNN depends on the dataset and corresponding tasks. Likewise, the decision is based on the experimental performance. For getting the transcript representation, in the proposed system, GRUs are used for modelling the sentence vector s_i.

$$\begin{aligned} \overleftarrow{h_i} &= \overleftarrow{\text{GRU}}(s_i, h_{i+1}) \\ \overrightarrow{h_i} &= \overrightarrow{\text{GRU}}(s_i, h_{i-1}) \end{aligned} \tag{5.4}$$

Similarly as in Equation 5.1, the backward hidden state $\overleftarrow{h_i}$ and forward hidden state $\overrightarrow{h_i}$ is added together for getting the sentence representation.

Then an attention layer is applied:

$$\begin{aligned} u_i &= \tanh(W_s h_i + b_s) \\ \alpha_i &= \frac{exp(u_i^T u_s)}{\sum_i exp(u_i^T u_s)} \\ v &= \sum_i \alpha_i h_i \end{aligned} \tag{5.5}$$

Where W_c and b_c is the weight vector and bias vector, respectively. u_s, the sentence-level matrix, is also initialised randomly as the word-level matrix. Finally, one fully-connected layer with a sigmoid function is applied for classification.

5.4 Experimental Setup

In this section, the experimental settings are introduced, including the used datasets, the baseline systems, the evaluation settings, and the configuration of the proposed system.

5.4.1 Datasets

In this chapter, the DementiaBank and the IVA_{33} datasets are used for exploring the performance of the proposed system. Both the information of the two datasets and the automatic transcripts generated from the ASR system is introduced.

5.4.1.1 Datasets Information

The HBANN system is designed for binary classification. For the DementiaBank dataset, only the recordings collected from the AD or HC groups are used in this chapter. In total, 222 samples from 89 HCs and 255 from 168 people living with AD are selected from the original 551 transcripts. For the IVA_{33} dataset, there are 33 transcripts in total including 17 transcripts from the HC group and 16 transcripts from the AD group. The dataset information is shown in Table 5.4.1.1. More detailed information can be found in Section 4.2.

In the experiment part, the IVA_{33} dataset is used as the extra data for training the system. However, as shown in the table, a mismatch exists on the average utterance length between the two datasets. Also, the speaker accents and data collection environment are different for the two datasets. Section 5.5 explores whether the IVA_{33} dataset can benefit the dementia detection system when being used as the extra training set.

Table 5.1: Information about the DementiaBank dataset and the IVA_{33} dataset.

#Dataset	Length	#Utterance	#Speaker	Utterance Length
DementiaBank(477)	7h40 mins	6124	257	4.50 seconds
IVA_{33}(33)	40 mins	264	33	9.04 seconds

5.4.1.2 Automatic Transcript Generation

For automatic transcription generation, the Kaldi [Povey *et al.*, 2011] toolkit hybrid time delay neural network (TDNN)-LSTM recipe is used for training the ASRs. For language models, the in-domain 3/4 grams is smoothed with the Kneser-Ney (KN) or Good Turing (GT) smoothing [Chen & Goodman, 1999]. Note that an additional 64 hours worth of conversational data is added (the Hallamshire dataset [Mirheidari *et al.*, 2019a]; 64 hours of conversational recordings between doctors and patients) to boost the acoustic model of the ASRs. Finally, automatic transcripts with a Word Error Rate (WER) of 41.6% on the DementiaBank dataset and 33.8% on the IVA_{33} dataset are used in this chapter (see Mirheidari [2018] for more details). The WER is relatively high, resulting from the dataset's poor quality and unclear pronunciation from AD.

To add punctuation in the ASR transcripts, a toolkit shared in github is used. It predicts placement and type of punctuation by using a BLSTM network with an attention layer. Further details can be found in Tilk & Alumäe [2016].

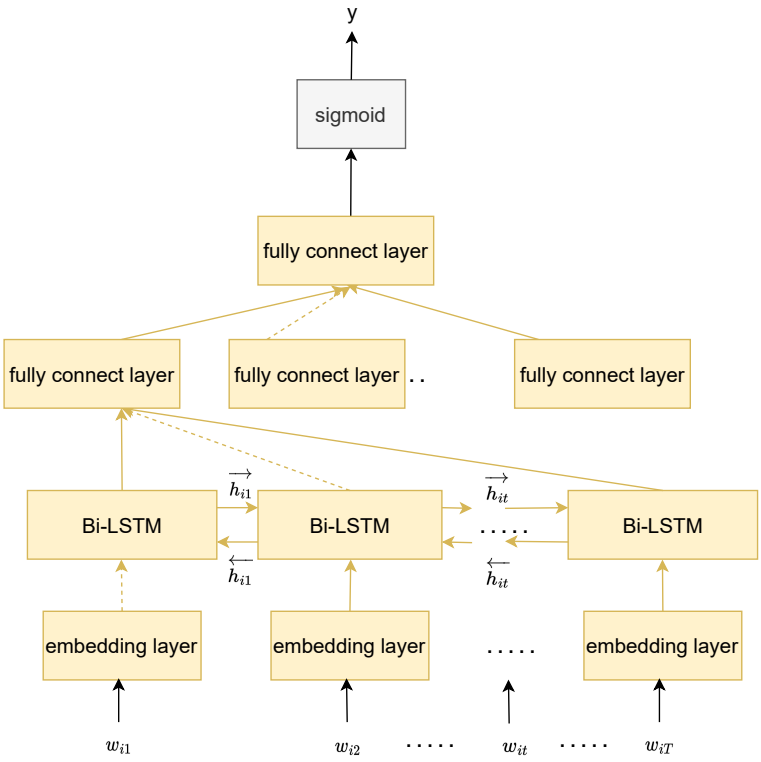


Figure 5.2: The structure of the baseline system: bi-LSTM.

5.4.2 Baseline Systems

In order to demonstrate the efficiency of the proposed *HBANN* system, two baseline systems are described: a *bi-LSTM* system and a *hierarchical bidirectional recurrent neural network (HBRNN)* system.

The bi-LSTM system treats the transcript as a sequence of words by removing all the punctuation. The input words are embedded into word vectors with the embedding layer initialised with the Glove pre-trained word embedding matrix. Following the BLSTM layer is a fully-connected layer with a *sigmoid* activation function. This system is designed to test the necessity of the hierarchical system, and the structure is shown in Figure 5.2.

The HBRNN system has the same structure as the HBANN described in Section 5.3

but without the attention mechanism in the word and sentence levels for taking the word and sentence vectors accordingly. The structure is shown in Figure 5.3. This system is designed to test the efficiency of the attention mechanism for the hierarchical system. The performance of the two baseline systems is shown in Section 5.5.

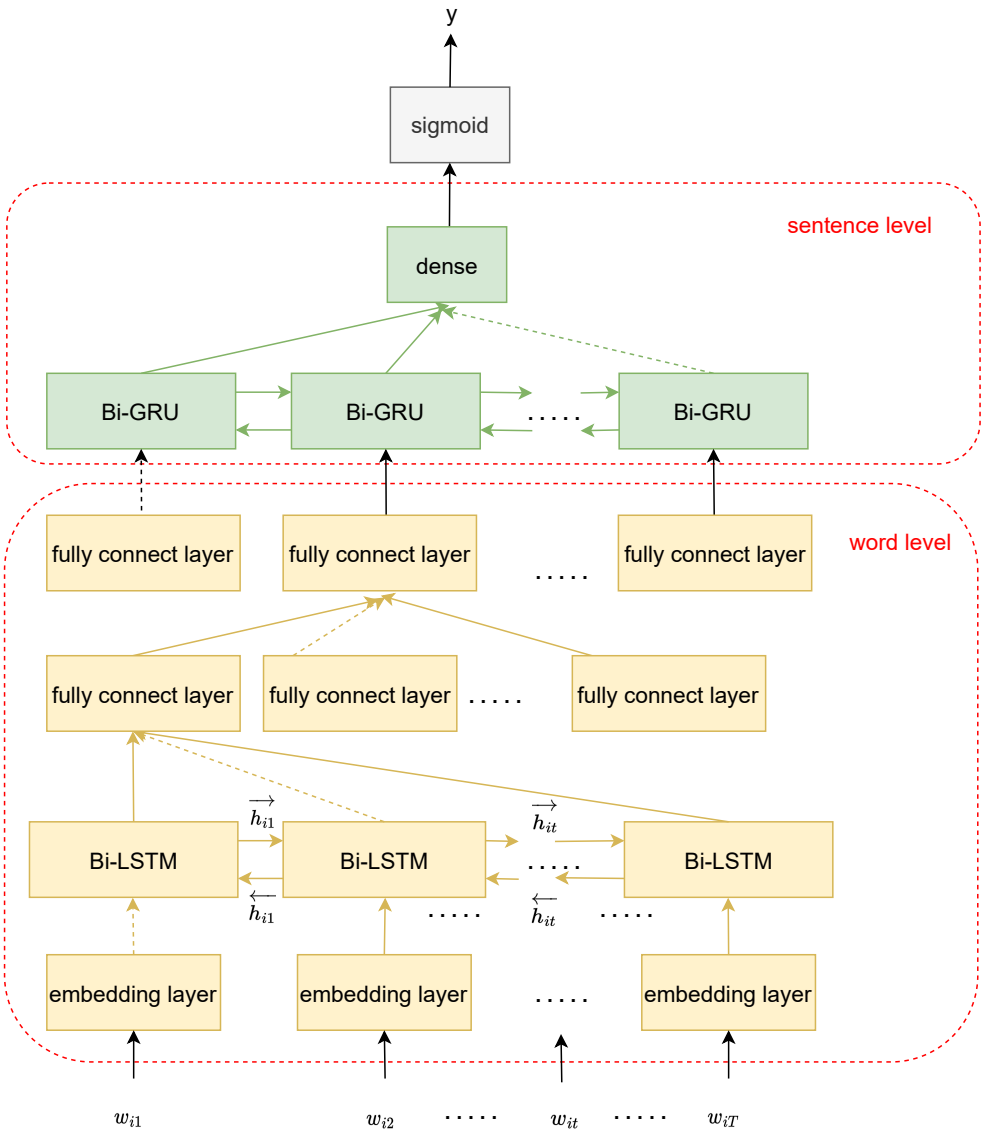


Figure 5.3: The structure of the baseline system: HBRNN.

5.4.3 Evaluation Settings

The DementiaBank dataset does not come with a specified training and test partition. For evaluation, A 10-fold cross validation (CV) setup is used. To ensure the comparison

of the system, 10-fold lists are **generated** and then **fixed** for all the experiments related to the DementiaBank dataset in this thesis. While generating 10-fold data lists, there is no overlap between speakers in the training set (8 folds), development set (1 fold), and test set (1 fold) at any time (speaker-independent). The lists in each fold include a training list, a development list and a test list. There is no reuse between the training set, development set and test set in each fold, and no reuse between the 10 folds' test sets. The list generated for each fold can ensure that transcripts from the same speaker are found in the same set, and the number of transcripts in the three sets of each fold is kept as balanced as possible. To avoid over-fitting, in this thesis, 10-fold CV is defined as follows: rather than use 9 folds of data for training the system, the training set is divided into the training set (8 folds) and the development set (1 fold). 10 models are trained using the 8-fold training data in each training set list, and the best-trained model for each fold is selected based on the performance (F-score) of the trained model on the development set. The reported test set result is averaged over the results from the 10 test sets in the 10 lists. As shown in Table 4.1, one speaker can sometimes correspond to more than one recording.

To explore whether the out-of-domain data can improve the performance of the designed system, the IVA_{33} data is used as the extra data. Rather than adding the IVA_{33} data to the development set (hoped to be as similar to the test set), all the data in the IVA_{33} data are added to the training set because it can increase the diversity of the training data. As a result, the 10-fold segmentation lists are kept unchanged except for adding the IVA_{33} data into the training set, so the results with IVA_{33} are comparable with the setup that uses only the DementiaBank dataset for training.

5.4.4 Model Configuration

Stanford's Global Vector (GloVe) pre-trained Word Embedding is trained on large datasets for capturing the semantic and syntactic meaning of a word. A 100-dimensional pre-trained Glove word embedding matrix is used for initialising the word embedding layer [Pennington *et al.*, 2014], similarly to the setup in Yang *et al.* [2016]. The selection is based on the performance of the HBANN system on the development set.

The proposed system is trained on a fixed number of 20 epoch, and the batch size is set to 20. The best-trained system is selected with the F-score on the development set. The number of the RNN units, including the LSTM and GRU, is set to 100, and the fully-connected layer dimension in word-level is set as 50. The attention layers' dimension of both the word and sentence levels are set to 30. An *adam* optimizer with a 0.001 learning rate is used [Kingma & Ba, 2014]. To avoid overfitting, dropout is applied to the output of all the functional layers. For all dropout layers, the dropout rate is set to 0.3. Considering the average sentence length, sentences shorter than 30 are zero-padded, and sentences with more than 30 words are truncated. The results of the test set are calculated by averaging the results across the 10-fold CV. For the bi-LSTM system, the number of words in a transcript is set as 30×30. All these parameters are decided based on the performance of the development set on the HBANN system.

5.5 Results and Analysis

In this section, both the manual and automatic transcripts generated by the ASR system are used respectively for examining the performance of the proposed HBANN system. First, the performance of the transcripts is tested with the proposed system HBANN and two baseline systems: bi-LSTM and HBRNN (Section 5.5.1). Then, the trained system is further analysed and shown in Section 5.5.2.

5.5.1 Experimental Results

First, the manual transcript is used for training and testing the system. The result is presented in Table 5.2. The experiment is composed of three parts to verify the efficiency of the hierarchical structure, the attention mechanism, and the punctuation restoration. As shown in the table, precision, recall and F-score are selected as the criteria. The first three rows correspond to the results achieved by the two baseline systems (bi-LSTM and HBRNN) and the proposed system HBANN on the manual transcripts. The last line corresponds to the results of the HBANN system on the manual transcripts, but with the automatic punctuation restoration, which is designed to test the effect of the automatic

punctuation restoration on the linguistic-based dementia detection.

Table 5.2: The classification results of the HBANN system, bi-LSTM system and HBRNN system on the manual transcripts of the DementiaBank dataset.

Punctuation	System	Precision%	Recall%	F-score%
Manually	Bi-LSTM	75.02	73.73	73.45
Manually	HBRNN	78.26	77.77	75.68
Manually	HBANN	**84.02**	**84.97**	**84.43**
Automaticly	HBANN	81.17	81.23	79.77

The observations are summarised by comparing the results from the proposed system HBANN and the two baselines:

- **Effect of the hierarchical structure**: By comparing the result from the bi-LSTM and the HBRNN systems, it is found that, after including the hierarchical mechanism, the F-score is improved from 73.45% to 75.68%. It demonstrates that the hierarchical neural network can extract more efficient linguistic information than the non-hierarchical system while doing the linguistic-based dementia detection. Therefore, it is inferred that considering the hierarchical structure of the transcripts into the linguistic feature extraction is beneficial.

- **Effect of the attention mechanism**: After adding the attention mechanism to the hierarchical system, the F-score is further improved from 75.68% to 84.43% by comparing the HBRNN and the HBANN system. The result shows that the attention mechanism can benefit the classification performance of the system. More analysis on the attention mechanism can be found in Section 5.5.2.3.

- **The influence of automatic punctuation**: The automatic transcripts output by the ASR system have no punctuation, which is required to be used as the input of the hierarchical system. Therefore, punctuation restoration technology is used on automatic transcripts. However, the effect of automatic punctuation restoration could not be examined on the automatic transcripts as there are no ground-truth labels available. Therefore, as an alternative, the influence of the automatic punctuation restoration is

evaluated on the manual transcripts. In the experiment, to add punctuation, the automatic punctuation restoration method described in Section 5.4.1.2 is used on the punctuation removed manual transcripts. The result shows that automatic punctuation restoration can cause an absolute 4% decline in the F-score (from 84.43% to 79.77%). As discussed in Yuan *et al.* [2020], the punctuation can be used as a proxy for the pause and disfluency. The mislocated punctuation can result in ambiguity, which may decrease the performance of the processed transcripts.

Then, the automatic transcripts with automatic punctuation are tested. Rather than only using the data from the DementiaBank dataset for training, the IVA_{33} dataset is also used for exploring whether the extra data can improve the performance of the proposed system on the test set.

Table 5.3: The detection results of the HBANN system and the baseline systems on the automatic transcripts of the DementiaBank dataset and the IVA_{33} dataset.

Training Set	**System**	**Precision%**	**Recall%**	**F-score%**
DementiaBank	Bi-LSTM	68.18	67.74	66.44
DementiaBank	HBRNN	74.03	74.80	72.11
DementiaBank	HBANN	**79.22**	76.33	74.37
DementiaBank+IVA_{33}	HBANN	78.83	**77.73**	**76.09**

The following conclusions can be drawn from Table 5.3:

- **Results on the automatic transcripts**: The performance of the bi-LSTM system, the HBRNN system, and the HBANN system remains consistent on the manual and automatic transcripts. As shown, the best performance in Table 5.3 is based on the HBANN system (74.37% F-score) by using the automatic transcripts of the DementiaBank dataset.

- **Manual and automatic transcripts comparison**: It is found that a gap exists between the F-score values of manual and automatic transcripts. It is inferred that the phenomenon originates from two sources: the effect of misrecognised words in the ASR transcripts and the punctuation restoration errors as shown in Table 5.2.

- **Effect of using additional training data**: Including the IVA_{33} dataset into the training set can increase the F-score of the test set from 74.37% to 76.09%, as shown by comparing the last two rows of Table 5.3. It shows that the proposed HBANN system still has the potential to be improved if proper training data can be included, even from a non-homogeneous dataset with an apparent mismatch. However, due to the limitation of available publicly dataset for dementia detection, the conclusion may be biased.

The system compares favourably with previous methods working on the manual transcripts of the DementiaBank dataset. The selected research based on the DementiaBank dataset is shown in Section 4.1.1. In Fraser *et al.* [2016] and Budhkar & Rudzicz [2018], the accuracy of 81.92% and the F-score of 77.50% was achieved, compared with 84.02% F-score for the HBANN system. Even though Karlekar *et al.* [2018] got a comparable result of 84.9%, it did not use the 10-fold CV and a speaker-independent way to evaluate the system, making the result from Karlekar *et al.* [2018] and those presented here incomparable. The result in this study on the DementiaBank dataset automatic transcripts is also considerable. In Ammar & Ayed [2018], an almost similar precision (79%) was achieved by feature selection based on both the acoustic and linguistic features, compared with the F-score of 76.09% by using the linguistic features only.

5.5.2 Result Analysis

The efficiency of the proposed HBANN system has been presented by comparing it with the baseline systems. In this part, firstly, the influence of stop words (such as "the", "their" that in the NLTK stop-words corpora on classification is explored. Secondly, the effect of word embedding is examined. Then, the word-level and sentence-level attention mechanisms are analysed to understand how attention influences classification. Finally, the word frequency is analysed by comparing the transcripts from the HC and the AD groups.

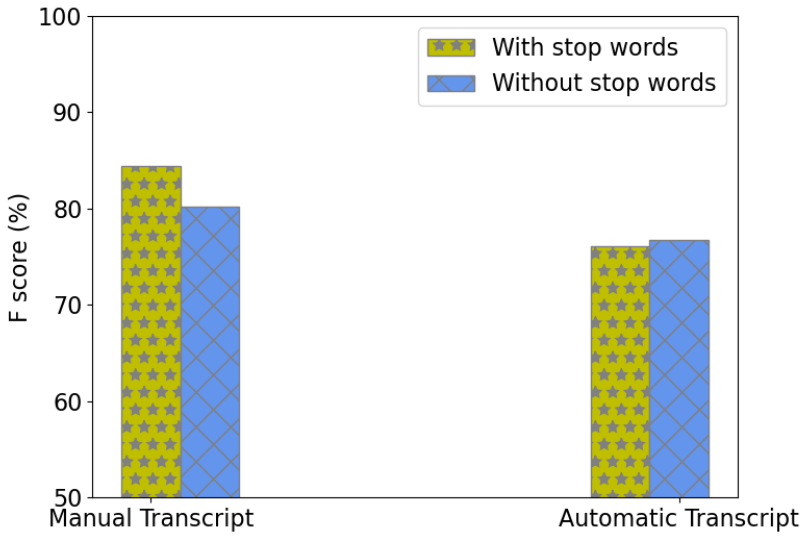


Figure 5.4: The effect of stop words on the manual and automatic transcripts for dementia detection.

5.5.2.1 Stop Words Analysis

In Adhikari *et al.* [2021], it is proposed that the *stop words* like "the", "an", are overused by people living with dementia. In order to analyse how the stop words influence the classification results in our study, the stop words are removed from the manual and automatic transcripts by using the NLTK stop-words corpora. Then, the processed transcripts are used as the input of the HBANN system. The classification results are shown in Figure 5.4. The effects of removing stop words from the manual and automatic transcripts are different. As shown, after removing the stop words, the F-score drops by 4.19% (from 84.43% to 80.02%) on the manual transcript but increases by 0.65% (from 76.09% to 76.74%) on the automatic transcript. It is inferred that the effect of stop words on the manual transcripts is more pronounced, as stop words may not be recognised correctly in the automatic transcripts.

5.5.2.2 The Effect of the Word Embedding Methods

In the experiment, as in Section 5.4.4, the word embedding layer initialised with the Glove pre-trained word embedding matrix is used for word embedding in the end-to-end system.

While training, the parameters in the embedding layer are set as trainable. The following experiments are designed to explore the different word embedding methods' effect on dementia detection. The three word embedding layer initialising methods are tested with both the manual transcripts and the ASR generated automatic transcripts.

Table 5.4: The results achieved using different word embedding initialisation and training methods on the manual and automatic transcripts.

Transcript	Embedding Layer	Precision%	Recall%	F-score%
Manual transcript	Randomised (trainable)	74.37	74.42	72.51
	Glove(fixed)	78.71	79.50	78.11
	Glove (trainable)	**84.02**	**84.97**	**84.43**
Automatic transcript	Randomised (trainable)	74.64	73.40	71.00
	Glove(fixed)	71.56	70.83	69.42
	Glove (trainable)	**78.83**	**77.73**	**76.09**

The results are shown in Table 5.4. The conclusions summarised from the Table 5.4 are as follows:

- For the embedding layer, which is initialised with the pre-trained Glove embedding matrix, the trainable embedding layer can improve the F-score from 78.11% to 84.43% on the manual transcripts and improve the F-score from 69.42% to 76.09% on the automatic transcripts. The result shows a mismatch between the performance of the fixed and the trainable embedding layer's setting.

- When the embedding layer is trainable, compared with the randomised parameters, initialising the embedding layer with the Glove pre-trained embedding matrix can improve the F-score from 72.51% to 84.43% on the manual transcripts and improve the F-score from 71.00% to 76.09% on the automatic transcripts. Thus, the result shows that though the Glove pre-trained embedding matrix is not perfect for our task, the general information it includes can make the system perform better than the randomly initialised embedding matrix.

- As shown, on the manual transcripts, a fixed embedding layer initialised with the Glove pre-trained embedding matrix (78.11%) performs better than a trainable embedding

layer initialised randomly (72.51%). However, for the automatic transcripts, the conclusion is the opposite. The trainable embedding layer initialised randomly performs better (71.00%). Therefore, it is inferred that the manual transcripts are more similar to the materials trained for the Glove (high-quality essays) pre-trained matrix. In comparison, the automatic transcripts are not that perfect due to the errors caused by the ASR system and the punctuation restoration. Therefore, the trainable embedding layer initialised randomly (71.00%) can perform better than the fixed embedding layer initialised with the pre-trained matrix (69.42%).

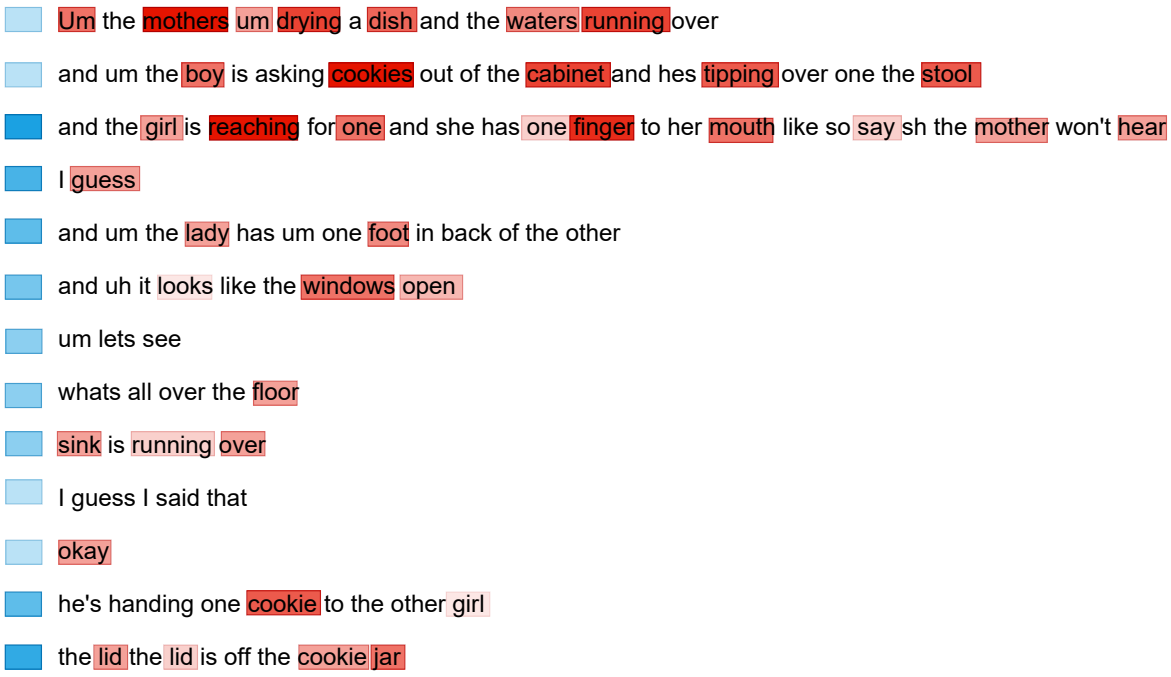


Figure 5.5: An example of visualising the word-level and sentence-level attention weights.

5.5.2.3 Attention Mechanism Analysis

A manual transcript is selected to visualise how the attention mechanism works for classification. The attention weights are visualised with shades of red. The darker the colour, the higher the weight. For the visualisation, the weights are normalised to $[0, 1]$ at the word and the sentence levels, respectively. The attention weights from the word-level and

sentence-level are shown in Figure 5.5.

As shown in the figure, at the word level, the informative words like *water* and *mother* have higher attention weights compared with the empty words (words that cover very little or no information [Almor *et al.*, 1999]) like *um* and *won't*. At the sentence level, the sentence with the highest attention weight in the transcript is *and the girl is reaching for one and she has one finger to her mouth like so say sh the mother wont hear*. Compared with the attention weights of *I guess I said that* and *okay*, the attention weights visualisation shows the sentence that includes more information has a higher attention weight compared with the sentence that includes less information.

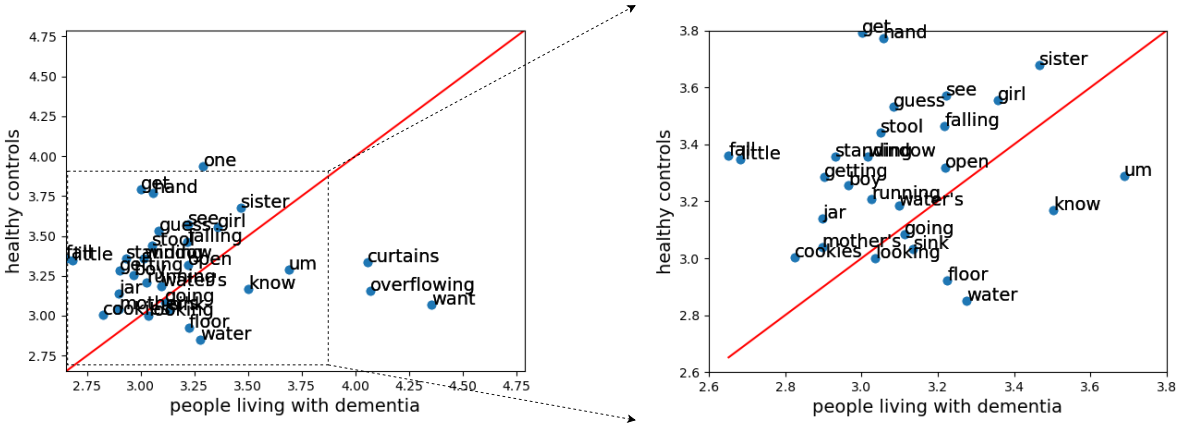


Figure 5.6: The extracted attention weights for the words from the HC and AD groups.

To understand the attention weights difference from the HC and AD groups, the attention weight for the selected words from the HC and AD are plotted in Figure 5.6. For clarity, part of the figure is enlarged and displayed on the right-hand side. All the attention weights are extracted from the 10-fold test sets. The attention of each word is estimated by averaging over all the attention weights of the same word. For each word, the averaged attention weight in the descriptions from the HC group are set as the x *axis*, and the averaged attention weights for the same word from the AD group are set as the y *axis*. The red line is the proportional function $x = y$. The words above and below the line are those words that achieve a higher or lower attention weight in the transcripts from the HC group compared with those from the AD group.

"Analysing the word attention weights from the AD and HC may be informative for understanding what the deep learning system has learned while doing the classification. By comparing the word attention weights from the AD and HC, it is found that some nouns, like *sister*, *hand*, *window* and *girl*, have higher attention weights from the HC than from the AD. Though the attention weight of the word **mother's!** (**mother's!**) from the HC is higher than the weight from the AD, but not a significant difference. How deep learning learns the attention weights for doing the classification tasks is expected to be further explored in future research.

5.5.2.4 Word Frequency Analysis

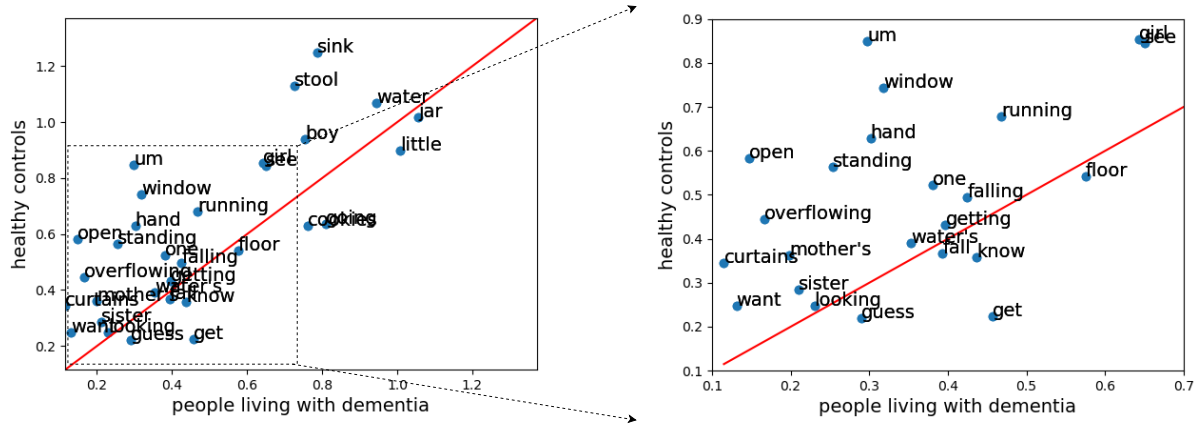


Figure 5.7: The word frequencies for the HC and AD groups.

The word frequency is estimated by averaging the number of each word in the transcripts from the HC and AD groups respectively. For visualisation, the *x-axis* and *y-axis* of each word in the figure is the word frequency from the HC and AD transcripts. The selected words frequency is shown in Figure 5.7. For clarity, an enlarged figure for a part of the original figure is plotted on the right-hand side. As shown, most of the words have a higher word frequency from the HC group than from the AD group, especially for nouns, like *sink*, *water*, *window*. It is known that HCs tend to have more nouns in their speech than people living with AD, as people living with AD often have word-finding difficulties. The conclusion is consistent with previous research [Graham *et al.*, 2001].

5.6 Summary

Automatic linguistic-based dementia detection is a relatively new research field. People with dementia show a decline in their speech at both word and sentence levels. In this chapter, the DementiaBank dataset is used to explore how to extract the linguistic information embedded in the transcripts for dementia detection.

For extracting the linguistic information, in this chapter, inspired by clinical diagnostic knowledge that linguistic ability declines at both the word-level and sentence level, an end-to-end hierarchical system was proposed for extracting linguistic information from the transcripts at both the word-level and sentence-level. At the same time, while clinical diagnosis, clinicians pay different attention to information units and pronouns, and the sentences with different idea densities are considered accordingly. To weigh the importance of the learned feature vectors, the attention mechanism was used to weigh the words and sentences respectively in the proposed system. The study achieved better performance on the DementiaBank dataset than existing studies. In the experiment, for evaluating the proposed system, both the manual transcripts provided in the DementiaBank datasets and the automatic transcripts generated by the ASR system were used as the system's input. The results showed that the proposed HBANN system is efficient for both the manual and the automatic transcripts. Furthermore, it is worth mentioning that using the extra IVA_{33} dataset for the system training can increase the system's classification performance on the DementiaBank test set, even though an apparent mismatch exists between the two datasets.

To demonstrate the efficiency of the proposed hierarchical structure and the attention mechanism on linguistic-based dementia detection, two baseline systems were designed: HBRNN and Bi-LSTM. The result shows that using both the attention mechanism and hierarchical structure can improve the performance of the linguistic system using manual or automatic transcripts. Analysis has been done to explore the effect of stop words on the classification results using manual or automatic transcripts. The results show that removing the stop words from manual transcripts can reduce detection but slightly increase automatic transcripts. The study demonstrated the efficiency of the trainable

word embedding layer initialised with the Glove pre-trained matrix, attention mechanism and hierarchical structure for the linguistic-based dementia detection with the designed experiments. To understand how the proposed system works, the attention weights, the embedding layer and the word frequency of the transcripts were analysed.

After the research proposed in this chapter published, more state-of-the-art structures, like BERT [Devlin *et al.*, 2018], were proposed to be used for linguistic-information modelling for dementia detection. In the paper accepted by Interspeech-2020 ADReSS special session, seven out of thirteen papers used the BERT for modelling the linguistic information [Balagopalan *et al.*, 2020; Farzana & Parde, 2020; Koo *et al.*, 2020; Pompili *et al.*, 2020b; Searle *et al.*, 2020; Syed *et al.*, 2020; Yuan *et al.*, 2020]. In parallel, nine out of eleven papers [Chen *et al.*, 2021; Pan *et al.*, 2021a; Pappagari *et al.*, 2021; Pérez-Toro *et al.*, 2021; Qiao *et al.*, 2021; Rohanian *et al.*, 2021; Syed *et al.*, 2021; Zhu *et al.*, 2021] accepted by the Interspeech-2021 ADReSSo special session extracted the linguistic features using the BERT. In comparison with the proposed HBANN system, BERT can perform better on the ADReSS dataset by comparing the result reported in Cummins *et al.* [2020] and Yuan *et al.* [2020].

Different from HBANN, which uses RNN to learn the context information, BERT can learn longer context-dependency with transformer. In addition, the BERT pre-trained model is available for initialising the parameters before fine-tuning on the relatively small datasets used for dementia detection. As shown in this chapter, using the IVA-33 dataset for modelling training can improve the system's performance, though there is an obvious mismatch between the IVA-33 and DementiaBank dataset. The result encourages us that using the out-of-set dataset has the potential to improve the system's performance. It is inferred that fine-tuning the pre-trained model has the potential to make up for the limitation of the dementia detection dataset, which should be explored further in future work.

In this chapter, the punctuation is added automatically to the transcripts output by ASR according to the context using the system proposed in Tilk & Alumäe [2016]. In comparison, in Yuan *et al.* [2020], the punctuation is added according to the pause length for including the speech rhythm information, which has been demonstrated to be efficient.

More analysis on how to use the speech rhythm information for dementia detection will be explored in Chapter 7.

Chapter 6

End-to-end Feature Extractor for Speech-based Dementia Detection

Contents

6.1 Introduction 101

6.2 End-to-end Feature Extractor 102

6.3 Experimental Setup 105

6.3.1 Dataset 105

6.3.2 Evaluation Settings 106

6.3.3 Model Configuration 107

6.3.4 Baseline Feature Sets 108

6.4 Results 108

6.4.1 Classification Results on the IVA_{3class} Dataset 108

6.4.2 Analysis of SincNet Filters from the IVA_{3class} Dataset 110

6.4.3 Classification Results on the DementiaBank Dataset 115

6.4.4 Dataset Comparison 115

6.5 Summary 116

In Chapter 5, linguistic-based dementia detection methods have been proposed, which are designed according to the clinical diagnostic process. For linguistic-based dementia detection, the system's input is transcripts, which need to be generated from recordings either manually or automatically because most of the collected data is audio recordings. The acoustic features are extracted directly from the audio recordings, but the extracting process is more challenging as the audio includes noise and the audio dimension is high. The feature extraction and selection depends on the classification task, the quality and the number of audio recordings. In previous research, arriving at a consensus for the best feature set for different classification tasks and datasets is always challenging when using the traditional features. Under this situation, the deep learning technologies are expected to be used for extracting the data-driven/task-driven acoustic features, so this chapter aims at investigating an answer to the first research question:"how can state-of-the-art deep neural networks be applied for speech- and language-based dementia detection? (**RQ1**)". The structure of this chapter is as follows:

Section 6.1 introduces the background of the speech-based dementia detection methods.

In **Section** 6.2 a proposed end-to-end feature extractor is described.

Section 6.3 contains the information about the experimental setup of our proposed system and the baseline acoustic features.

Section 6.4 summarises the classification results and analyses our proposed system.

Section 6.5 contains the summary of this chapter.

6.1 Introduction

The research concerning automatic speech-based dementia detection is reviewed in Section 3.2. As mentioned, traditional acoustic features can be classified into two categories depending on the feature type: basic acoustic features, like MFCC [Alhanai *et al.*, 2017], F_0 [Meilán *et al.*, 2014], Jitter and Shimmer [Lopez-de Ipiña *et al.*, 2015], and specifically designed features informed by medical knowledge, like the conversational analysis features proposed in Mirheidari *et al.* [2019a]. While the basic acoustic features contain information about the speaker's cognitive status, they cannot describe the task-specific symptoms well when being used alone. That is the reason why a long list of acoustic features is needed for dementia detection [Fraser *et al.*, 2016; Hernández-Domínguez *et al.*, 2018; Luz, 2017; Orimaye *et al.*, 2017; Yancheva *et al.*, 2015]. For example, the Mel-scale filter bank that is designed to mimic auditory and physiological evidence of how humans perceive speech signals [Davis & Mermelstein, 1980] is used broadly but cannot always guarantee to be the best filter bank for the target task [Ravanelli & Bengio, 2018a]. Also, one single feature cannot describe the symptom comprehensively. Therefore, for improving the classification performance, some researchers choose to extract very large feature sets (often in the thousands), and the most suitable features are selected [López-de Ipiña *et al.*, 2015; Weiner & Schultz, 2018] for the specific dataset and classification task. However, the selected acoustic feature set cannot ensure consistent performance across different classification tasks and datasets. On the other hand, the specially designed features require an exact translation from an expert's medical knowledge into a mathematical expression, which is challenging.

The performance of a typical classification pipeline is highly dependent on the quality of the front-end features. Recently, many machine learning tasks have deployed end-to-end methods that automatically learn features or a joint system to obtain the feature representation and classification models [Trigeorgis *et al.*, 2016; Tzirakis *et al.*, 2018]. Furthermore, as shown in Section 3.2.2, extracting the target information directly from the raw waveform by neural networks has been an active and promising area of research, especially for mainstream speech research fields like speaker recognition and emotion

recognition. Moreover, for some dementia detection tasks, compared with the traditional acoustic features, end-to-end solutions have shown advantages in achieving a more efficient information representation [Chen *et al.*, 2018].

For dementia detection, end-to-end neural networks have shown their efficiency in various tasks as a front-end feature extractor compared with traditional acoustic features [Hinton & Salakhutdinov, 2006; Sainath *et al.*, 2015b]. However, most neural networks appear as a *black box*, which means it is hard to analyse and interpret any learned representations that could lead to meaningful insights for clinical diagnosis.

When designing an end-to-end system, the first layer is always very important for the performance of systems working directly on raw waveform input as it deals with the high-dimensional and noisy input [Ravanelli & Bengio, 2018b]. For processing the raw waveform, CNNs are usually used as the first layer of the end-to-end system to extract the acoustic features from the raw waveform. The function of an CNN can be thought of as data-driven finite impulse-response set of filterbanks followed by a nonlinearity function [Sainath *et al.*, 2015b]. However, the learned filter banks in the first convolutional layer with the raw waveform as the input lack interpretability [Ravanelli & Bengio, 2018a]. Under these considerations, an end-to-end feature extractor is designed in Section 6.2 for training the interpretable filters with the raw waveform as the input.

6.2 End-to-end Feature Extractor

In Sainath *et al.* [2015a], it was found that combining CNNs, LSTMs, and DNNs for speech processing in a unified architecture allows for the exploitation of their complementary natures. In 2018, a novel CNN structure named *SincNet* was proposed in Ravanelli & Bengio [2018b]. The filters in the SincNet are defined with a set of parametrised *Sinc* functions, making the filters more *interpretable*. Also, the SincNet benefits from having fewer parameters to learn, making them converge faster [Ravanelli & Bengio, 2018b]. These characteristics make the SincNet suitable as the first layer in our designed system. In this chapter, the SincNet is applied as the first layer of the designed system followed by a CNN layer (**C**), an LSTM layer (**L**) and an attention layer (**A**). The designed system,

as shown in Figure 6.1, is referred to as *Sinc-CLA* in the following. The system is used as the feature extractor by extracting acoustic features from either the fully connected or attention layers.

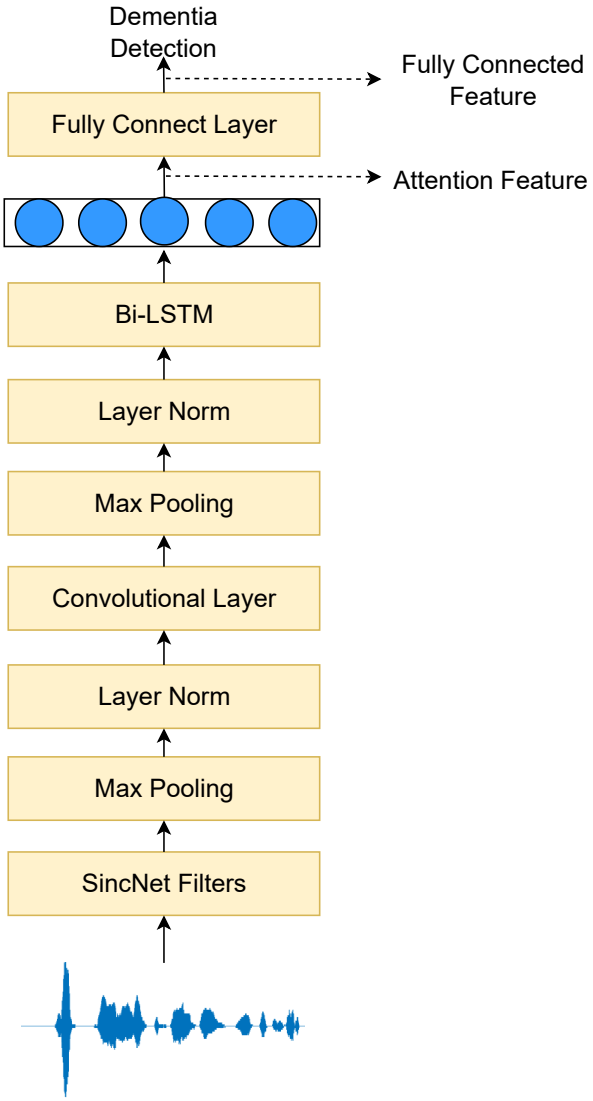


Figure 6.1: The structure of the Sinc-CLA feature extractor.

The first functional layer of the system is the SincNet layer, followed by the max-pooling layer plus normalisation. For the N defined filters in the SincNet layer, the output of the SincNet layer for the i_{th} filter, $i \in [1, N]$ is defined as follows:

$$\begin{aligned} h_i[n] &= x[n] * g[n, f_{i1}, f_{i2}] \\ &= x[n] * [2f_{i2}sinc(2\pi f_{i2}n) - 2f_{i1}sinc(2\pi f_{i1}n)] \end{aligned} \tag{6.1}$$

where $x[n]$ is the n_{th} chunk of the raw waveform signal; $g[n, f_{i1}, f_{i2}]$ is used to represent the function of the i_{th} filter-bank, and f_{i1} and f_{i2} are the low and high cut-off frequencies that need to be learned while training. The sinc function is defined as $sinc(x) = sin(x)/x$. To avoid the ripples in the passband and attenuation in the stop band, a Hamming window [Clark, 2005] is applied on $g[n, f_{i1}, f_{i2}]$. In Eq. 6.1, the N filters are initialised with the cut off frequencies of the Mel-scale filter-bank in order to include the human perception into the system initialisation.

The second part of the system is a standard 1-D convolutional layer, a max-pooling layer, plus layer normalisation. CNNs have demonstrated their ability to extract robust and invariant representations when facing the typical frequency variations of acoustic signals by applying local filters and pooling mechanism [LeCun *et al.*, 1999].

The output $H[n]$ of the second normalised layer is used as the input to the third part of the proposed system, the Bidirectional LSTM (BLSTM). LSTMs are good at capturing the temporal evolution of speech signals and model the sequence information from the time series [Chung *et al.*, 2014]. Moreover, the BLSTM can utilise both the forward and backward information of the input feature.

Then, an attention layer and a fully connected layer are applied for feature weighting and mapping. The attention mechanism has lately been used in different fields and achieved a great deal of success as described in Section 3.2.2.4. It was also used for weighting the linguistic features in the system proposed in Chapter 5. The main idea behind the attention mechanism is to apply higher attention weights to the more critical dimension of the feature vector for classification. The system training is based on minimising the loss between the predicted label and the ground-truth label.

After training, the parameters in the system are fixed to enable the extraction of the trained features. To test the feature representation ability of the feature extractor, as illustrated in Figure 6.1, features extracted from either the fully connected layer or the attention layer are evaluated, respectively. The learned features are named as *fully connected feature* (the output of the fully connected layer) and *attention feature* (the output of the attention layer) respectively. The features are extracted from the end-to-end system trained with the training data and the specific target, so the extracted features

are data-driven and task-specific. For each classification task, one feature extractor is trained. More details of the feature extractor setting can be found in Section 6.3.

6.3 Experimental Setup

This section presents the experimental setup, including the datasets used for our proposed system, the evaluation settings, the model configuration and baseline feature sets.

6.3.1 Dataset

The features extracted for dementia detection are mostly traditional features in previous research for the IVA dataset [Mirheidari, 2018; Mirheidari *et al.*, 2019a,b]. Similarly, as described in Section 4.1.1, in previous research, the acoustic information-based feature extraction methods on the DementiaBank dataset are mainly based on traditional features. The main reasons are the high background noise of the audio recordings in the DementiaBank dataset and the limited audio recordings for training an end-to-end system. Though with these difficulties, in this chapter, both the DementiaBank dataset and the IVA_{3class} dataset are used for testing the performance of the proposed end-to-end feature extractor.

The information about the IVA_{3class} dataset is shown in Table 4.8. The IVA_{3class} dataset includes the 88 recordings from three diagnostic categories: HC, MCI and ND. The average duration of the recordings in the IVA_{3class} dataset is about 9 minutes which is too long to utilise directly as the input of the Sinc-CLA feature extractor. A similar problem was described in Warnita *et al.* [2018], and they chose to segment the input with manual information. As the overall purpose of this research is to investigate fully automatic approach, instead, it was chosen to cut the recording into 2-second chunks for constructing a fully automatic system. Each chunk is assigned a label corresponding to its diagnostic category.

The information about the acoustic recording in the DementiaBank dataset is shown in Table 6.1. Our research in this chapter aims at exploring the proposed system on the binary classification task. Only the recordings from HCs and people living with AD are utilised. The experimental results are comparable with the results in Chapter

5 as the experimental setting is consistent. For processing the audio recordings in the DementiaBank dataset, similarly to the IVA_{3class} dataset, the recordings are cut into 2-second chunks. Each chunk is assigned a label corresponding to its diagnostic category.

Table 6.1: The speaker information, recording information and audio information for the DementiaBank Dataset.

Diagnostic Category	Number of Speakers	Number of Recordings	Audio Duration
AD	168	222	4h12min
HC	89	255	3h28min
Total	257	477	7h40min

6.3.2 Evaluation Settings

The 10-fold CV is used on the two datasets, and each fold is fixed for all the presented experiments. As shown in Table 4.8 and Table 6.1, some speakers contributed more than one recording, and these are kept in the same partition, ensuring speaker independence. The CV used in this chapter is similar to that described in Section 5.4.3. For each fold, the number of recordings in the three partitions (training, development, and test) is set as balanced as possible in terms of the diagnostic category. For the DementiaBank dataset, the 10-fold CV list is the same as the one used in Chapter 5. While training the feature extractor, the training set includes 8 folds of data, while the development set and the test set include 1 fold data.

In the experiment, a typical classification pipeline is used to evaluate the extracted features. The front-end features are either the baseline feature sets or the features learned by the Sinc-CLA, followed by a back-end classifier. The LR and SVM, the most commonly used classifiers in acoustic-based cognitive impairments detection fields, are adopted as our classifiers as in previous research [Edwards *et al.*, 2020; Luz *et al.*, 2018; Satt *et al.*, 2013, 2014]. The kernel type in the SVM is set as *rbf*. For each data fold, the features from training and development sets (9 folds) are used to train the back-end classifier, and the test set is used for evaluation. The presented result is averaged across the 10-fold test set.

The classification accuracy of the chunks is referred to as the *chunk-level accuracy*. The predicted recording label is estimated by majority voting over the predicted chunk-level labels belonging to the same recording, and the accuracy is referred to as *recording-level accuracy*. To verify our system, the classification tasks include the HC vs. ND, the HC vs. MCI and the HC vs. people living with either ND or MCI on the IVA_{3class} dataset. For the DementiaBank dataset, the classification task is the HC vs. AD.

6.3.3 Model Configuration

The segmented chunks in the training set are fed into the designed feature extractor (Sinc-CLA). The SincNet layer is composed of N=80 filters of length L=125 samples. The parameters for the filters in the SincNet layer are initialised with the cut-off frequencies of the Mel-scale filter-bank as in Ravanelli & Bengio [2018b]. The standard convolutional layer uses 60 filters of length 5. The max-pooling size of the two convolutional layers is 3. The number of units in the BLSTM is 50. The output of the BLSTM layer is a 100 dimensional feature, which is the concatenation of the two 50 unidirectional LSTM outputs. The dimension of the attention matrix is set as 30. The output of the attention layer is a 100 dimension vector. The fully connected layer composes 1024 neuron units. In the model, all hidden layers use the *leaky-ReLU* [Maas *et al.*, 2013] non-linearities. The *rmsprop* [Tieleman & Hinton, 2012] is applied as the optimizer with a learning rate of 0.01. While training, the mini-batch size is set to 30 and the epoch is set to 40. The parameters related to the SincNet are set according to the parameters used in Ravanelli & Bengio [2018b] at first, and then tuned according to the performance of the development set. Similarly, the other parameters are set according to the performance of the development set.

After the feature extractor is trained, the parameters are fixed, and the 2-second chunks are input into the Sinc-CLA feature extractor. The features output by the attention layer and layer (named "attention feature" and "fully connected feature" in the following) are used for the classification experiments.

6.3.4 Baseline Feature Sets

Research has shown promising results for using features initially proposed for emotion recognition in systems for automatic assessment of cognitive impairments [Luz *et al.*, 2020a; Warnita *et al.*, 2018]. Therefore, IS10 [Schuller *et al.*, 2010] and ComParE [Eyben *et al.*, 2013] features, which have achieved outstanding results for dementia detection (as reviewed in 3.2.1), are adopted as the baseline feature sets in our experiment. The features are extracted using the OpenSMILE [Eyben *et al.*, 2013] toolkit. Compared with the Low Level Descriptor (LLD) features, the statistical suprasegment feature can provide better performance on our task. To get the suprasegment feature for each 2-second chunk, the mean, maximum, minimum, median, and standard deviation are calculated across time on the LLD feature matrix as in Alhanai *et al.* [2017]. Then, a list of 380-dimension (76×5) features based on the IS10 feature set and 650-dimension (130×5) features based on the ComParE feature set are generated.

6.4 Results

In this section, the classification results are presented, and the analysis of the learned filters is shown. To test the efficiency of our proposed system, either the fully connected feature or attention feature is used as the learned task-specific feature. For analysis, the outputs of the SincNet layer and the learned parameters of the SincNet filters are plotted. The classification results on the baseline feature sets (the IS10 and ComParE) and the dense/attention features are calculated by averaging across the 10-fold test set. Finally, both the chunk-level and recording-level F-scores are calculated.

6.4.1 Classification Results on the IVA$_{3class}$ Dataset

Table 6.2 shows the classification results on the IVA$_{3class}$ dataset. As shown in the table, the performance of the fully connected and attention features do not differ much. For example, for the HC vs. ND task, the SVM based classification F-score on the fully connected feature is 88.21%, and on the attention feature is 88.35% F-score. Compared with the IS10 and the ComParE feature sets, the classification results of the fully connected

Table 6.2: The F-score (%) for chunk-level classification on the IVA$_{3class}$ dataset.

Classifier	Feature	HC vs. ND	HC vs. MCI	MCI+ND vs. HC
LR	ComParE	77.08	68.33	75.15
	IS10	81.34	70.08	77.51
	Fully Connected	**88.39**	78.19	84.26
	Attention	88.15	77.87	84.18
SVM	ComParE	72.58	67.26	70.70
	IS10	78.28	70.04	75.64
	Fully Connected	88.21	**79.27**	84.23
	Attention	88.35	78.88	**84.56**

feature and attention feature are superior for the three classification tasks. Specifically, for the HC vs. MCI task, the best chunk-level classification F-score is 79.27% achieved by the fully connected features classified by SVM, compared with the best baseline result: 70.08% F-score achieved by IS10 classified by LR.

By comparing the results from the three classification tasks, it is found that the hardest classification task is HC vs. MCI. The best result on this classification task is achieved by the fully connected feature classified by the SVM classifier. The easiest classification task is the HC vs. ND, corresponding to 88.39% F-score. The result is consistent with the review in Section 2.1 that the diagnosis of the MCI from HC is more challenging than the ND vs. HC task as the symptoms of the MCI are less obvious.

Table 6.3: The F-score (%) for the recording-level classification on the IVA$_{3class}$ dataset.

Classifier	Feature	HC vs. ND	HC vs. MCI	HC vs. MCI+ND
LR	ComParE	88.09	81.18	81.60
	IS10	93.25	81.60	84.31
	Fully Connected	**98.29**	84.09	**93.18**
	Attention	96.58	85.74	**93.18**
SVM	ComParE	89.83	77.21	77.13
	IS10	93.25	81.28	82.57
	Fully Connected	96.58	85.61	92.06
	Attention	96.58	**87.27**	**93.18**

The recording-level classification results with two baseline feature sets and two data-driven feature sets are shown in Table 6.3. Compared to the results shown in Table 6.2,

the performance of the features at the recording-level is better, but the advantage of the learned features is consistent compared with the baseline features under the same situation at the chunk-level. Specifically, the fully connected features and attention features perform better than the ComParE and IS10 feature sets on the three classification tasks. For example, as shown in the table, the best classification result between the three tasks is on the HC vs. AD task, which is 98.29% F-score. Even for the most difficult classification task between the recording from the HC and MCI, the best result between the four types of features is improved from 79.27% at the chunk-level to 87.27% at the recording-level.

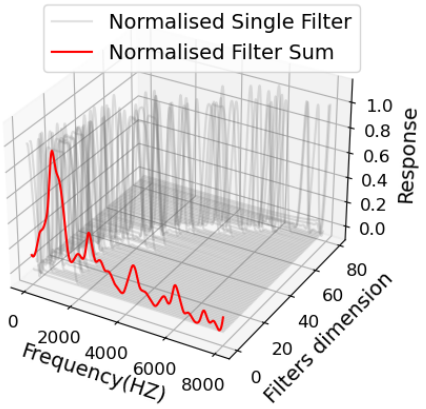
6.4.2 Analysis of SincNet Filters from the IVA$_{3class}$ Dataset

The filters in SincNet is interpretable by exploring the parametrised *sinc* functions. A case study and statistical analysis of the trained filters are shown in this section to better understand the trained filters.

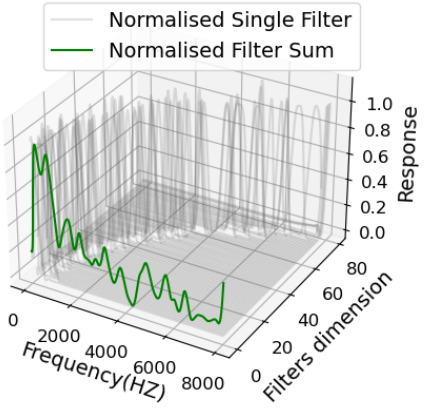
6.4.2.1 Case Study of Learned *Sinc* Filters

For the case study, one out of ten trained SincNet filters is selected (each one corresponding to one CV fold) for each classification task, and the averaged response of the audio chunks is plotted over the same class. The three filters trained for the three classification tasks are shown in the 3-D figures. As shown in Section 6.3.3, there are 80 filters in the SincNet layer, and each filter corresponds to the "Filter dimension" axis. The magnitude frequency response of each trained filter and their cumulative response is shown in the " Frequency (HZ)" axis. The normalised response of each filter is shown in the "Response" axis. In the figures, the grey line is the filter response. The red, green and blue lines represent the Cumulative Frequency Response (CFR) of the trained filters, which is calculated by averaging all the single filters.

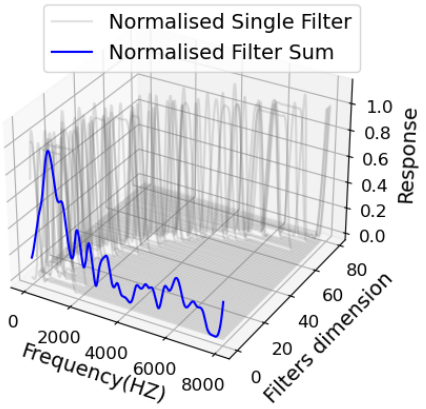
As shown, for all the three classification tasks, more filters are trained to locate the low-frequency bands. The analysis reflects that, for classifying the recordings from people living with or without cognitive impairments, the information embedded in the low frequencies is somehow more important than the high frequencies. The difference between the CFRs learned from the three classification tasks is further analysed in Section 6.4.2.2.



(a) The HC vs. ND task



(b) The HC vs. MCI task



(c) The HC vs. MCI+ND task

Figure 6.2: Visualisation of the learned filters for the three classification tasks.

6.4.2.2 Statistical Analysis of the Learned SincNet Filter Parameters

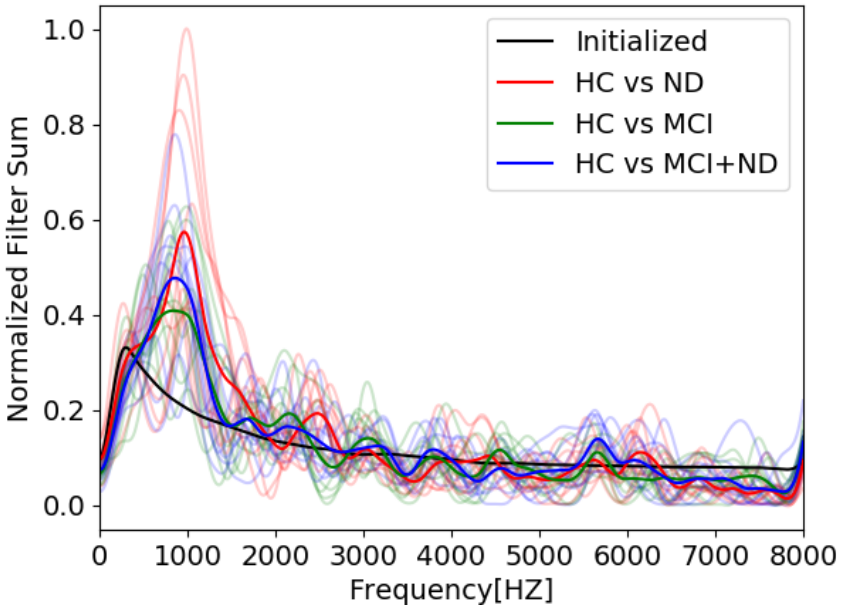


Figure 6.3: Cumulative frequency response of SincNet filters on the three classification tasks; bold lines are the average response for the 10-fold CV and thin lines are the response for every fold trained system.

Similarly to Ravanelli & Bengio [2018a], all the learned filters from the 10-fold CV are shown in Figure 6.3. The initialised and the three learned CFRs of the SincNet layer are plotted. The black line corresponds to the initialised CFR (Mel-scale filter-bank), and the different colored lines refer to different classification tasks after training. The filter sum is normalised with the highest response. The bold lines are the average response for the 10-fold CV, and the thin lines are the response for every fold trained system. By analysis, the conclusions can be summarised as below:

1. By comparing with the initialised CFRs (Mel-scale filter-bank), there is more fluctuation in the CFRs learned from the three classification tasks. It is explained that while training the filters, the information embedded in the input data has been learned. Also, the learned filters show the difference for three different tasks, especially in the low-frequency zone, which means that different classification tasks can influence the desired filter banks. In other words, the Mel-scale filter bank is not perfect for the classification

tasks shown in this chapter. On the other hand, the distinction also reflects that the fluctuation learned by the filters in SincNet is related to the cognitive status rather than the acoustic conditions during recording (e.g. background noise).

2. By observing the CFRs of the three tasks, it can be seen that the frequency responses concentrate on the low frequencies, which is consistent with prior knowledge [Alhanai *et al.*, 2017; Meilán *et al.*, 2014]. However, even though the low-frequency information related features have been included in the IS10 and ComParE feature sets, such as F_0 related features, they cannot achieve as good results as the features learned by our designed feature extractor (shown in Section 6.4).

3. Furthermore, compared with the other two tasks, the CFRs of the low-frequency zone is higher for the HC vs. ND classifier. This may represent that for classifying the more severe symptoms, as seen in the ND vs. HC classification task, more concentration should be put on low frequencies for classification.

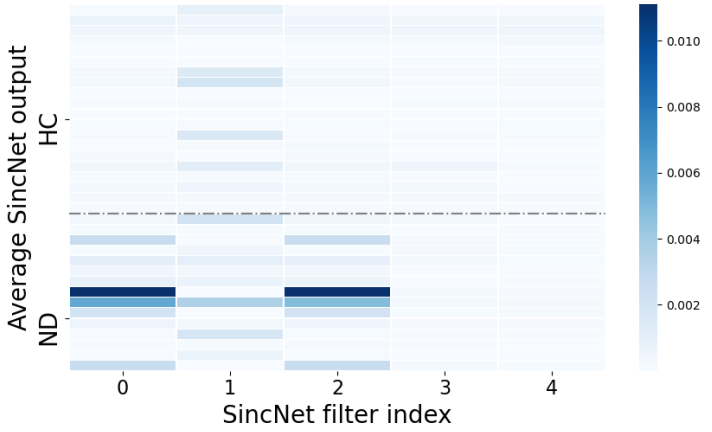
It is found that these conclusions are consistent with the case study in Section 6.4.2.1, which can make the conclusion derived from the case study and the statistical study more convincing.

6.4.2.3 Analysis of the SincNet Layer Outputs

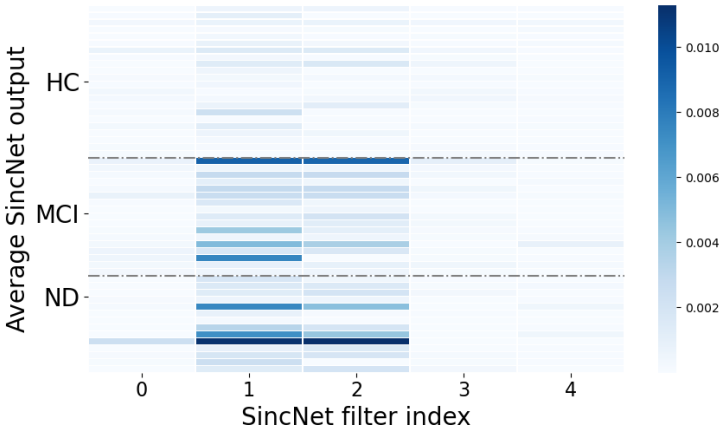
The output of the SincNet layer is a $H \in [frame\_num \times filter\_num]$ matrix, where filter_num equals 80 in our experiment. The analysis of the SincNet output can also help us interpret the frequency-related information better, which may be informative for the clinical cognitive impairments assessment. To understand what information has been output by the SincNet filters, the averaged $filter\_num$ (80) dimensional vector for each recording is calculated by averaging H over the frames. According to the initialisation principle of the SincNet (initialised with Mel-bank filters), the SincNet learned filters are ordered according to the frequency (from low to high using the Mel-scale initialisation).

In Figure 6.4, only the first 5 out of 80 dimensions of the averaged vector is plotted as they are more distinctive than the remaining dimensions. The index of "Filters" refers to the index number of the filters in the SincNet layer. For example, the axis equal to 0 means

(a) The HC vs. MCI classification task.



(b) The HC vs. ND classification task.



(c) The HC vs. MCI+ND classification task.

Figure 6.4: The representation averaged across frames of the first five SincNet filters output; only the recordings from the first fold training set are shown.

the first filter among the 80 SincNet filters. Each row corresponds to the filter response of one recording. The darker the color, the higher the output value. As shown in the figures, the high values and main differences are concentrated on the outputs of the first several filters for the three tasks. Similarly to the description in Martínez-Sánchez *et al.* [2017], the increase in the value of low-frequency ranges can result from the progress of dementia. For example, the output of the SincNet layer in the feature extractor trained for the HC vs. MCI+ND classification task is shown in Figure 6.4(c). As shown, the outputs of the first several filters corresponding to the recordings from ND and MCI are mostly with higher values, compared with the outputs from HC.

6.4.3 Classification Results on the DementiaBank Dataset

The audio recordings from the DementiaBank dataset are also used as the input of the Sinc-CLA feature extractor to test our proposed system's performance on a publicly available dataset. Table 6.4 shows the recording-level classification results. As shown in the table, different from the superior performance achieved by the fully connected feature and attention feature on the IVA_{3class} dataset, the results achieved by the fully connected feature and attention feature are not as convincing on the DementiaBank dataset. Specifically, the best results on the DementiaBank dataset is achieved by the ComParE feature set classified by an SVM classifier (61.51% F-score). In comparison, the best result from the trained features is the 51.49% F-score. The audio recording samples from the two datasets are analysed in Section 6.4.4 to explore the reason for the performance mismatch of the trained features on the IVA_{3class} dataset and the DementiaBank datasets.

6.4.4 Dataset Comparison

In this chapter, the performance of the proposed systems is examined on the IVA_{3class} and the DementiaBank datasets. For exploring the reason for the performance mismatch on the two datasets, two audio samples are selected and plotted in Figure 6.5(a). As shown, the audio recording sample from the DementiaBank dataset has a higher background noise level. Furthermore, when listening to the recordings, it is found that the speech from ADs

Table 6.4: The binary classification F-score (%) result for the recording-level classification on the DementiaBank dataset.

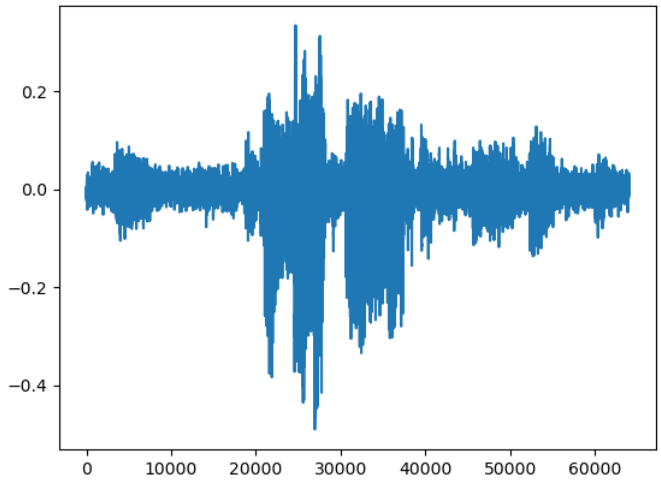
Classifier	Feature	Recording-level
LR	ComParE	**58.74**
	IS10	57.91
	Fully Connected	51.49
	Attention	49.33
SVM	ComParE	**61.51**
	IS10	58.04
	Fully Connected	51.49
	Attention	49.33

tends to include more frequent and longer pauses. The background noise and the pauses increase the difficulties of extracting high-quality acoustic features. In comparison, the audio recordings from the IVA_{3class} dataset include less background noise than those from the DementiaBank dataset.

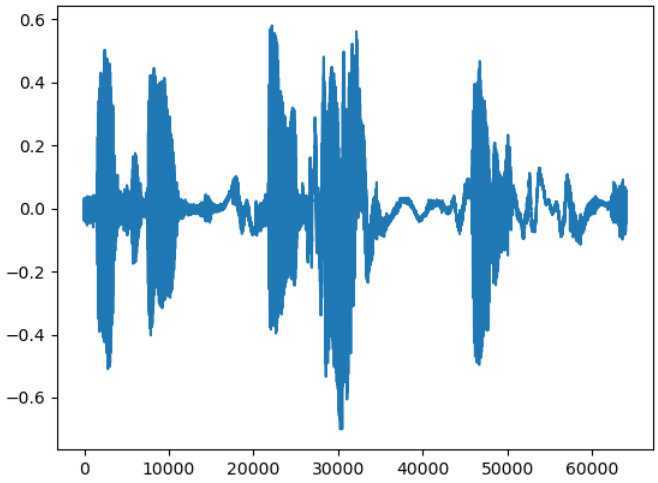
Also, the Signal to Noise Ratio (SNR) is estimated for the two datasets. The SNR for the DementiaBank dataset is -42.33 db and for the IVA_{3class} dataset is -17.64 db. The higher the SNR, the better the audio recording qualities. Compared with the audio recordings in the IVA_{3class} dataset, the audio recordings in the Dementiabank dataset are of lower quality. It may be that a raw waveform based end-to-end system, like the one proposed in this chapter, cannot handle this high noise-level. The next chapter will explore how to extract the high-performing acoustic features from the low-quality audio recordings in the DementiaBank dataset by introducing the medical knowledge into designing and extracting the features.

6.5 Summary

This chapter worked on extracting acoustic information with deep learning technologies for dementia detection. In this chapter, a feature extractor named Sinc-CLA was designed using state-of-the-art deep learning technologies for extracting the trained features from the raw waveform. The target was to classify the recordings from people living with



(a) A 5 seconds recording sample from the DementiaBank dataset



(b) A 5 seconds recording sample from the IVA_{3class} dataset

Figure 6.5: Recording samples from the two datasets used in this chapter.

or without related neurodegenerative disorders (ND, and MCI). In this chapter, three classification tasks were explored using the designed Sinc-CLA feature extractor: HC vs. ND, HC vs. MCI and HC vs. MCI+ND. Compared with the popular feature sets (IS10 and the ComParE), the features extracted from the Sinc-CLA feature extractor achieved superior performance on the three classification tasks with the IVA_{3class} dataset. Sinc-CLA is the first system to extract acoustic information from the raw waveform used for dementia detection.

Analysing the CFRs of the SincNet layer gave us evidence that low-frequency information is critical for classifying the MCI and ND recordings from the HC recordings. The intuition of the learned filters and their output made the result more convincing. However, for the low-quality acoustic recordings (the DementiaBank dataset), the proposed Sinc-CLA cannot achieve satisfactory performance. It might be caused by the poor quality of the audio recordings. The next chapter will explore how to extract high-performing features from low-quality acoustic recordings. In addition, this chapter focused mainly on binary classification. In Chapter 8, the system proposed will work on four different diagnostic classes to better reflect real-world clinical practice.

After this study was published, some other deep learning structures were used to extract acoustic features from raw waves directly. For example, wav2vec2.0 [Baevski *et al.*, 2020] and VGGish [Hershey *et al.*, 2017] were used as the feature extractors for extracting the acoustic features for dementia detection [Balagopalan & Novikova, 2021; Gauder *et al.*, 2021; Pan *et al.*, 2021a; Wang *et al.*, 2021; Zhu *et al.*, 2021]. Compared with the traditional pipeline systems that use the front-end features followed by the classifiers, the deep learning structures can usually be better. In the Interpseech-2020 ADReSS and Interspeech-2021 ADReSSo challenges, the best acoustic-only results were both achieved using deep learning structures [Gauder *et al.*, 2021; Koo *et al.*, 2020]. Inspired by the research in this chapter and the deep learning-based studies reviewed in this paragraph, how to highlight the information embedded in the low frequency of speech when designing the deep neural networks for dementia detection is expected to be explored in future studies.

Chapter 7

High-performing Acoustic Feature Extraction

Contents

7.1 Introduction 121

7.2 Background 121

7.3 Methodology for Acoustic Feature Design 123

7.3.1 Data Analysis 123

7.3.2 Feature Construction 125

7.3.3 Feature Classification 127

7.4 Experimental Setup 129

7.4.1 Pre-processing of the Audio Recordings and Transcripts 129

7.4.2 Classifier Configuration 130

7.5 Results 131

7.5.1 Acoustic-based Results 131

7.5.2 Linguistic-based Result 132

7.5.3 Combined Feature Results 133

7.6 Summary 133

In Chapter 5 and Chapter 6, methods for linguistic-based dementia detection and acoustic-based dementia detection were proposed respectively. In Chapter 6, the method proposed for acoustic feature extraction performs well on the IVA subset (IVA_{3class}) but cannot get a satisfactory performance on the DementiaBank dataset. The comparison between the audio recordings from the IVA_{3class} and DementiaBank datasets in Chapter 6 showed that the quality of the audio recordings from the DementiaBank dataset is lower than the quality of the audio recordings from the IVA_{3class} dataset. As a result, the Sinc-CLA system proposed using state-of-the-art deep learning technologies could not ensure a consistent performance. Therefore, this chapter explores how to extract better high-performing acoustic features from the DementiaBank dataset for dementia detection using medical technologies. This chapter explores the answer to the second research question: "how can the known clinical dementia detection knowledge help in constructing an automatic dementia detection systems and extracting useful features? (**RQ2**)". The structure of this chapter is as follows:

Section 7.1 introduces the research topic and motivation of this chapter.

Section 7.2 summarises the related previous dementia detection methods and existing difficulties on acoustic feature extraction.

Section 7.3 proposes to use the outputs of the ASR system for extracting the high-performing acoustic features for dementia detection.

Section 7.4 contains the information about the experimental setup of our proposed system and the baseline acoustic features.

Section 7.5 summarises the classification results and corresponding analysis of our proposed system.

Section 7.6 contains the summary of this chapter.

7.1 Introduction

As reviewed in Section 2.2, both the linguistic and acoustic abilities can be affected, even in the early stages of dementia. When diagnosing whether a person is living with or without dementia, clinicians generally use both acoustic and linguistic information embedded in the speech to make the diagnosis more robust, as discussed in Section 2.3. Hesitations and unclear pronunciations in the speech from AD are some of the symptoms used for clinical diagnosis. For automatic dementia detection, the collected data is mostly audio recordings, which can be used directly for acoustic feature extraction. However, extracting the linguistic features requires the transcript, which is generated by an ASR system or transcribed manually. An automatic system typically works on audio recordings, and all processing steps must be done automatically, including the speech-to-text transcription by an ASR system and the feature extraction.

In some previous research, both the linguistic and the acoustic information have been used jointly for automatic dementia detection [Campbell *et al.*, 2020; Gónzalez Atienza *et al.*, 2021]. However, audio recordings with high pause rates and/or unclear pronunciation decrease the performance of the ASR system. For an ASR system, the output is not only the words but also the estimated time alignment and a confidence score for each word. Whether the ASR outputs can be used for improving the quality of the extracted features is explored in this chapter. Specifically, this chapter explores how to extract high-performing acoustic features for dementia detection and then combine the extracted acoustic features with the linguistic system proposed in Chapter 5 to improve the dementia detection performance.

7.2 Background

The audio recordings from people living with dementia tend to be challenging to recognise as the word articulation might be 'blurred'. As a result, the output of the ASR system tends to have more word transcription errors and lower confidence scores (more detailed information can be found in Section 7.3.1). Confidence scores are helpful when evaluating the reliability of the automatic transcripts [Coucke *et al.*, 2018; Weiner *et al.*, 2017; Yu

et al., 2010], lending evidence to the benefit of relying more on the high-quality speech segments for automatic cognitive assessment. In previous research, manual or automatic selection of sub-segments of recordings with a relatively *higher* acoustic quality was shown to improve the performance of the extracted features [Luz *et al.*, 2020a; Warnita *et al.*, 2018]. In this chapter, the estimated confidence score is used as a proxy measure for the quality of the spoken segments and the reliability of the recognised words.

As described in Section 2.2.1, *rhythm* is a speech property to do with the temporal organisation of syllables and units composed of several individual phonetic segments such as vowels and consonants. It can be partially described by related statistical parameters such as speech unit duration and the number and duration of pauses. As mentioned in Section 3.2.1, speech rhythm can be affected by dementia and has been used for dementia detection in previous research [Angelopoulou *et al.*, 2018; Ash *et al.*, 2012; Martínez-Sánchez *et al.*, 2017; Pistono *et al.*, 2016; Satt *et al.*, 2014; Skodda & Schlegel, 2008].

To extract the rhythmic parameters, manually identifying the word location is time-consuming and error-prone, especially at scale. Speech and pause duration has been used to design features for dementia detection [Jarrold *et al.*, 2014; Satt *et al.*, 2013, 2014; Weiner *et al.*, 2017; Yuan *et al.*, 2020]. In Satt *et al.* [2014], the words position in a sentence is estimated by the *PRAAT* [Boersma, 2011] for estimating the rhythm-related features (semantic fluency). Similarly, a voice-activity-based speech pause detection method has been used for estimating the pause duration, and speech duration for designing the speech pause-based features [Weiner *et al.*, 2017]. Compared with using a different toolkit for estimating the word location in the speech, the ASR system, which is necessary for automatic audio recording transcribing, can also be used for extracting the rhythm-related features. In Tóth *et al.* [2015], the speech rate and hesitation are estimated by extracting the phonetic level segmentation and annotation from the ASR system. Jarrold *et al.* [2014] also proposed to extract the pause-related features with the ASR system. After this study was conducted, Yuan *et al.* [2020] proposed to use the pause and disfluency annotation as the extra information of the manual transcripts while using the BERT for linguistic information modelling. In this chapter, the benefits of adding rhythm-related features extracted from the outputs of the ASR system is explored.

7.3 Methodology for Acoustic Feature Design

In this section, an acoustic feature set is designed by extracting rhythm-related features and high-performing acoustic features from the audio recordings. Finally, the systems constructed for classification are proposed. First though, an audio recording segment selected from the DementiaBank dataset is analysed to help motivate the proposed approach.

7.3.1 Data Analysis

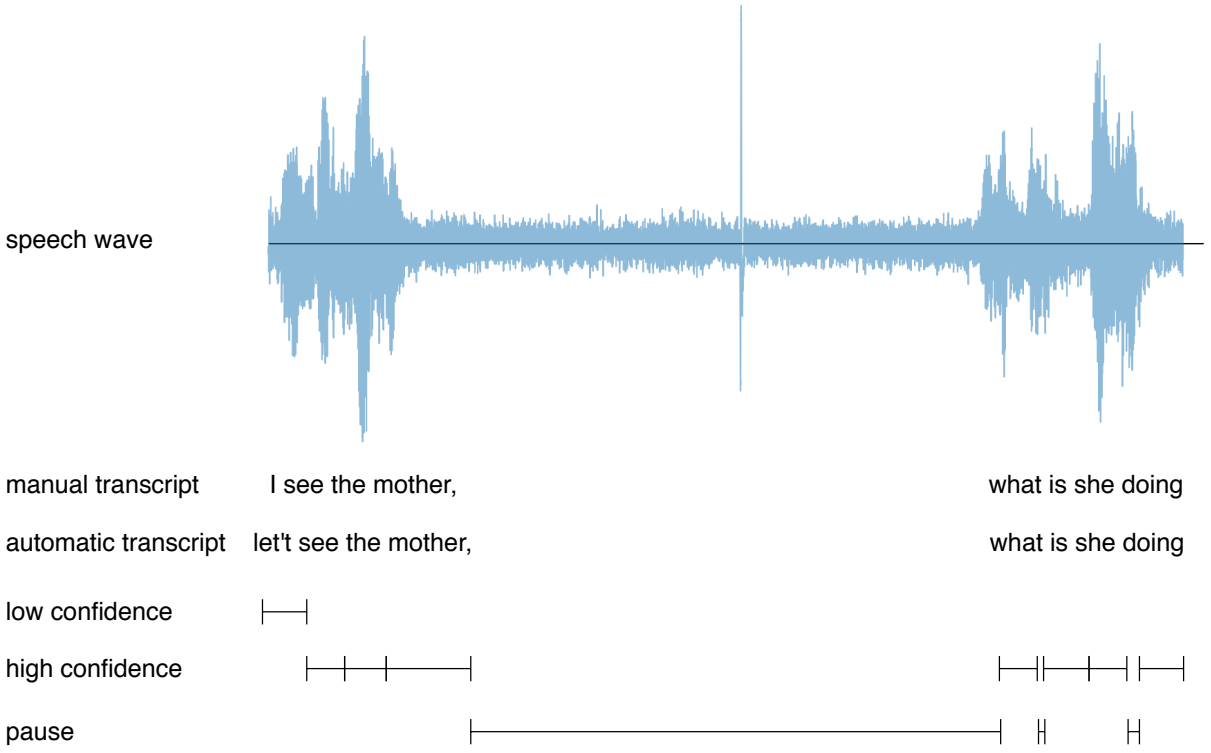


Figure 7.1: A piece of speech segment from the DementiaBank dataset, together with the manual transcript, automatic transcript, estimated confidence scores of each word and the time alignment information; the confidence threshold is set equal to 0.95 for classifying the speech segments into low confidence and high confidence.

A part of the waveform from the DementiaBank dataset is plotted in Figure 7.1, together with its corresponding manual and automatic transcripts. It is clear that the segment contains a lot of noise and has a long pause. Acoustic features are often extracted across the whole signal (speech+pause), although some others have tried to identify and

exclude the pause part [Haider *et al.*, 2020; Luz *et al.*, 2020a; Warnita *et al.*, 2018]. The assumption in this chapter is that the features extracted from the pause segments are of less good quality.

In our research, a confidence score threshold is set first. The pauses and the words with confidence scores lower than the threshold are categorised as *low quality segments*. In the figure, comparing the manual transcript to the ASR automatic transcript, it can be seen that the word *I* has been misrecognised as *let's*, and this word also has a confidence score lower than the confidence threshold. Our approach would exclude these speech segments by using the word alignments and identified by thresholding the confidence score. To get the *high confidence/quality segments*, the pause between two high confidence words is neglected if the duration is shorted than 0.1s, like the pause between *what* and *is*, and between *she* and *doing*. Finally, using this approach, the two speech segments and their corresponding transcribed words such as *see the mother*, and *what's she doing* are selected for further processing.

Before extracting the acoustic features from the audio recordings, it is worth analysing the information output from the ASR decoding of the DementiaBank dataset and how this might vary for the AD and HC groups. The ASR system (more information in Section 7.4) provides word identities, estimated confidence scores of the transcribed words, and word alignments. The information is used to calculate several parameters presented in Table 7.1, including the averaged word duration, the averaged pause duration between words, the averaged number of words in the transcript, and the averaged confidence score of the transcribed words. As shown, both the word and pause duration are longer on average in the AD group than in the HC group, while the number of words per transcript is lower in the AD group than in the HC group, which is consistent with the analysis in Pistono *et al.* [2016]. It has been proposed in previous research that people living with dementia tend to speak slower and also have more pauses as well as longer pauses in their speech than the HCs [Gayraud *et al.*, 2011; Roark *et al.*, 2011; Singh *et al.*, 2001]. Also, the confidence score is higher for the HC group than the AD group, which is likely to be the result of the more unclear pronunciation of people living with dementia. The discrepancy of these parameters indicates that using ASR decoding output might be informative for

classifying the recordings from the AD or HC.

Table 7.1: The average and variance of word duration, pause duration, number of words in the transcript and word confidence scores calculated for the HC and AD groups in the DementiaBank dataset.

Parameter (Mean&Variance)	**HC**	**AD**
Word duration (s)	.535 (0.984)	.642 (1.632)
Pause duration (s)	.545 (1.091)	.663 (1.835)
#Words/transcript	97 (3193)	84 (2476)
Word confidence score	.916 (0.029)	.882 (0.038)

7.3.2 Feature Construction

In previous research, the pause duration, speech duration and the number of words are analysed, and their related information, like the average speech duration and the average number of words in the speech, is used as the acoustic features, such as in König *et al.* [2015]; Roark *et al.* [2011]; Rohanian *et al.* [2021]. That is, in the feature extraction process, the time sequence information in the speech has been discarded. However, it is interesting to explore whether the related time-sequential information embedded in the speech and pause patterns is also informative for dementia detection. Therefore, *include duration*, *exclude duration* and *include word numbers (word index)* are used to locate each word in the speech. For example, in Figure 7.1, the *include word number* of the first word is 0 because the first-word *let's* has a low confidence score, and its corresponding speech segment will not be used for acoustic feature extraction. After calculating the three indexes of each word in the speech, each word is represented with a point in a three-dimensional space, and therefore a track in the three-dimensional space can represent an audio recording.

To visualise the designed three-dimension rhythm-related, 30 recording samples (15 AD and 15 HC) were randomly selected from the DementiaBank dataset and their 30 tracks plotted in Figure 7.2. In this three-dimensional space, $x-axis$, $y-axis$ and $z-axis$ correspond to the *include duration*, *exclude duration* and *include word numbers*. The confidence score threshold (*conf.*) is set as 0.95 (more details in Section 7.4). As

shown in the figure, the tracks of the samples from the two classes are distinctive and the samples from each class are grouped.

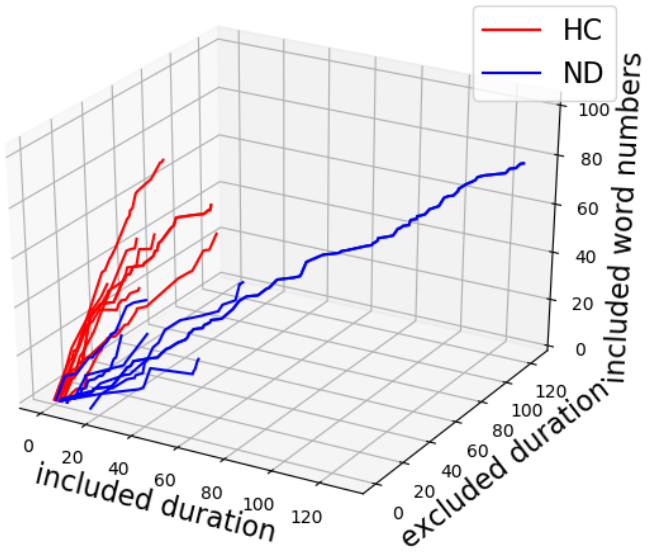


Figure 7.2: Visualisation of the designed three-dimension rhythm-related acoustic features from 30 selected recordings from HC and AD respectively.

By comparing the tracks corresponding to the recordings from the HC and AD, it is found that the recordings from the HC group tend to have a longer speech duration (more talkative) but shorter pause duration than the tracks from the AD group, which is consistent with the known medical knowledge [Banovic *et al.*, 2018; Deb *et al.*, 2007; Turner *et al.*, 1995]. Also, in previous research, it has been mentioned that some people living with dementia may start expressing themselves in a more wordy way as they cannot explain themselves with accurate and concise language [Avila & Porto, 2020]. This phenomenon can also be found in the figure: a very long track corresponds to people living with dementia. However, most of the other tracks corresponding from the AD are shorter than the tracks for the HC because people living with AD tend to speak less as their speech ability declines. Inspired by this phenomenon, the tracks in the three-dimensional space is used as the designed three-dimensional rhythm-related feature. Thus, each speech segment corresponds to a three-dimensional rhythm-related feature vector. As an exam-

ple: the sentence in Figure 7.1 will have two high-quality speech segments selected, so two three-dimension rhythm-related feature vectors are extracted.

To represent the acoustic information of each recording, the IS10, which achieved the best result on the DementiaBank dataset reported in Warnita *et al.* [2018], is adopted as the frame-level feature set. The dimension of the IS10 Low Level Descriptor (LLD) is 76. Considering the varying lengths of the segments, the suprasegmental acoustic feature vector for each speech segment is generated by averaging the frame-level features over the high-quality speech segment. The sentence segments in Figure 7.1 correspond to two 76-dimension IS10 feature vectors.

The 3-dimension rhythm-related feature vector is concatenated with the 76-dimension IS10 feature vector for representing the acoustic information embedded in the speech segments. Finally, each selected high-quality speech segment corresponds to a 79-dimension acoustic feature vector x. Thus, if each audio recording contains T selected high-quality speech segments, it corresponds to a feature matrix $X \in T \times 79$.

7.3.3 Feature Classification

In this section, two feature classifier systems are designed to utilise the extracted acoustic features alone or use both the acoustic feature and linguistic feature jointly in the designed combined system.

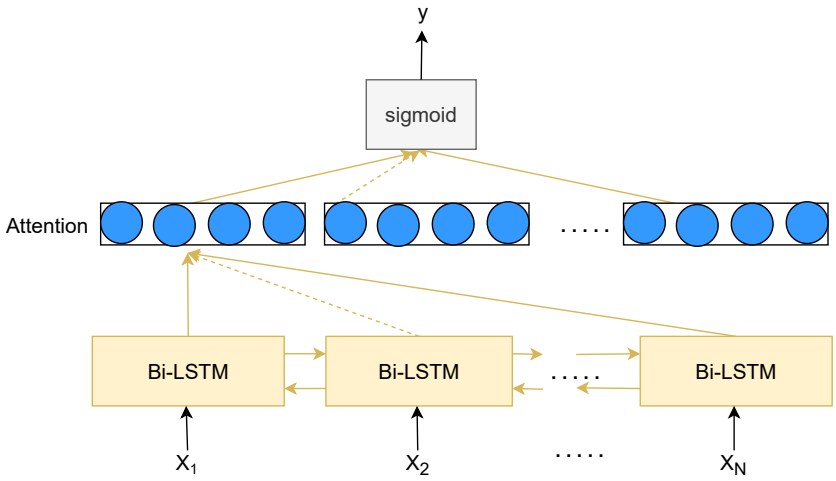


Figure 7.3: The neural network based classifier designed for classifying the extracted acoustic features for dementia detection.

7.3.3.1 Acoustic Feature Classification

The first classifier is composed by a Bidirectional LSTM (BLSTM) with an attention mechanism as shown in Figure 7.3. The BLSTM contains two sub-networks for the forward and backward sequence information modelling. The output of the i_{th} segment h_i in BLSTM is represented by an element-wise sum on the outputs of the two sub-networks as described in Section 5.3.2. Then, the output vector is used as the input of the attention layer. The attention function is defined as in Section 5.3.

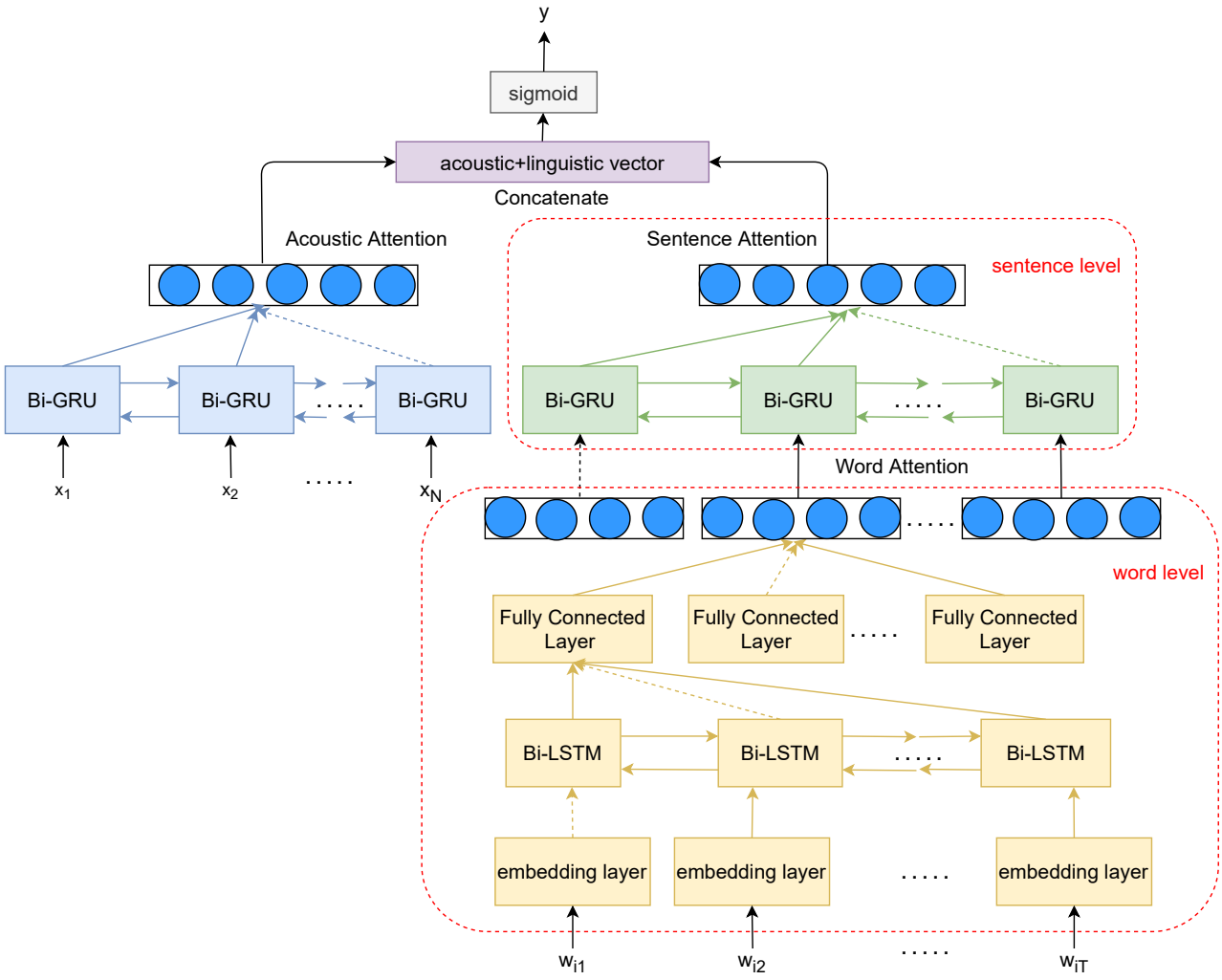


Figure 7.4: The combined system for utilising the acoustic information and linguistic information jointly for dementia detection.

7.4 Experimental Setup

7.4.0.1 Combined Feature Classification

For utilising the linguistic and acoustic information jointly for dementia detection, a combined system is built as shown in Figure 7.4. As shown, the combined system is composed of the linguistic system composed in Chapter 5 and the acoustic system built in Section 7.3.3.1. The learned document-level linguistic and acoustic feature vectors are concatenated before classifying by the final fully connected layer with a *sigmoid* function.

Our experiment is the binary classification based on the DementiaBank dataset. The selected audio recordings are the same as the recordings used in Chapter 5 and Chapter 6. More detailed information can be found in Section 4.1.1 and Section 5.4.1. In this section, the ASR system, the pre-processing procedure of the audio recordings and the ASR transcripts is introduced first. Then, the configuration of the classifiers proposed in Section 7.3.3 is presented.

7.4.1 Pre-processing of the Audio Recordings and Transcripts

To train the ASR system, a 10-fold CV approach is used. 9 folds of the DementiaBank dataset are used for training and 1 fold for the test. The Kaldi's Librispeech [Povey *et al.*, 2011] recipe is followed to train a source Time delay neural network acoustic model. Then the transfer learning technique proposed by Manohar *et al.* [2017] ('transferring all layers') is used for fine-tuning the pre-trained ASR system. The acoustic model is adapted to the data in each fold (following a similar approach to Mirheidari *et al.* [2020] using only one epoch of training to get the best results). The language models are trained using the four-gram models gained from the transcripts in each fold interpolated with the four-gram of the Librispeech data set. To boost both the acoustic and language models, an extra dataset is added to the training set in each fold (the Hallamshire dataset [Mirheidari *et al.*, 2019a]; 64 hours of conversational recordings between doctors and patients). An average 32.3% WER is achieved for the ASRs, compared with the performance of the ASR system in Chapter 5 (41.6% WER). More detailed information can be found in Mirheidari [2018].

As in Chapter 5, though automatically adding punctuation can decrease the performance compared with manual punctuation, it can benefit the results compared with using the ASR transcripts directly without considering the sentence boundary. To add punctuation in the ASR transcripts, the toolkit shared in github is used, and more information can be found in Tilk & Alumäe [2016]. After punctuation, only the words with confidence scores higher than the *conf.* threshold are left in the transcripts. For the acoustic features, the 76 dimensional IS10 frame-level features are extracted with the OpenSMILE toolkit [Eyben *et al.*, 2013] from the selected segments before calculating the mean vector across time. In addition, the 3-dimension rhythm-related feature is extracted from each segment.

7.4.2 Classifier Configuration

For each fold, the systems described in Section 7.3.3 is trained with an 8-fold training set and evaluated on the 1-fold development set. For training, the epoch is set to 20. The best model is selected based on the F-score of the development set. All the results reported in our experiment are averaged across the 10-fold of the test set. The number of the BLSTM unit is set to 50, and the attention layer dimension is set to 10. The batch size is set to 20. The maximum number of segments in each recording is set to 50, and the recordings with fewer than 50 segments are zero-padded. Dropout with a rate of 0.5 is applied after the BLSTM and attention layer to avoid over-fitting. The network is optimised using *Adam* optimizer [Kingma & Ba, 2014] with the L2 regularization ($\lambda = e^{-6}$). A random seed is used. All the parameters above are set according to the performance of the system on the evaluation set. In the combined system, apart from using a different number of epochs (30) and dropout rate (0.3), all other parameters are kept unmodified as in the hierarchical attention network in Chapter 5 and the acoustic system described in Section 7.3.3.1. All the parameters are tuned according to the averaged F-score of the 10-fold development set.

7.5 Results

In this section, the results of the acoustic-only system is presented first. Then, the performance of the combined system utilising both acoustic information and linguistic information is presented.

7.5.1 Acoustic-based Results

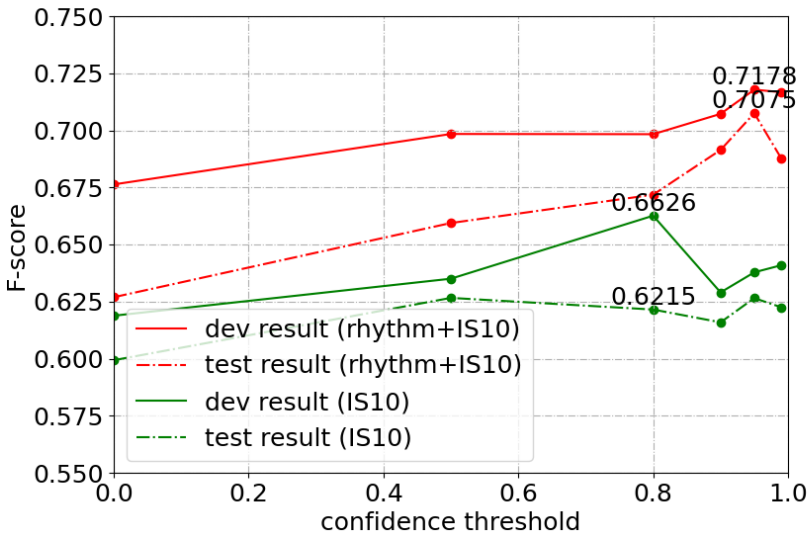


Figure 7.5: The relationship between the F-score and word confidence threshold on the evaluation and test set.

For evaluating the influence of changing the confidence score threshold on the classification performance, the relationships between the threshold and F-score for the development and test sets are shown in Figure 7.5. The 76-dimension IS10-only feature set and the 79-dimension combined feature set (the combined IS10 and rhythm-related features) are evaluated. The confidence threshold is set to 0.0 means that all the speech segments are used for extracting the IS10 acoustic features. The conclusions are summarised as follows:

1. The F-score with IS10-only corresponding to 0.0 threshold is 60.00% F-score on the test set, compared with 57.91% F-score with the LR classifier and 58.04% F-score with the SVM classifier in Section 6.4.3 (Chapter 6). The audio recordings selected with a

0.0 threshold means that all the speech segments are selected using the time alignment information without considering the confidence scores. The results show the efficiency of our proposed classifier and the application of time alignment information estimated by the ASR system.

2. Comparing the red and green lines with the same threshold shows the efficiency of the proposed rhythm-related features. Using the rhythm-related features as the extra information, the best evaluation result improves from 66.26% F-score to 71.78% F-score on the evaluation set, and the corresponding test set result improves from 62.15% F-score to 70.75% F-score.

3. By comparing the results from the same feature sets, it is found that increasing the threshold within a specific range can benefit the classification performance of the features. For example, the best F-scores on the development set are achieved when the threshold increases to 0.8 and 0.95 on IS10 and IS10+rhythm-related feature sets, respectively (66.26% F-score and 71.78% F-score).

7.5.2 Linguistic-based Result

The results in Figure 7.5 show the efficiency of the proposed approach on the acoustic features extraction. For testing whether the confidence score can be used for selecting the transcribed words before being used for extracting the linguistic information, the hierarchical system proposed in Chapter 5 is further explored in this chapter. In the following experiment, the confidence score threshold is fixed at 0.95.

The result with the proposed hierarchical system is shown in Table 7.2. Firstly, the transcripts with the lower WER generated by the ASR system described in Section 7.4.1 is used as the input of the hierarchical system. In comparison with the result reported in Section 5.5.1, the F-score with the new transcripts is absolutely increased by 1.18% (from 74.37% F-score in Chapter 5 to 75.55% F-score). After applying the segment selection approach proposed in this chapter, the system using the transcripts with high confidence words only achieves an F-score of 77.25%. For the automatic transcripts, the words with low confidence scores are more likely to be recognised mistakenly, resulting in ambiguity.

The result demonstrates that the words with low confidence scores can decrease the quality of the transcripts for classification.

Table 7.2: The F-score (%) of the Linguistic-based system on the DementiaBank dataset.

Transcript	F-score (Development)	F-score (Test)
Original	78.92	75.55
High quality selected	84.75	77.25

7.5.3 Combined Feature Results

The performance of the combined system proposed in Section 7.4.0.1 is also tested. After the combination, the system achieves a 78.34% F-score, which is better than the acoustic-only system (70.75% F-score) or the linguistic-only system (77.25% F-score). Compared with previous research, though better results were reported in Fraser *et al.* [2016]; Yancheva & Rudzicz [2016]. However, those systems were not fully automatic, and instead used manual transcripts thus avoiding the effect of erroneous ASR transcripts. In comparison, previous studies that did not use the manual transcripts for linguistic information extraction achieved a 67.21% F-score (Luz [2017]) and 62% F-score (Luz *et al.* [2020a]) which is lower than results presented here.

7.6 Summary

The data for training the speech-based dementia detection systems are usually audio recordings. An ASR system is inevitable for constructing an automatic linguistic-based dementia detection system. In this chapter, inspired by medical knowledge, a three-dimensional rhythm-related feature was designed using the ASR decoding outputs, including the confidence scores and time alignment information.

Before designing the three-dimensional rhythm-related feature, the difference between the recordings collected from the HC and AD on word duration, pause duration, and word confidence scores estimated by the ASR system is analysed. Visualising the three-dimensional rhythm-related feature showed that the designed feature is distinctive for

classifying the recordings from HC and AD.

To extract high-performing acoustic features from low-quality audio recordings, the confidence score threshold is used for selecting high-quality speech segments. The result showed that increasing the confidence score threshold within some range for high-quality speech segments selection was efficient for improving the performance of the extracted acoustic features. Then, The high-performing acoustic features extracted from the high-quality speech segments are combined with the designed rhythm-related features for further improving the acoustic system's performance on dementia detection. Finally, an automatic system was composed by utilising the acoustic and linguistic information for dementia detection on the DementiaBank dataset. Compared with the study reported in Chapter 6, the study proposed in this chapter can provide better performance on the DementiaBank dataset.

Chapter 8

Multi-class Classification for Dementia Detection

Contents

8.1 Introduction 137

8.2 Background 138

8.3 System Construction 139

8.3.1 Using Low- and High-quality Speech Segments 140

8.3.2 Adding Traditional Features 141

8.4 Experimental Setup 144

8.4.1 Evaluation Setting 144

8.4.2 Model Configuration 145

8.5 Results 146

8.5.1 Exploring the best speech input to the system 146

8.5.2 Results with the TR-1 Feature 148

8.5.3 Results with the TR-2 Feature 149

8.5.4 Results Comparison 150

8.6 Summary 151

In clinical practice, it is of great importance to be able to distinguish all the major diagnostic categories like MCI, FMD and ND from the HCs as the treatment of FMD, MCI and ND is different. However, the automatic classification of the four categories is difficult due to the very similar symptoms in people's speech for the different groups. Chapter 5, Chapter 6 and Chapter 7 concentrated on the binary classification. Considering the clinical practice, this chapter explores how to design the end-to-end system for classifying the recordings from the MCI, FMD, ND and HC groups, which corresponds to the first and third research questions:"how can state-of-the-art deep neural networks be applied for speech- and language-based dementia detection? (**RQ1**)" and "how to design a framework for more clinically relevant diagnostic scenarios? (**RQ3**)". In this Chapter, a balanced set of 60 recordings (15 FMD, 15 ND, 15 MCI and 15 ND) is chosen for the study, and named as the IVA_{60} Dataset. A feature extractor is designed for improving the performance of clinical practice classification tasks on the IVA_{60} dataset. The structure of this chapter is as follows:

Section 8.1 introduces the clinical practice and the defined classification tasks.

Section 8.2 presents the related research background of the multi-class classification and the IVA_{60} dataset.

Section 8.3 proposes the end-to-end feature extractor.

Section 8.4 contains the information about the experimental setup of the proposed system and the baseline system.

Section 8.5 summarises the classification results and the analysis of the proposed system.

Section 8.6 contains the summary of this chapter.

8.1 Introduction

Functional Memory Disorder (FMD) describes a condition where an individual can experience poor memory function that is not caused by dementia but by poor concentration or inability to cope with too much information [Schmidtke *et al.*, 2008]. In general, people who go to their GP with worries about their memory often have FMD but the GP lacks the expert knowledge to diagnose this and rule out dementia. They, therefore, refer these patients to secondary care resulting in expensive tests. Therefore, automatic methods for distinguishing FMD from MCI and ND is of great interest. Before being diagnosed with ND, people with early signs of cognitive decline often get diagnosed with MCI. As mentioned in Chapter 2, dementia can be caused by different kinds of Neurodegenerative Disorder (ND), like Alzheimer's Disease (AD) and Parkinson's Disease. More detailed information can be found in Section 2.1.1.

The classification between the MCI, HC, ND and FMD groups can be regarded as a four-class classification task (four-way). In the clinic, making a four-way diagnosis is very difficult considering the similar symptoms between the MCI and ND and FMD. In practice, the four-class classification task can be replaced by a three-class classification (three-way) task, which regards HC and FMD as one class because people living with FMD and HCs do not have dementia. Moreover, the MCI and ND can also be regarded as one class (two-way). One reason is that they share similar symptoms, but more importantly, the MCI+ND vs HC+FMD is the distinction between those that need to be referred to secondary care and those that do not. In this study, all of the three scenarios described above are summarised in Table 8.1 and adopted for exploring the designed feature extractor.

Table 8.1: The three scenarios designed for classifying the audio recordings from the HC, FMD, MCI and ND.

Name	Scenarios
Four-way	HC vs. FMD vs. MCI vs. ND
Three-way	HC+FMD vs. MCI vs. ND
Two-way	HC+FMD vs. MCI+ND

Chapter 6 and Chapter 7 proposed two methods for extracting acoustic features for

the binary classification on the IVA subsets and the DementiaBank dataset. In Chapter 6, the acoustic features were extracted from the raw waveform directly using the designed Sinc-CLA end-to-end system. The method proposed in Chapter 7 demonstrated that utilising the time alignment information and word confidence scores estimated by the ASR system implemented for audio recording transcription enables the extraction of high-performing features for dementia detection. Compared with the binary classification studies in Chapter 6 and Chapter 7, the multi-class classification task is more complicated. Therefore, an end-to-end feature extractor is designed to extract acoustic features for multi-class classification tasks in this chapter inspired by the studies in Chapter 6 and Chapter 7.

8.2 Background

Similarly to Section 7.1, the output of the ASR system is used to calculate a number of parameters as presented in Table 8.2 for the IVA_{60} dataset. As shown in the table, the averaged *word duration* and *pause duration* of the recordings from the HC group are shorter than the two parameters from the other three classes, and the variances of the *word duration* and *pause duration* are also lower than the corresponding variances from the FMD, MCI and ND groups. It demonstrates that the duration of speech and pause from the HC group is more stable than the other three categories, and the averaged pronunciation of the word is shorter, which is consistent with the review in Chapter 2.

Also, the automatic transcript from the MCI group includes the fewest number of words, which is 464 on average. In comparison, HCs are most talkative, followed by the ND group, whose automatic transcripts include 805 words per transcript on average. It is inferred that people living with ND tend to have word-finding difficulties, resulting in more frequent self-repair [McNamara *et al.*, 1992] and empty speech [Nicholas *et al.*, 1985b]. In the table, the averaged word confidence score from the HC group is the highest at 0.936. It is inferred that the words produced by HCs are the clearest, so they are the easiest to recognise by the ASR system. The averaged word confidence scores from the MCI, ND and FMD groups are very similar.

The duration mismatch between short and long segments of speech causes difficulties to the feature modelling, which is also a difficult research topic in speaker identification [Kye *et al.*, 2020; Ma *et al.*, 2016a; Sarkar *et al.*, 2012]. In Ma *et al.* [2016a], a twin model is proposed for modelling the short and long utterances separately depending on the distribution mismatch of the extracted features. Similarly, in dementia detection, Fritsch *et al.* [2018] proposed to train two language models for the AD and HC groups, respectively, depending on their linguistic mismatch. The system proposed in Chapter 7 discarded all the audio segments with confidence scores lower than the confidence score threshold. However, the audio recordings for assessing cognitive decline are limited, and an end-to-end system is a data-driven system that requires a large amount of data for the system training. Discarding the segments with low confidence scores could be problematic considering the limited available data for the end-to-end system training. Also, the mismatch between the low- and high-quality speech segments might be informative for dementia detection. Therefore, in this chapter, the low- and high-quality speech segments are used separately as the input of the designed feature extractor. The hypothesis is that the distribution of the low- and high-quality speech segments is different. To take the mismatch into consideration in our system, a twin model is constructed in Section 8.3.

Table 8.2: The parameter analysis for the recordings and automatic transcripts of the HC, MCI, FMD and ND groups in the IVA_{60} dataset.

Parameters (Mean&Variance)	**HC**	**MCI**	**ND**	**FMD**
Word duration (s)	0.518±(0.547)	0.744±(1.859)	0.719±(2.073)	0.718±(2.047)
Pause duration (s)	0.017±(0.019)	0.037±(0.077)	0.034±(0.045)	0.037±(0.064)
Confidence score	0.936±(0.023)	0.904±(0.034)	0.903±(0.034)	0.900±(0.034)
#Words/transcript	822±(123884)	464±(165498)	805±(797407)	650±(219237)

8.3 System Construction

In this section, a feature extractor is first designed to extract the acoustic-based trained features (named as the **TR-1** features) from the low- and high-quality speech segments.

Then, the traditional features and trained features are combined using the feature extraction system to extract the combined trained features (named as the **TR-2** features). Finally, both the TR-1 features and TR-2 features are used for the three scenarios introduced in Section 8.1.

8.3.1 Using Low- and High-quality Speech Segments

A twin-CCLA end-to-end feature extractor system is designed as in Figure 8.1. The system is designed for representing the mismatch between the low- and high-quality speech segments separately. As shown, the designed system is composed of two CNN blocks, a bi-directional LSTM layer and an attention layer in each branch.

Figure 8.2 is plot to investigate whether the distribution of the low- and high-quality speech segments is different. The analysis is based on the four-way scenario. Firstly, the vectors are output by the attention layers (100-dimension) with the low- and high-quality speech segments as the inputs. The two attention vectors are dimension reduced into two-dimension vectors by the Principle Component Analysis (PCA) algorithm [Wold *et al.*, 1987]. To inspect the distribution of these vectors they are plotted in Figure 8.2. Each point in the figure represents the extracted vectors after dimension reduction. The figure shows that the distribution of the processed vectors of high and low confidence scores are clustered within the same group and distinguished between the groups. It demonstrates the hypothesis: *the distribution of low- and high-quality speech segments is different.*

For utilising the distribution mismatch of the features learned from the speech segments with low and high confidence scores, the Euclidean distance between the vectors' output of the attention layer is calculated. Specifically, the two vectors output from the attention layers are concatenated with the 1-dimension Euclidean distance vector to work as the input of the fully connected layer. After training, the TR-1 is extracted from the 64-dimension fully connected bottleneck layer by inputting the audio recordings, as shown in Figure 8.1.

The calculated Euclidean distances and their absolute distances for the four-way scenario are shown in Figure 8.3. As shown, the distributions of the Euclidean distances for the HC, FMD, MCI and ND groups are different, indicating the learned Euclidean

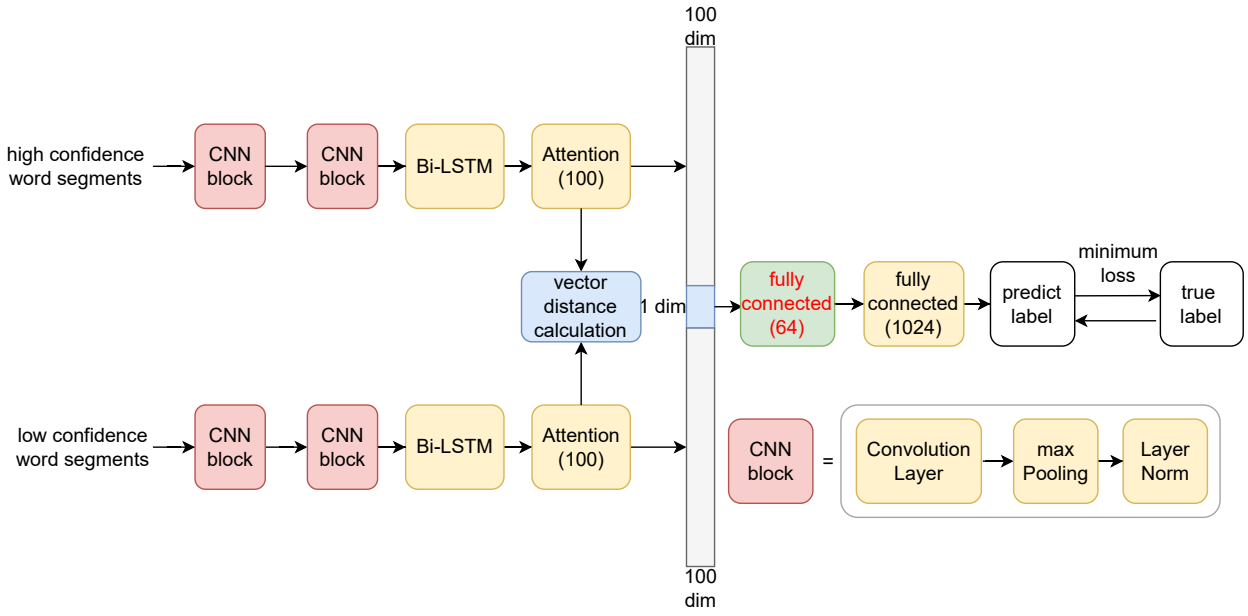


Figure 8.1: The structure of the designed twin-CCLA feature extractor for extracting the TR-1 feature.

distance is distinctive when being used as a 1-dimension feature. As shown, the absolute peak distance from the MCI group is the largest, meaning the distribution of the features learned from the low and high confidence scores has the most significant mismatch among the four classes. The Euclidean distance's effect on the classification accuracy is also examined in Section 8.5.1.

8.3.2 Adding Traditional Features

As discussed in Section 2.2, both speech and language abilities can be affected by dementia. In the previous research (reviewed in Section 3), the combination of speech and language features for dementia detection has shown their efficiency. The studies in Chapter 7 have also demonstrated that the combination of using the acoustic and linguistic information can improve the system's performance compared with using the information embedded in the speech and language separately.

In Mirheidari *et al.* [2017], some features are proposed inspired by conversation analysis (20-dimension, named as CA) and linguistic information (7-dimension, named as WV-PCA) on the IVA_{60} dataset and have shown their efficiency. The feature list is shown

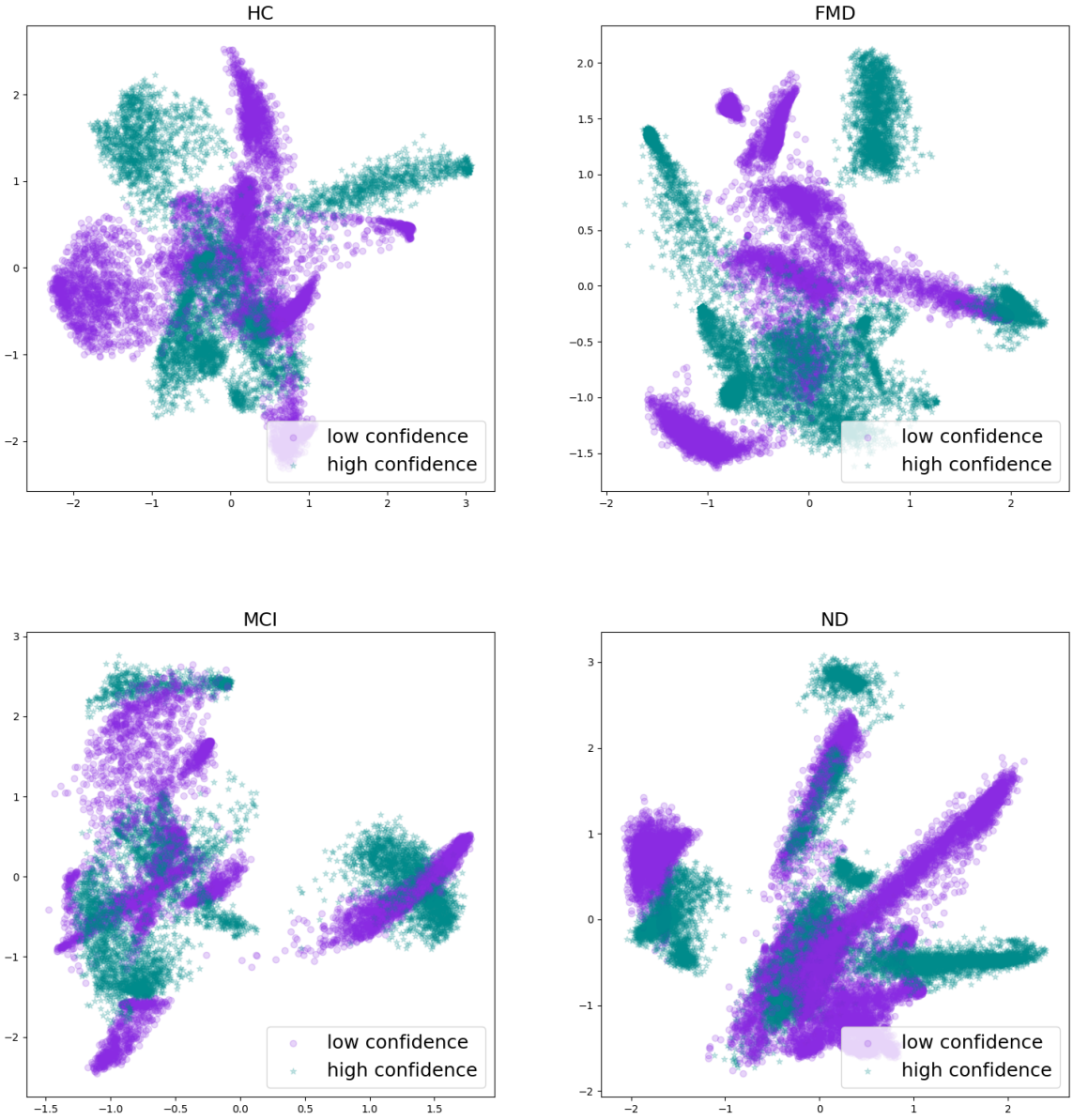


Figure 8.2: Comparison of the extracted attention vectors from the low- and high-quality speech segments of the HC, FMD, MCI and ND.

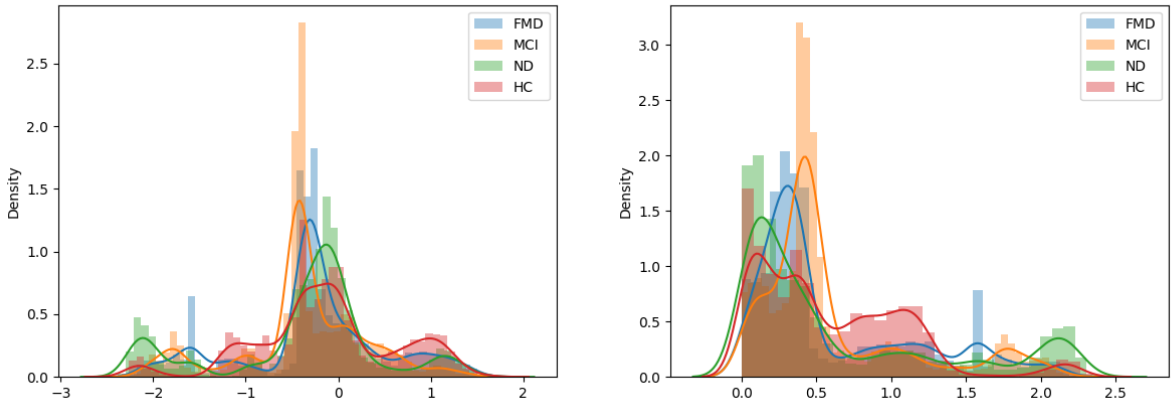


Figure 8.3: The Distribution of the Euclidean distance between the vectors' output of the attention layer from the four-way scenario.

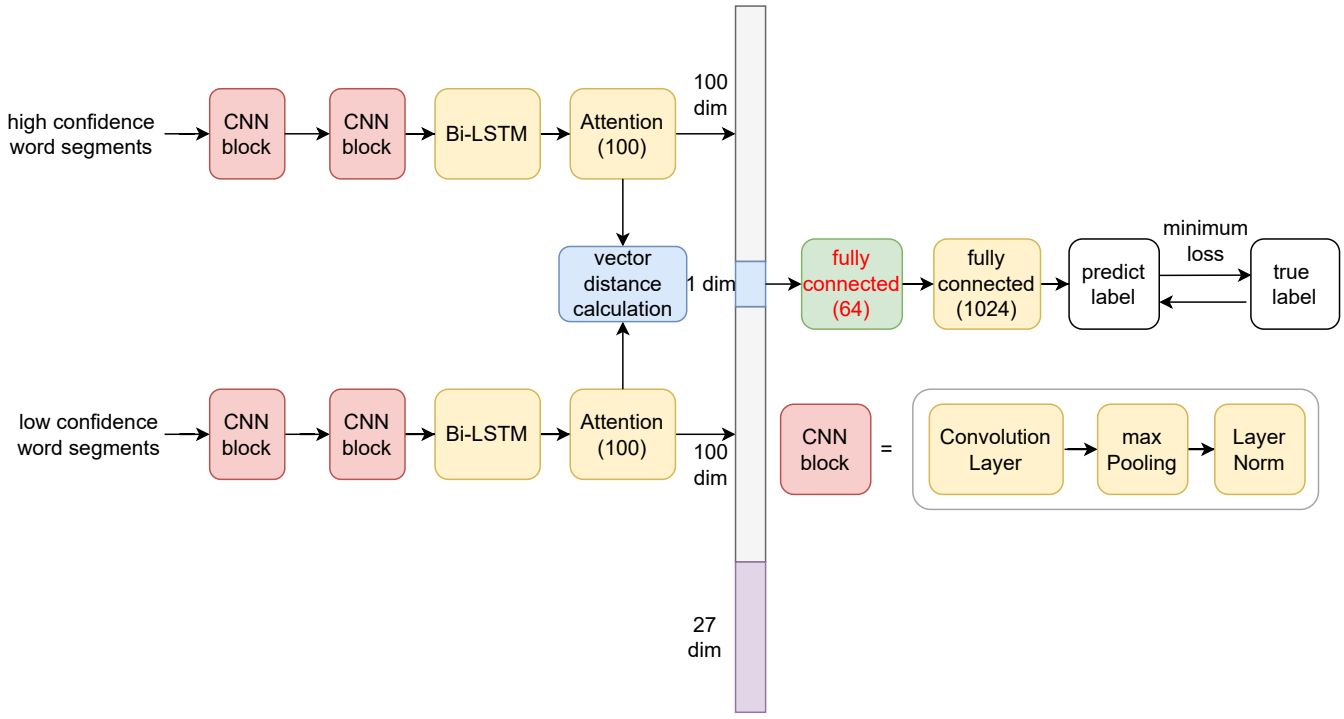


Figure 8.4: The structure of the feature fusion system designed for extracting the TR-2 feature.

in Appendix A. To combine with the proposed system, as shown in Figure 8.1, the 27-dimension feature vector is concatenated with the 201-dimension acoustic feature output from the attention layer. Then, TR-2 is extracted from the 64-dimension fully connected layer as shown in Figure 8.4.

8.4 Experimental Setup

In this section, the experimental setup is provided. The evaluation setting of the dataset is summarised first. Then, the model configuration of the proposed feature extractor shown in Section 8.3.1 and Section 8.3.2 is presented.

8.4.1 Evaluation Setting

As mentioned at the beginning of this chapter, the IVA_{60} dataset is used. The IVA_{60} dataset includes the same number of recordings from the MCI, HC, ND and FMD. More detailed information about the dataset can be found in Section 4.2.3.

To provide a reliable result, the 10-fold CV is used on the relatively small dataset, and each fold is fixed for all the experiments presented in this chapter. The number of recordings in the three partitions (training set, evaluation set, and test set) of each fold is as balanced as possible in terms of the diagnostic category. To evaluate the extracted features, a typical classification pipeline is used. The trained system is selected on the 1-fold evaluation set after training with the 8-folds training set. After training, the 64-dimension feature output from the fully connected layer is used as the front-end feature. LR, one of the most commonly used classifiers in acoustic-based cognitive decline detection fields, is adopted as the back-end classifier. For testing the performance of the trained features with the pipeline system, the features extracted from the training and evaluation sets (9 folds) are used to train the back-end classifier, and the test set is used for evaluation.

In the twin-CCLA end-to-end systems, the parameters that need to be learned while training is initialised randomly, which can cause a variation in the final results. To ensure the comparisons are fair, all the results are averaged over the 5 sessions. The result for each session is averaged across the 10 folds' test sets. Both the mean and the variance of

the results over the 5 sessions are presented in Section 8.5.

As in Chapter 7, 0.95 is used as the *confidence score threshold (conf.)*. The words with a confidence score higher or lower than 0.95 are referred to as the *high confidence words* and *low confidence words*. Segments with high/low confidence scores are audio recordings composed of high/low confidence words, respectively. To get the segments with high/low confidence scores, as in Chapter 7, the pause between two high/low confidence words is neglected if the duration is shorter than 0.1s.

8.4.2 Model Configuration

The audio recording from the IVA dataset includes the speech from a neurologist, a participant and accompanying person(s). Only the audio recordings from the participants are extracted and concatenated into an audio recording. The average duration of the recordings in the IVA_{60} dataset is about 9 minutes, which is too long to be utilised directly as the input of the designed feature extractor. Therefore, similarly to the procedure used in Chapter 6, each recording is cut into 2 seconds chunks. Each chunk is assigned a label corresponding to its diagnostic category. Finally, the segments corresponding to the words with high and low confidence scores are concatenated respectively for being used as the input of the twin-CCLA system.

Similarly to Chapter 6, the first convolutional layer is composed of N=80 filters of length L=125 samples. The second convolutional layer uses 60 filters of length 5. The max-pooling size of the two convolutional layers is 3. The number of units in the BLSTM is 50. The output of the BLSTM layer is the 100 dimensional feature, which is the concatenation of the two 50 unidirectional LSTM outputs. The dimension of the attention matrix is set as 30. The output of the attention layer is a 100 dimension vector. The two fully connected layers comprise 64 and 1024 neuron units, respectively. In the model, all hidden layers use the Leaky-ReLU non-linearities [Maas *et al.*, 2013], and the *rmsprop* [Tieleman & Hinton, 2012] is applied as the optimizer with a learning rate of 0.001. While training, the mini-batch size is set to 100 and the epoch to 80. All the parameters of the network are selected according to the performance of the development set. The F-score is used as the criteria. After the feature extractor is trained, all the parameters are fixed,

and the 2 second chunks are categorised with confidence scores before inputting into the feature extractors for generating the trained features.

The ASR system in this chapter is based on the system described in Mirheidari *et al.* [2020]. The system is built with a Kaldi script. The language models for the ASR are trained as four-grams with Turing smoothing interpolated with the four-gram language model from the Librispeech dataset [Panayotov *et al.*, 2015] (60% weight for the train set and 40% weight for the Librispeech language model), then using the "transferring all layers" technique [Manohar *et al.*, 2017] for training an adjusted language model. The WER is 25.1% for the IVA_{60} dataset.

8.5 Results

This section includes three parts. First, the results from the convolutional-CLA (CCLA) and twin-CCLA feature extractors are presented in Section 8.5.1, which is aimed at investigating what kinds of audio segments should be used as the input of the feature extractors. Then, the results from the pipeline systems which use the trained features are presented in Section 8.3.1 and Section 8.3.2. Finally, the analysis of the designed systems and comparison of the results from different trained features are shown in Section 8.5.4.

8.5.1 Exploring the best speech input to the system

The experiment is designed based on the four-way classification task to investigate what type of segments can improve the system's performance. The CCLA, which is the same as the Sinc-CLA proposed in Chapter 6 excepted for the first layer, is used as the baseline system for comparison. Compared with the performance of the system using the SincNet as the first layer, the system using the CNN as the first layer is 0.65% absolute better for the four-way scenario on the IVA_{60} dataset. Therefore, in this chapter, the CNN rather than SincNet is selected as the first layer of the systems.

The experimental results of using the CCLA feature extractor are shown in Table 8.3. For the CCLA system, it has three different kinds of inputs: *full waveform* refers to the audio segments without pre-processing in the IVA_{60} dataset; *Speech only* refers to both

the low- and high-quality speech segments. The speech only segments are generated by using the time alignment information for selecting the segments without the silence and long pauses. The *High quality* corresponds to the high-quality speech segments, which is achieved by the same process as in Chapter 7. In other words, only the speech segments with confidence scores higher than the threshold are selected as the input.

Table 8.3: Performance of the CCLA feature extractors designed with different types of waveform input on the four-way classification task. *full waveform*: the audio segments without pre-processing; *speech only*: low- and high-quality speech segments; *high quality*: the high-quality speech segments.

System Input	Accuracy(%)	Precision(%)	Recall(%)	F-score(%)
Full waveform	40.34±(2.17)	39.17±(5.11)	40.34±(2.17)	38.05±(2.88)
Speech only	44.16±(5.70)	41.42±(3.96)	44.17±(5.69)	41.58±(4.55)
High quality only	38.90±(2.79)	38.90±(5.01)	42.33±(2.79)	39.62±(4.00)

Table 8.4: Performance of the twin-CCLA feature extractors designed with different types of waveform input on the four-way classification task.

System Input	Euclidean	Accuracy(%)	Precision(%)	Recall(%)	F-score(%)
Low&high quality	Without	41.33±(1.39)	41.03±(2.65)	41.33±(1.39)	40.18±(1.89)
	With	44.69±(5.05)	44.30±(4.31)	44.67±(5.05)	**42.75**±(3.95)

By comparing the results from the CCLA with different kinds of input, it shows that removing the silence and long pauses in the audio recordings can improve the four-way classification performance (increasing the F-score from 38.05±(2.88)% to 41.58±(4.55)%), but removing the segments with low confidence scores does not result in further improvement in the F-score. Instead, the F-score drops to 39.62±(4.00)%.

For the twin-CCLA system, the low- and high-quality speech segments are used as the input separately. For exploring the efficiency of the Euclidean distance, the performance of features extracted from the system with or without Euclidean distance is presented in Table 8.4. As shown, using the Euclidean distance can improve the F-score from 40.18±(1.89)% to 42.75±(3.95)%. By comparing with the results in Table 8.3, using the low- and high-quality speech segments separately can improve the F-score from 41.58±(4.55)% to 42.75±(3.95)%. The result demonstrates that modelling the low- and high-quality speech segments separately can improve the efficiency of the learned features.

8.5.2 Results with the TR-1 Feature

The performance of the pipeline system with TR-1 as the front-end feature and LR as the back-end classifier is presented in this section. The three clinical scenarios shown in Table 8.1 are all explored, and the results are shown in Table 8.5. Each row in the table corresponds to a clinical scenario. The IVA_{60} dataset is tested with the same 10-fold CV list as the experimental results shown in Table 2 in Mirheidari *et al.* [2020]. In Mirheidari *et al.* [2020], 8 feature sets were proposed for four-way classification scenario on the IVA_{60} dataset. The best result among the 8 types of feature sets in Mirheidari *et al.* [2020] is 50.7% F-score on the IS09 feature set [Schuller *et al.*, 2009]. As shown in Table 8.5, the result is 47.66±(1.74)% F-score for the four-way classification task, which would rank second in Mirheidari *et al.* [2020] compared to the features with augmentation and rank third compared to the features without augmentation.

Table 8.5: The LR classification results with the TR-1 feature extracted by the Twin-CCLA system with low- and high-quality speech segments as the inputs.

Task	Accuracy(%)	Precision(%)	Recall(%)	F-score(%)
Four-way	47.67±(1.49)	50.47±(2.05)	47.67±(1.49)	47.66±(1.74)
Three-way	53.33±(1.18)	47.34±(1.82)	53.33±(1.18)	48.83±(1.71)
Two-way	78.33±(2.64)	78.57±(2.51)	78.33±(2.64)	78.28±(2.67)

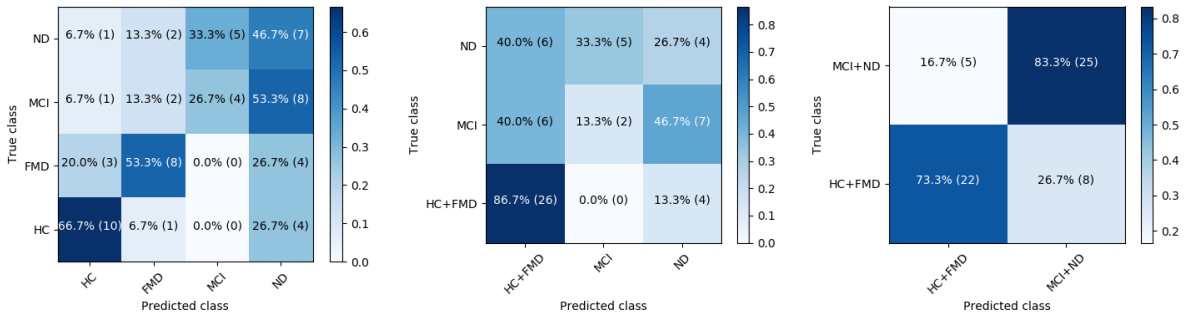


Figure 8.5: The confusion matrices of the classification results from the twin-CCLA extracted TR-1 features classified by the LR classifier.

The confusion matrices for the three clinical scenarios are shown in Figure 8.5. For the three scenarios shown in the sub-figures, the predicted labels are estimated by the system corresponding to the medium F-score among the five sessions in Table 8.5. As shown, for

the four-way classification task, distinguishing between the MCI and ND is difficult. For the three-way classification task, recordings from the ND and MCI tend to be recognised mistakenly as HC+FMD. Compared with the other classes, classifying the recordings from the MCI is the most difficult. Specifically, only 15 out of 4 and 2 recordings are recognised correctly for four-way and three-way scenarios, respectively, which is consistent with the literature review in Chapter 3 that the detection of the MCI is more difficult than the ND detection as the symptoms are less clear.

8.5.3 Results with the TR-2 Feature

The end-to-end based feature extractor described in Section 8.3.2 uses both the combination of raw waveform and the 27 dimension features described in Mirheidari *et al.* [2020] (20-dimension Conversation Analysis (CA) feature together with 7-dimension word vector with the PCA (WV-PCA)) for training the feature extractors. The features trained by the feature extractor are classified with the LR and the results are shown in Table 8.6.

Table 8.6: The LR classification results with the TR-2 feature extracted by the twin-CCLA system with the CA feature, WV-PCA feature, and low- and high-quality speech segments as the inputs.

Task	Accuracy(%)	Precision(%)	Recall(%)	F-score(%)
Four-way	59.00±(1.90)	57.74±(2.62)	59.00±(1.90)	57.66±(2.32)
Three-way	62.33±(3.25)	61.95±(1.82)	62.33±(3.25)	61.99±(2.40)
Two-way	83.33±(2.36)	83.39±(2.32)	83.34±(2.36)	83.33±(2.37)

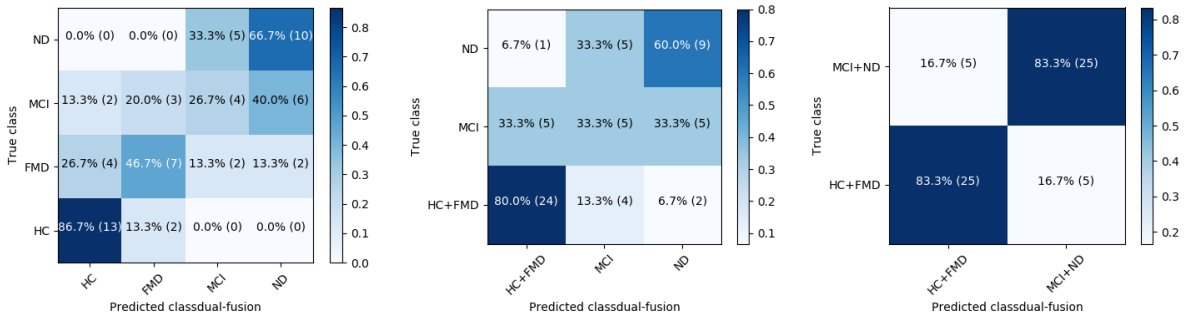


Figure 8.6: The confusion matrices of the classification results using TR-2 features classified by the LR classifier.

As shown, the performances of the TR-2 feature on the three scenarios are all improved

after including the CA feature and the WV-PCA feature for the system training, which can be seen by comparing to the results in Table 8.6. Specifically, for the four-way classification task, the F-score is 57.66±(2.32)%. In comparison, the best out of the 8 feature sets in Table 2 in Mirheidari *et al.* [2020] is the 53.9% F-score before augmentation. Also, the F-score of the 747-dimension combined feature set is 54.9%, which is not as good as the 64-dimension TR-2 feature proposed in our study. However, TR-2 cannot perform as well as the combination feature set after argumentation (the 59.8% F-score).

The confusion matrices for the three scenarios using the TR-2 feature is shown in Figure 8.6. Comparing to the confusion matrices shown in Figure 8.5, for all three scenarios, it shows that fewer recordings from the MCI and ND groups are mistakenly recognised as the recordings from the HC and FMD groups. Therefore, it is inferred that the conversation analysis and linguistic information can help distinguish the MCI and ND from the HC and FMD. Also, for the four-way scenario, after including the conversation analysis and linguistic information, no recordings from the ND group are mistakenly recognised as the HC or FMD groups, and no recordings from HC are mistakenly recognised as the MCI or ND.

8.5.4 Results Comparison

For comparison, the results from Section 8.3.1 and Section 8.3.2 are summarised in Figure 8.7. As shown, the influence of the traditional features is not as significant on the two-way scenario as on the three-way and four-way scenarios. Specifically, the absolute F-score for four-way and three-way is improved by 10% and 13.16% respectively, compared with an absolute 5.05% F-score improvement for the two-way scenario. It is inferred that the acoustic information embedded in the TR-1 feature is more distinctive for the two-way classification task compared with the three-way and four-way classification tasks. In comparison with the two-way classification task, while doing the more difficult three-way and four-way classification tasks, not only the acoustic information, the conversation analysis and linguistic information are also essential for achieving the desirable classification performance.

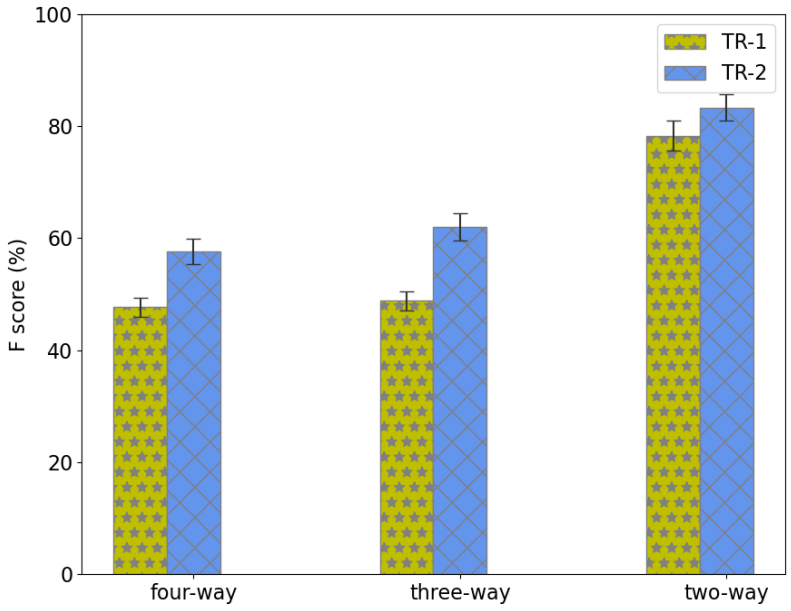


Figure 8.7: The F-score for two-, three- and four-way systems using TR-1 and TR-2 features on the IVA_{60} data.

8.6 Summary

Detecting people living with MCI, FMD, ND from HC is difficult due to the similar symptoms exhibited in the speech by MCI, FMD and ND. However, the distinction between them is of important clinical relevance as the follow-up treatment procedures are different from MCI and ND and FMD. In this chapter, three scenarios were designed considering the clinical diagnostic process.

The defined three scenarios and the designed corresponding frameworks are related to clinical settings and are expected to be used in clinical practice. Specifically, the designed two-way framework can be used in clinical practice when facing a person who worries about their memory and aims to detect whether they are living with a disease (FMD, ND and MCI) or not. If the aim is to detect whether a person needs to be referred to secondary care, the designed three-way framework can be used for classifying the collected recordings into HC, FMD and MCI+ND.

The IVA_{60} dataset that includes 15 recordings each from the HC, MCI, FMD and ND

groups was used for the study. To extract the features from the audio recordings in the IVA_{60} dataset, different types of speech segments were used as the input of an end-to-end feature extractor. After exploring, the twin-CCLA feature extractor was designed as the feature extractor by using the low- and high-quality speech segments separately. The mismatch between the low- and high-quality speech segment was demonstrated to exist by feature analysis. Then, the Euclidean distance is used for evaluating the mismatch and used as a one-dimension vector during the feature construction.

Two feature sets were extracted based on the twin-CCLA feature extractor. The difference between them is the input information. The trained features from the twin-CCLA feature extractor with the raw waveform as the only input were referred to as **TR-1** and used as the front-end features of the pipeline system. The performance of the TR-1 features was tested in the three scenarios. To include more information from the audio recordings, the conversation analysis and linguistic information features designed in Mirheidari *et al.* [2020] were used as the extra input of the twin-CCLA system for combining with the acoustic information embedded in the raw waveform. The trained features were referred to as **TR-2**. The results of all three scenarios showed that the TR-2 could perform better than the TR-1. By comparing the performance of the TR-1 and TR-2, the influence of the conversation analysis and linguistic features on the three scenarios was analysed.

As reviewed in Section 3.2, the mainstream dementia detection methods can be categorised into traditional feature-based systems and end-to-end technologies based systems. In this thesis, Chapter 6 explored how to use the end-to-end feature extractor for learning the acoustic features; Chapter 7 explored how to use the designed and traditional acoustic features to improve the performance of the dementia detection system. This chapter fuse the features learned from the end-to-end feature extractor with the designed acoustic features to improve the performance of the learned features, which has answered RQ1.

Chapter 9

Multi-task Estimation of Age and Cognitive Decline from Speech

Contents

9.1 Introduction 155

9.2 Background 155

9.3 Data Analysis 157

9.4 Multi-task System Construction 160

9.4.1 End-to-end System 160

9.4.2 Pipeline System 162

9.5 Experimental Setup 162

9.5.1 Datasets 163

9.5.2 Evaluation Setting 164

9.6 Results 165

9.6.1 Baseline Results 165

9.6.2 End-to-end System based Result 166

9.6.3 Pipeline System based Result 168

9.7 Summary 169

An individual's speech properties may change over time due to ageing, but changes may also occur due to an illness, such as those that cause cognitive decline. The changes due to ageing and cognitive decline result from two independent processes but are highly correlated. In our previous research in Chapter 6, Chapter 7, and Chapter 8, only the cognitive decline is considered for acoustic based feature analysis, without consideration for the person's age. In this chapter, both the age and cognitive decline are estimated simultaneously in a system based on acoustic features, which addresses the fourth research question: "age and cognitive decline are confounding factors; how can the age information be used to improve the performance of the dementia detection system? (**RQ4**)". The structure of this chapter is as follows:

Section 9.1 introduces the background of the speech-based cognitive decline and age estimation, and their relationship.

Section 9.2 presents the related research background of this chapter.

Section 9.3 analyses the relationship between the age and cognitive states on acoustic features commonly used in previous research.

Section 9.4 presents the constructed multi-task systems.

Section 9.5 contains the information about the experimental setup of our proposed system and the baseline acoustic features.

Section 9.6 summarises the classification results and analysis of our proposed system.

Section 9.7 contains the summary of this chapter.

9.1 Introduction

As described in Section 2.1.1, cognitive decline, associated with early signs of many neurodegenerative disorders, is caused by slow progressive loss of neurons in the central nervous system and can lead to an irreversible selective loss of brain functions [Haider *et al.*, 2019], resulting in speech changes even decades before diagnosis. The effects of cognitive decline on speech have been reviewed in Section 2.2.1. The relationship between acoustic features and Mini Mental Status Examination (MMSE) has been analysed in Fu *et al.* [2020]. With ageing, the subsystems which make up the human speech production system undergo progressive physiological change affected by the decreasing rate and strength of muscle contraction [Kelly & Harte, 2011], resulting in acoustic changes, including the changes in F_0, the rate and intensity of speech, the quality of the speech and the stability of the speech [Linville, 1996; Reubold *et al.*, 2010].

The changes caused by ageing and cognitive decline in speech result from two independent processes but are highly correlated. Fu *et al.* [2020] demonstrated that utilising information such as age and education can improve the estimation of the Mini Mental Status Examination (MMSE), a commonly used set of questions for screening cognitive function. More detailed information about the MMSE can be found in Section 2.3.2. In addition, previous research demonstrated that the acoustic features (e.g. speaking duration, F_0, x-vector) that have been shown effective for diagnosing pathological speech and estimating age are similar. For example, Meilán *et al.* [2014] proposed that some acoustic features linked to AD are also associated with normal ageing. In this chapter, both the age and MMSE is estimated by using the traditional acoustic features or the raw wave-based end-to-end system, respectively.

9.2 Background

Automatic age estimation can be either a regression (ageing is a continuous progress) or a classification task (consider each specific age or age range as a class). For classification, the earlier studies are based on the use of Perceptual Linear Prediction (PLP) and MFCC [Mahmoodi *et al.*, 2011] as the input of an SVM for the classification procedure. A

Gaussian Mixture Model (GMM) based method was proposed for learning the age-specific information followed by an SVM for classification or regression [Dobry *et al.*, 2011]. The efficiency of the F_0 and formats, as well as other prosodic features, has been demonstrated for the age estimation [Spiegl *et al.*, 2009]. State-of-the-art approaches involve popular speaker embedding like the i-vector [Dehak *et al.*, 2010] or x-vector [Snyder *et al.*, 2018] as the front-end features followed by a regression stage, like an SVM based regression [Sadjadi *et al.*, 2016] or a shallow Neural Network [Fedorova *et al.*, 2015; Ghahremani *et al.*, 2018; Kalluri *et al.*, 2019] based regression. The results reported in Kalluri *et al.* [2019] are 7.60 and 8.63 RMSE for male and female respectively, as well as the 4.92 Mean Average Error (MAE) on the IS10 feature set [Ghahremani *et al.*, 2018] (more detailed information in Section 3.2.1), which includes the F_0, Jitter, Shimmer.

For the automatic acoustic-based cognitive decline estimation methods, the performance of the typical pipeline systems for cognitive decline depends on both the front-end acoustic feature and back-end estimation stage. As reviewed in Section 3.2, in addition to the paralinguistic acoustic feature sets [Haider *et al.*, 2019], the x-vector and i-vector are also efficient for pathological speech detection [López *et al.*, 2019; Quintas *et al.*, 2020]. In recent years, inspired by the outstanding performance of deep neural networks used in numerous speech-based research areas reviewed in Section 3.2.2, deep neural networks are demonstrated to be efficient for dementia detection, as demonstrated in Chapter 5 and Chapter 6.

The information between related tasks can be shared using Multi-task learning (MTL) methods. The MTL methods can be categorised into *hard parameter sharing* and *soft parameter sharing* for hidden layers. Hard parameter sharing represents the neural networks that share the hidden layers between tasks while keeping several task-specific output layers. Each task has its own model for the soft parameter sharing method, and the distance between the model parameters is restricted while training.

The MTL has led to successes in many applications of machine learning, from natural language processing and speech recognition to computer vision and drug discovery [Ruder, 2017]. Specifically, in Collobert & Weston [2008], a single convolutional neural network is constructed for predicting the part-of-speech tags, named entity recognition, seman-

tic roles labelling, chunking and language modelling. In Deng *et al.* [2013], two kinds of multi-task speech recognition systems are introduced. Firstly, mixed-band acoustic recordings with 16-kHz and 8-kHz sampling rates are both recognised with one single system. Secondly, a DNN system is constructed for multilingual or cross-lingual speech recognition. Considering the high correlation between the age and MMSE, this chapter addresses a novel problem of the simultaneous estimation of the two changes in speech properties utilising multi-task learning.

9.3 Data Analysis

In order to analyse the relationship between the age and MMSE in acoustic features, multiple datasets are used, including the publicly available datasets (the Trinity College Dublin Speaker Ageing (TCDSA) dataset [Kelly *et al.*, 2014], the DementiaBank dataset [Becker *et al.*, 1994]) and the IVA dataset. The information of the DementiaBank dataset and the IVA dataset can be found in Section 4.1.1 and Section 4.2, respectively. In our experiment, the IVA subset that includes the age or MMSE information is selected for the analysis in this section. The dataset information, including the number of recordings, number of speakers and the age range, is summarised in Table 9.1. In total, there are 406 recordings from HCs and 370 recordings from people with cognitive decline (CD)[1]. The age range is from 19 to 96. Only the recordings from the DementiaBank and the IVA datasets include the MMSE value. More detailed information about the TCDSA dataset can be found in Section 9.5.1. Although this collective dataset contains different accents, recording environments and speech content, it can still provide some intuition for the correlations observed between the effect of age and MMSE on acoustic features.

First, the influence of age on the typical acoustic features is analysed for people living with or without cognitive decline. The features are extracted automatically using the open-source *myspsolution.praat* toolkit [Jadoul *et al.*, 2018]. The extracted acoustic features from the HC and CD groups are averaged over the age separately. As shown in Figure 9.1, the commonly used traditional features, including *F0_median*, *speaking*

[1]In this chapter, "cognitive decline" is defined as belonging to the ND group (IVA) or the AD group (DementiaBank)

Table 9.1: The information for the datasets used for data analysis; CD: cognitive decline.

Dataset	**Number of Recordings (HC vs. CD)**	**Number of Speakers (HC vs. CD)**	**Age Range**
TCDSA	179 vs.71	23 vs. 3	[19, 96]
DementiaBank	222 vs. 255	89 vs. 168	[46, 90]
IVA	5 vs. 44	5 vs. 40	[26, 87]
In total	406 vs. 370	117 vs.211	[19, 96]

duration, *number of pauses* and *number of syllables* from people living with or without cognitive decline seems to have a weak inverse correlation trend when plotted against age. For example, the number of syllables from people living with cognitive decline and healthy controls is reversed. With ageing, the number of syllables decreased in the speech from people living with cognitive decline. In previous research [Guinn & Habash, 2012; Luz *et al.*, 2018], the result shows that there are fewer syllables in AD than in non-AD, but the relationship between age and cognitive decline on syllables was not analysed.

The relationship between the age and MMSE is also explored by calculating the Euclidean distance between the x-vectors of HCs and people diagnosed with different MMSE values. For calculating the *anchor* (i.e., average) x-vector representing the healthy controls, only x-vectors from the people in the TCDSA dataset not known to have any cognitive health issues are used for calculating the average representation, which is 179 recordings in total. The averaged x-vector is used as the anchor x-vector. It is hypothesised that the distance between the anchor x-vector and x-vectors averaged across speakers with a particular MMSE value is larger for lower MMSE values (indicating more severe cognitive decline cases).

As shown in Figure 9.2(a), the plotted distance for each MMSE is estimated by averaging the distance between the anchor x-vector and the x-vectors of speakers with the corresponding MMSE. However, the values in Figure 9.2(a) has not considered age. Under this situation, only one anchor x-vector is calculated. To analyse the correlation between the age and MMSE values, multiple age-specific anchor x-vectors are calculated by averaging the x-vectors from people of the same age. The anchor x-vector corresponding to any missing age values is estimated by averaging the x-vectors of its neighbour ages. The average distance of the x-vectors for each MMSE value and healthy anchor

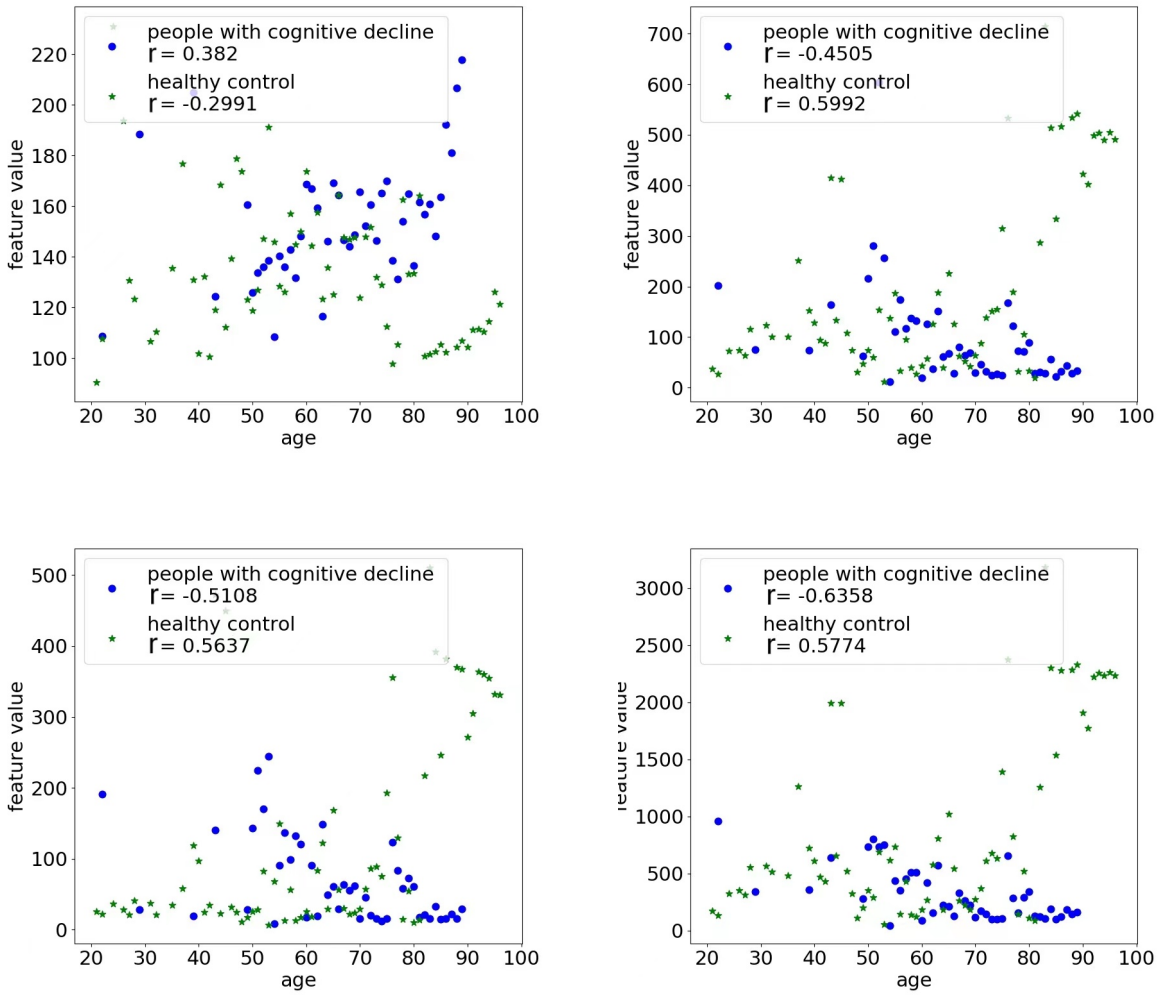
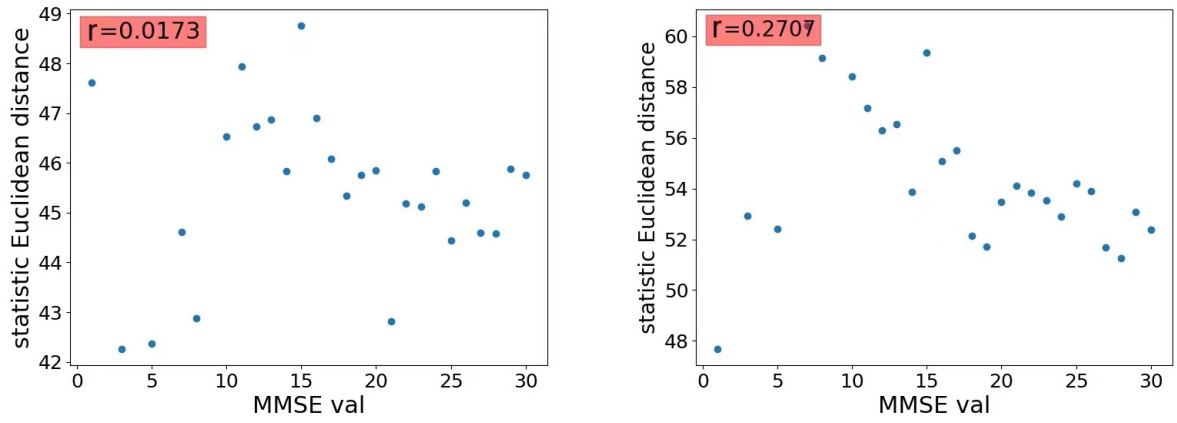


Figure 9.1: The correlation between age and acoustic features (top left: F0_median, top right: speaking duration, bottom left: number of pauses, bottom right: number of syllables) with different cognitive status.

age-specific x-vector is shown in Figure 9.2(b). By comparison, it is found that the relationship (Pearson's correlation) between healthy and people living with cognitive decline becomes stronger after considering age. Specifically, the p-value increased from 0.0173 to 0.2707 after considering age. The comparison between Figure 9.2(a) and Figure 9.2(b) demonstrates the correlation exists between the age and MMSE when considering x-vector acoustic features.

The relationship between the change caused by the age and MMSE on speech (show in Figure 9.1 and Figure 9.2) encourages us to explore how their estimation might be



(a) age un-specific

(b) age specific

Figure 9.2: The Euclidean distance between the anchor x-vectors and x-vectors extracted from people with different MMSE values.

improved by training a joint, single system, like a multi-task system. The mainstream architectures for the age and MMSE estimation are the pipeline systems and end-to-end systems. For the pipeline system, speaker embeddings, like the x-vector or i-vectors, have achieved excellent performances for both the age or MMSE estimation. On the other hand, the efficiency and interpretability of the Sinc-CLA end-to-end system for neurodegenerative related disorder classification have been demonstrated in Chapter 6. Therefore, in this chapter, both of the two mainstream structures are adopted for multi-task learning.

9.4 Multi-task System Construction

In this section, both the Sinc-CLA end-to-end system proposed in Chapter 6 and a traditional pipeline system is designed as multi-task classification systems for the age and MMSE estimation.

9.4.1 End-to-end System

Previous studies in Chapter 6 have shown that the Sinc-CLA architecture has a good performance and interpretability in classifying the recordings from people living with

mild cognitive impairment, neurodegenerative disorders, or healthy controls. The multi-task Sinc-CLA system introduced in this chapter is shown in Figure 9.3. The two tasks share the SincNet Layer and CNN layers, but the Bidirectional LSTM (BLSTM) layer and its following layers are separately trained with a specific target (the age or MMSE). The detailed description of each functional layer can be found in Section 9.5.2 of this chapter and Chapter 6.

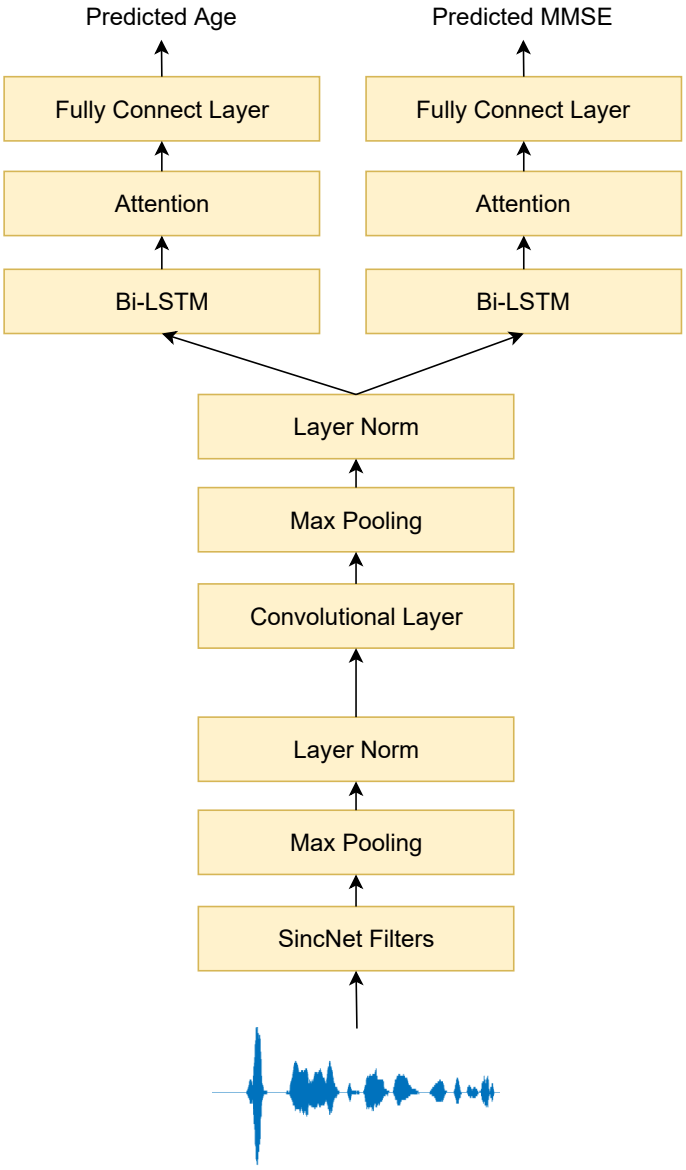


Figure 9.3: The structure of the multi-task Sinc-CLA system for the age and MMSE estimation.

9.4.2 Pipeline System

For constructing the pipeline system, the x-vector or i-vector speaker embeddings were adopted as the front-end features for the age and MMSE estimation. To make use of the age information in the estimation of the MMSE and likewise using the cognitive decline to improve the age estimation, a multi-task shallow neural network comprised of two shared fully connected layers, and one separated output layer is designed for the front-end feature regression, as shown in Figure 9.4.

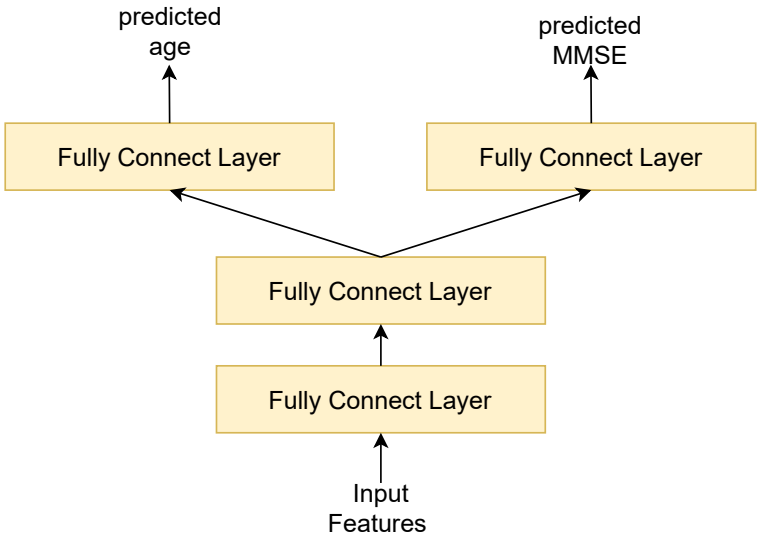


Figure 9.4: The structure of the multi-task pipeline system for the age and MMSE estimation.

9.5 Experimental Setup

The target of our experiment is to estimate the age and MMSE at the same time with the two designed systems introduced in 9.4. For evaluating our experiments, both the IVA dataset and the publicly available dataset ADReSS are used for testing. More detailed information is introduced in Section 9.5.1. The evaluation setting is presented in Section 9.5.2.

9.5.1 Datasets

In Section 9.3, the Trinity College Dublin Speaker Ageing (TCDSA) is used for data analysis. The TCDSA dataset was designed primarily to investigate the effect of the ageing-related vocal change on speaker verification [Kelly *et al.*, 2014]. The main portion of this dataset contains speech recordings of 26 adults (15 males and 11 females) across a period of between $25-58$ years per speaker. Among the 26 speakers, three (Thatcher, Reagan and Neill) were diagnosed with mental health problems in their later years and the others are not known to have any cognitive health issues. Therefore, only the recordings from the people without any cognitive health issues were adopted for the analysis in Section 9.3.

The target of our experiment is to estimate the age and MMSE for the IVA dataset with a system trained on the publicly available DementiaBank dataset. For the IVA dataset, in our experiment, only the audio recordings from the participants that has associated labels with MMSE and age information are used, which is a total of 34 recordings with ages ranging from 45 to 80. The subset of the IVA dataset is referred to as the $IVA_{age\&MMSE}$ dataset. Further information about the data can be found in Section 4.2.4. Likewise, the subset of recordings with both the age and MMSE labelled are selected from the DementiaBank dataset. After selection, as presented in Section 4.1.1, 459 recordings from 286 speakers with ages ranging from 46 to 95 are left. The 459 recordings are separated into 5 folds[1] for the CV application. In the 5-fold CV, 4 folds are used for training, 1 fold for validation (hyperparameter optimization), and the recordings from the $\text{IVA}_{age\&MMSE}$ dataset are used for testing (test set). The results presented in this chapter are averaged across the 5 results obtained from the test set on each of the systems corresponding to the 5 folds.

To compare, the ADReSS dataset is also applied in this chapter. To train the system, the 108 speakers in the training set are divided into 9 folds as in Cummins *et al.* [2020]. The result presented is averaged across the result estimated by the 9 trained systems. The dataset used for the experiment in Section 9.6 are shown in Table 9.2.

[1]Partitioning of folds available on request

Table 9.2: Detailed information for datasets used for the multi-task learning system training.

Dataset	Number of Recordings	Number of Speakers	Age Range
ADReSS	156	156	[50, 79]
DementiaBank	459	286	[46, 95]
$\text{IVA}_{age\&MMSE}$	34	34	[45, 80]

9.5.2 Evaluation Setting

The Kaldi Toolkit is adopted for the x-vector and i-vector based speaker embedding extraction. The detailed information about the system settings can be found in Snyder *et al.* [2018]. To train an x-vector extractor and an i-vector extractor, the combination of speaker recognition evaluation (SRE) datasets (SRE 2004, SRE 2006 train set and SRE 2008) and Linguistic Data Consortium (LDC) datasets (LDC2001S13, LDC2004S07, LDC98S75, LDC99S79 and LDC2002S06) are used [Snyder *et al.*, 2018]. In total, 141k acoustic recordings are used. The trained DNN extractors and total variability space can map each recording in the DementiaBank and $\text{IVA}_{age\&MMSE}$ datasets into a 512 dimension x-vector and a 600 dimension i-vector, respectively.

For the Sinc-CLA system, the parameter setting of each layer is the same as in Chapter 6, except for the loss function, which is the Mean Average Error (MAE) in the current regression system. While training, the mini-batch size is set to 80 and the epoch is set to 100. The parameters are selected according to the performance of the DementiaBank evaluation set. For regression, both the age and MMSE value is normalised to the [0,1] range before estimation. To train the system, each recording is cut into multiple 2-second chunks and assigned a label corresponding to its normalised age or the MMSE value. The predicted value for the test recording is the average of the estimated value of all the corresponding chunks from that recording.

The single-task shallow neural network is composed of two fully connected layers with 64 units and a 1-unit output layer. The output layer of the multi-task shallow neural network is two separate 1-unit fully connected layers for each regression task. All hidden layers use the leaky-ReLU non-linearities. To train the system, the *rmsprop* is applied as

the optimiser with a learning rate of 0.01. While training, the batch size is set to 80 and the epoch is set to 300. For the two multi-task systems, the weight of age-based MAE and MMSE based MAE share the same weights when added together as the loss criteria for parameter tuning. The parameters are also tuned according to the performance of the DementiaBank evaluation set.

9.6 Results

In this section, the results achieved from the baseline systems are summarised first, followed by the end-to-end system and the pipeline system.

9.6.1 Baseline Results

As the baseline system, the x-vector, i-vector and the ComParE features (6373-dimension including energy, spectral, MFCC, and voicing related low-level descriptors (LLDs)) extracted by the OpenSMILE toolkit are regressed with the SVM for the age or MMSE estimation. The results are shown in Table 9.3. In our experiment, the Root Mean Square Error (RMSE) is utilised as the criteria for comparing the performance of the baseline and proposed approaches. As shown in Table 9.2, the age range of the IVA$_{age\&MMSE}$ dataset is between 45 and 80.

Table 9.3: The results from the SVM based regression.

Target	Feature Type	RMSE
Age	ComParE	5.23
	I-vector	4.99
	X-vector	**4.83**
MMSE	ComParE	5.34
	I-vector	**5.03**
	X-vector	5.36

Though some previous research showed that the i-vector can provide better or similar performance in various pathological related research tasks [Quintas *et al.*, 2020] for the MMSE estimation, our experimental results show that the i-vector achieves a better result for the MMSE estimation (5.03 RMSE) compared with the x-vector (5.36 RMSE) and

ComParE (5.34 RMSE) in Table 9.3. However, for the age estimation, x-vector achieves the best performance compared with the ComParE and i-vector, which is a 4.83 RMSE. In Section 9.6.3, both the i-vector and x-vector are used for testing their performance on the single-task and multi-task pipeline systems.

9.6.2 End-to-end System based Result

Table 9.4: Results from the single-task and multi-task Sinc-CLA network.

Target	Task Type	RMSE
Age	Single	5.17±(0.32)
	Multi	5.40±(0.21)
MMSE	Single	4.44±(0.21)
	Multi	4.43±(0.14)

The results of the end-to-end system trained for single-task and multi-task targets are shown in Table 9.4. For measuring the performance and stability of our proposed system, both the average RMSE and standard deviation RMSE over the five folds are shown in the table. From Table 9.4, the following conclusions are drawn:

- By comparing the results for the same task from single and multi-task systems within Table 9.4, it is found that the multi-task learning can improve the RMSE of the MMSE estimation from 4.44 to 4.43, but causes a small decline in the age estimation.

- By comparing the results in Table 9.4 and Table 9.3, it is found that the performance of the end-to-end system can ensure a better MMSE estimation but performs worse on the age estimation compared with the baseline system (5.17 vs. 4.83).

- The best performance of the multi-task end-to-end system over the 5 folds is the 4.85 (5.17±0.32) RMSE on the age estimation, which is almost the same as the performance of the baseline result (the 4.83 RMSE).

To illustrate the critical information for the age and MMSE estimation learned by the SincNet filters, the cumulative frequency responses (CFRs) of the SincNet from the three

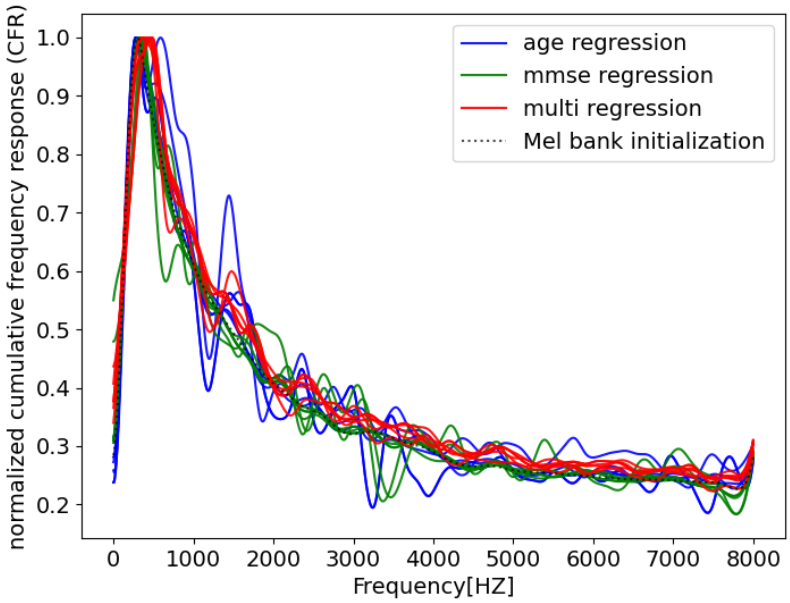


Figure 9.5: The learned normalised Cumulative Frequency Response from the three tasks.

systems are plotted in Figure 9.5. Similar to Figure 6.3, lines in the same color correspond to the same regression task training for the 5-fold CV.

The black line is the initialised Mel-bank CFRs. The fluctuations shown in the trained CFRs for different regression tasks represent the specific information learned in the system training. Comparing the initialised Mel-bank CFRs with learned CFRs can demonstrate that the Mel-bank is not perfect, though the MFCC is popular as the acoustic feature for dementia detection-related research. By comparing the CFRs from different regression tasks, it is found that the different fold-specific CFRs appear to be less variable for the multi-task regression than for the single-task system. It is consistent with the results shown in Table 9.4 that the results from the multi-task system are more stable than the results from the single-task systems.

By comparing the CFRs shown in Figure 9.5 with the CFRs shown in Figure 6.3, it is found that the CFR lines corresponding to the age and MMSE estimation tasks are clearly overlapping, which demonstrates that the critical information learned by the age estimation and MMSE estimation tasks is similar. The result encourages us to do more exploration on how to use the confounding factors between age and cognitive decline to

improve the performance of the dementia detection system in future studies. Considering the results shown in Table 9.4, MMSE estimation is improved at the cost of a small decrease in age estimation, how to balance the multi-task learning to improve every single task at the same time should be considered in the future.

9.6.3 Pipeline System based Result

The results from the single-task and multi-task neural network-based systems (The system introduced in Section 9.4.2) are shown in Table 9.5. By comparing the results from the x-vector, it is found that both the age and MMSE estimation can be improved from 4.89 to 4.64 (age) and from 4.50 to 4.35 (MMSE) when utilising the multi-task learning. By comparing with the results in Table 9.4 and Table 9.3, it is found that the x-vector based multi-task learning can achieve the best performance. In addition, the results from the i-vector are consistent with our expectation that the multi-task learning performs better than the single-task learning, though not as well as the result from the SVM based regression in Table 9.3. Comparing the performance of the x-vector and i-vector under the same evaluation setting demonstrates the efficiency of the x-vector with the shallow neural network.

Table 9.5: The results with speaker embedding features on single-task/multi-task pipeline system estimation.

Target	**Task Type**	**RMSE (x-vector)**	**RMSE (i-vector)**
Age	Single	4.89±(0.08)	5.88±(0.39)
	Multi	**4.64**±(0.11)	5.47 ± (0.29)
MMSE	Single	4.50±(0.09)	8.32±(0.42)
	Multi	**4.35**±(0.22)	7.25±(0.37)

9.6.3.1 Task Weight Comparison

For exploring the influence of the weight of the two tasks on the regression performance, the weights of the two tasks are changed while doing the multi-task estimation with the pipeline system. The results are shown in Table 9.6. As shown in the table, the best

result is the from the balanced weights (age:MMSE = 1:1) as in Section 9.6.3. It shows that adding the weights of the specific task while training the system cannot ensure a better performance.

Table 9.6: The regression results (RMSE) with speaker embedding features on single-task/multi-task pipeline system estimation.

Target / Weight (Age: MMSE)	**1:1**	**1:2**	**2:1**
Age	**4.64**±(0.11)	4.80±(0.13)	5.04±(0.58)
MMSE	**4.35**±(0.10)	4.67±(0.79)	4.48±(0.27)

9.6.3.2 Results on the ADReSS Dataset

To check the performance of our proposed method on the publicly available ADReSS dataset, the following experiment is designed. The proposed x-vector based shallow neural network is used on the ADReSS dataset MMSE estimation task. Similar to the previous results, the multi-task learning can improve the estimation of the age and MMSE: the RMSE values obtained on the MMSE estimation is 5.85 with the multi-task learning, compared with the baseline 6.14 shared in Luz *et al.* [2020a] and 5.92 in Syed *et al.* [2020] (acoustic features only).

9.7 Summary

The changes caused by ageing and cognitive status are two independent processes but can result in similar physiological changes, like the decreasing rate and strength of muscle contraction. These changes can both lead to acoustic changes, including the changes in F_0, the rate and intensity of speech, quality of the speech and stability. When designing the acoustic-based dementia detection system, including the age information in a dementia detection system construction by designing a multi-task estimation system is expected to be beneficial for the performance of the designed system.

This chapter presented a multi-task method by utilising either the end-to-end system or a shallow neural network-based pipeline system for estimating the MMSE and age

for people living with or without cognitive decline. Firstly, the analysis of the acoustic features from people living with or without cognitive decline revealed the relationship between the two confounding factors. Then, the result from the $\text{IVA}_{age\&MMSE}$ and ADReSS datasets demonstrated that applying the multi-task learning techniques using x-vectors can achieve better results than the single-task architecture or the SVR based pipeline systems on the MMSE estimation. The end-to-end Sinc-CLA system can achieve better results on the MMSE estimation with the multi-task system compared with the single-task system but cannot get a better result on the age estimation.

In the previous research, age information has also been used as the input of the dementia detection system [Fu *et al.*, 2020]. However, age and cognitive status are two independent processes but are confounding factors that can affect speech abilities. In other words, age and cognitive status share similar statuses. As a result, a multi-task estimation system is proposed in this chapter regarding age and MMSE as the two outputs of the designed system. In future work, other methods for including the age information into the dementia detection system construction is expected to be explored.

Chapter 10

Conclusions and Further Work

Contents

10.1 Conclusions 172

10.2 Limitations 177

10.2.1 Limited Publicly Available Data 177

10.2.2 Inconsistent Settings and Performance 178

10.3 Future Work 179

10.3.1 Simplify the Automatic System Structure 179

10.3.2 Exploring Longitudinal Applications 180

10.3.3 Utilising Transfer Learning Technologies 180

10.4 Concluding Remarks 181

10.1 Conclusions

The thesis aims at constructing automatic systems for speech and language-based dementia detection utilising deep learning technologies. State-of-the-art technologies were adopted for constructing the systems. Both the acoustic and linguistic features were extracted automatically from the audio recordings for dementia detection. The publicly available datasets (DementiaBank and ADReSS) and the dataset (IVA) collected by the Royal Hallamshire Hospital (Sheffield, UK) were both used for the studies. The results confirm the potential of using the automatic system as clinical assistance for dementia detection. The four research questions put forward in 1.2 will be restated together with their solutions proposed in this thesis.

RQ1: how can state-of-the-art deep neural networks be applied for speech- and language-based dementia detection?

The cutting edge deep learning technologies were successfully used for extracting the linguistic and acoustic information embedded in the audio recordings for dementia detection, like the research proposed in Chapter 5, Chapter 6 and Chapter 8. Though most of the deep neural network-based end-to-end systems are black boxes and hard to interpret, the end-to-end systems designed in our studies allowed us to do some analysis and interpretation of what had been learnt in the proposed systems.

The audio recordings in the DementiaBank dataset collected from people describing the cookie theft picture were used as the material for the study presented in Chapter 5. To extract the linguistic information from the audio recordings, a hierarchical system named HBANN was proposed for extracting the hierarchical information from the manual and the automatic transcripts. First of all, the transcripts are embedded into word vectors using the word embedding layer. In the HBANN, the bidirectional LSTM was used at both the word and sentence levels to extract the sequential information embedded in the transcripts, followed by the attention mechanism for weighting the importance of the learned feature vectors. As a result, the information embedded in each sentence was represented by a vector by processing the word vectors at the word level, and the information embedded in the transcript was represented by a vector by processing the

sentence vectors at the sentence level. The efficiency of the hierarchical structure and the attention mechanism were demonstrated by comparing with the baseline systems. In addition, the attention weights were plotted for interpreting how the attention mechanism works. The setting of the word embedding layer was also analysed in the study, which can help us to understand the difference between the HC and AD on the language ability better.

In Chapter 6, an end-to-end system named as *Sinc-CLA* was designed with the raw waveform as the input for extracting the data-driven acoustic features (trained features) for different classification tasks. The aim of the study is the binary classification between the recordings from people living with or without cognitive decline (ND, and MCI) on the IVA_{3class} dataset. The proposed Sinc-CLA system was composed of four functional network layers: SincNet layer, convolutional layer, bidirectional LSTM layer and attention layer. The SincNet was proposed based on the CNN, but it was designed to be used as the first layer for processing raw waveform [Ravanelli & Bengio, 2018b]. A unified structure composed by SincNet, CNNs, LSTMs, and DNNs was designed for speech processing and proven to be efficient for exploiting the complementary natures of the multiple functional layers inspired by Sainath *et al.* [2015a]. The popular traditional feature sets (IS10 and the ComParE) were used as the baseline features. The trained features achieved superior performance on the IVA_{3class} dataset compared with the baseline feature sets on the three classification tasks. To analyse this, the Cumulative Frequency Response (CFR)s and the outputs of the SincNet layer were plotted and analysed in our study, making the trained features more interpretable. The conclusion is consistent with the research proposed in Alhanai *et al.* [2017]; Meilán *et al.* [2014].

In Chapter 8, for modelling the speech with low and high confidence scores estimated by the ASR system respectively, a twin based end-to-end system was designed inspired by the previous speech-based research [Fritsch *et al.*, 2018; Ma *et al.*, 2016a]. A feature extractor was designed by using CNNs, LSTMs, and DNNs and the attention mechanisms, similarly as the system proposed in Chapter 6. The Euclidean distance evaluated the mismatch between the speech with low and high confidence scores. The effect of modelling the segments with low and high confidence scores separately or indiscriminately was eval-

uated on the IVA_{60} dataset. The result showed that modelling the categorised segments with the twin system can improve the performance of the learned features for dementia detection. More than using the acoustic information embedded in the raw waveform, the linguistic features and conversational analysis features were used together with the raw waveform as the input for improving the performance of the learned features. Compared with the features proposed in Mirheidari *et al.* [2020], the features extracted from the proposed feature extractor can provide a superior result on the IVA_{60} dataset.

RQ2:how can the known clinical dementia detection knowledge help in constructing an automatic dementia detection systems and extracting useful features?

When designing the systems, the medical knowledge used by the clinicians was utilised in designing the deep learning systems and extracting the features. For example, the hierarchical system proposed in Chapter 5 was designed by considering the hierarchical structure of transcripts (word level and sentence level), similarly to that done for a clinical diagnosis. In addition, the symptoms of unclear pronunciation and frequent/long pauses that is present in speech from AD was utilised for designing and extracting the features in Chapter 7.

In Chapter 5, a hierarchical attention-based system was proposed for extracting the linguistic information from the transcripts inspired by the clinical knowledge. Specifically, the linguistic ability decline that existed at both the word and sentence levels was used for the clinical diagnosis. Under this inspiration, a hierarchical system was constructed for learning the hierarchical information embedded in the transcripts. In addition, while diagnosing, clinicians take the words and the sentences into account separately. For example, clinicians tend to pay more attention to the informative units than the empty words in the transcripts. For weighting the words and sentences while modelling, the attention mechanism was applied to the word vectors and sentence vectors according to the classification target in the system. The correlation between the words and sentences can reflect the semantic abilities of people, so the RNNs were used for extracting the sequence information from the time series transcripts. The attention weights and word frequencies were analysed in order to understand the designed system. This was the first

study that introduced the hierarchical structure into the end-to-end system for dementia detection.

The audio recordings collected from people living with dementia are likely to include long pause rates and unclear pronounced words. The recordings also vary in acoustic quality and the degree of background noise when collected in a home-based environment, resulting in difficulties for high-performing acoustic feature extraction. An ASR system is generally indispensable for extracting the linguistic information from the audio recordings for dementia detection. In addition to the transcribed words, the time alignment and confidence score estimated for each word is also estimated by the ASR system. To extract high-performing acoustic features, in Chapter 7, a three-dimensional rhythm-related feature was designed using the ASR decoding outputs, including confidence scores and time alignment information. The designed rhythm-related features were combined with the acoustic features extracted from the high-quality audio segments for dementia detection on the DementiaBank dataset. The performance of the combination of the rhythm-related and acoustic features was examined by changing the confidence score threshold. The result showed that increasing the confidence score threshold within some range can improve the system's performance on dementia detection. After combining with the linguistic-based system proposed in Chapter 5, the performance of the automatic dementia detection was improved further.

RQ3: how to design a framework for more clinically relevant diagnostic scenarios?

In the clinical diagnosis, diagnosing MCI, FMD and ND from HC is of great importance as the treatment of FMD, MCI and ND is different. However, doing the four-way classification scenario is difficult due to similar and overlapping symptoms for people living with MCI, ND and FMD. In clinical diagnosis, the four-way classification scenario is always replaced by the three-way scenario (regarding FMD and HC as one class) or the two-way scenario (further regarding MCI and ND as one class) to guide the follow-on treatment. In Chapter 8, the IVA_{60} dataset that includes 15 recordings from HC, FMD, MCI and ND respectively was used for the study. Considering the clinical practice, the three scenarios mentioned above were designed for classifying the recordings in the IVA_{60}

dataset. For extracting the acoustic features, inspired by the research in Chapter 6 and Chapter 7, the confidence scores and time alignment information estimated by the ASR system was used for categorising the audio recordings into high-quality and low-quality audio segments. Then, a twin-CCLA system was designed to extract acoustic features by using the mismatch between high-quality and low-quality audio segments. The Euclidean distance evaluated the mismatch between the high-quality and low-quality audio segments. For combining with the linguistic information and conversational analysis information, the features were trained with the designed twin-CCLA feature extractor by inputting the raw waveform and the conversation analysis and word vector-based features proposed by Mirheidari *et al.* [2020].

RQ5:age and cognitive decline are confounding factors; how can the age information be used to improve the performance of the dementia detection system?

Ageing and cognitive decline result from two independent processes but are highly correlated, and both are shown on the acoustic characters. When doing cognitive decline estimation on the longitudinal collected dataset, like the IVA dataset, the age information can be used for benefiting the performance of the system. In Chapter 9, a multi-task method was proposed for estimating the Mini Mental Status Examination (MMSE) (a commonly used set of questions for screening cognitive function) and age for people living with or without cognitive decline. The Sinc-CLA end-to-end system and a shallow neural network-based pipeline system were designed to realise the multi-task method. The proposed systems were tested on the $IVA_{age\&MMSE}$ datasets. For the multi-task based pipeline system, i-vector [Dehak *et al.*, 2010] and x-vector [Snyder *et al.*, 2018] proposed for speaker identification were used as the front-end features.

Though the relationship between the age and MMSE was stated in the previous research [Fu *et al.*, 2020; Meilán *et al.*, 2014], this was not based on any quantitative analysis. In our study, the relationship between the age and MMSE on the acoustic features were analysed quantitatively. The analysis demonstrated that including age information can improve the correlation between the acoustic features and the MMSE value. The performance of the two multi-task estimation systems was also listed in Chapter 9. The

result demonstrated that applying multi-task learning techniques with x-vectors as the front-end features followed by a shallow neural network classifier can improve the age and MMSE estimation compared with estimating those two separately. Furthermore, the same system was used on the ADReSS dataset, which also achieved a superior result on the MMSE estimation task.

10.2 Limitations

The limitations of the studies in this thesis can be summarised into two parts: a limited amount of publicly available data and the inconsistent settings and performance of the proposed systems on different datasets.

10.2.1 Limited Publicly Available Data

For speech and language-based dementia detection, the DementiaBank dataset is the largest publicly available dataset, which includes 222 samples from 89 HC and 255 samples from 168 people living with AD. In terms of the size of datasets normally used for speech technology work, this is relatively small. In addition, the quality of the audio recordings in the DementiaBank dataset makes the work challenging. Therefore, the application of complex deep neural networks is restricted. In addition, researchers have tended to use different training and test partitions of this dataset and apply different rules with respect to whether speakers with repeat sessions are kept purely in one of these partitions (ensuring all test data is from unseen speakers). This is particularly important when working with multiple folds.

In contrast, the IVA dataset is a relatively large dataset and has been used for some studies in this thesis. For this dataset though, there are restrictions in terms of data sharing which means there are difficulties in comparing with methods proposed in other research.

Since 2020, an Interspeech Challenge: Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) has been organised for providing researchers with a benchmark dataset for linguistic- and acoustic-based dementia detection tasks. More data has

been provided in the ADReSSo Interspeech-2021 Challenge, which encourages more researchers to work on the speech-based dementia detection research field. Hopefully, more datasets can be published with a standard evaluation framework in the near future.

10.2.2 Inconsistent Settings and Performance

The studies in this thesis includes five experimental chapters. In Chapter 6, the first end-to-end system is proposed for extracting the acoustic information from raw waveform for dementia detection. As mentioned in the Summary of Chapter 6, the trained features extracted from the Sinc-CLA system are demonstrated to be more efficient than the popular feature sets (the IS10 and ComParE) on the IVA_{3class} dataset. However, the conclusion derived from the DementiaBank dataset is the opposite. It is inferred that the proposed system cannot ensure a considerable performance on the audios with poor qualities. To help address this issue, the study in Chapter 7 is proposed for extracting high-performing features from the audio recordings in the DementiaBank dataset.

In Chapter 6, SincNet is used as the first layer of the proposed system for extracting task-driven features from the raw waveform, while a CNN is used as the first layer in Chapter 8. As shown in Chapter 8, many experiments have been done on the three scenarios. Each scenario is trained by five sessions using the 10-fold CV method. The trained feature extraction requires both time and computing resources. Therefore, the selection of the first layer is based on the performance of the four-way classification task, which cannot ensure the CNN is always the best choice for all three scenarios. The studies aim to extract high-performing acoustic features for dementia detection, so limited research has been done on improving the SincNet based end-to-end system.

The outputs of the ASR system are used as the auxiliary information both in Chapter 7 and Chapter 8. However, the rhythm-related feature designed in Chapter 7 is not used in Chapter 8 in the thesis. The main reason is the different types of questions in the DementiaBank and IVA dataset. For the IVA datasets, the audio recordings include eight conversational questions, two verbal fluency tests, and the picture description task. However, in the DementiaBank dataset, each audio recording only includes the cookie theft picture description task, which provides long and continuous speech for extracting

the rhythm-related features. In comparison, the recordings collected from IVA include the speech from the participants, their families, and the clinicians. The experiments were based on the audio recordings collected from the participants only, generated by manual segmentation and concatenation, which may influence the quality of the rhythm-related feature. Also, the time limitation restricted any further exploration on how to design similar rhythm-related features for the IVA dataset.

10.3 Future Work

In this section, three future work is summarised according to the studies in this thesis and the recent development of the research in the related fields. The discussions in this section is expected to provide the readers with some inspiration.

10.3.1 Simplify the Automatic System Structure

In the work in this thesis, to extract the linguistic information from the audio recordings, an ASR system is needed for transcribing the audio recordings into automatic transcripts if an automatic system is expected to be constructed. However, the linguistic information is also embedded in the collected audio recordings. For selecting the best transcript from the ASR system, the Word Error Rate (WER) is used as the criteria. However, the transcripts with the lowest WER cannot ensure the best performance for dementia detection. In Pan *et al.* [2021a], multiple ASR hypotheses were extracted from different paths in the lattice together with their corresponding confidence scores in the ASR system as the input of the linguistic feature modelling system. This was proven to be efficient for dementia detection. However, due to time limitations, further research has not been investigated. In Zhu *et al.* [2021], the audio embedding extracted from the wav2vec rather than the transcript is used as the input of the BERT. In the future study, a more straightforward and efficient way to extract the linguistic information from the audio recordings should be investigated, rather than using the ASR transcripts selected on the based of the WER for providing the linguistic information.

10.3.2 Exploring Longitudinal Applications

As mentioned in 10.2, in our studies, the limitation of the available longitudinal data restricts the study on the disease progress tracking. However, longitudinal disease progress tracking is of great research importance and practical value. In Yancheva *et al.* [2015], the cognitive decline estimation takes the identity information into the system design, which has improved the performance of the designed system. In Chapter 9, the relationship between age and cognitive decline was explored. In the Interspeech-2021 ADReSSo Challenge, disease tracking based on the longitudinally collected data is one of the three tasks in the challenge, and two teams have done some exploration on this task [Syed *et al.*, 2021; Zhu *et al.*, 2021]. Longitudinal tracking is practical for constructing a clinical assistant dementia detection tool and deserves to be explored further.

10.3.3 Utilising Transfer Learning Technologies

Training a system from scratch requires both a large amount of data and plenty of time. As an alternative, transfer learning can build accurate models for different tasks with the same pre-trained model. Over the last several years, the fields of NLP [Howard & Ruder, 2018; Peters *et al.*, 2018; Radford *et al.*, 2018], speech processing [Shivakumar & Georgiou, 2020; Tomashenko *et al.*, 2019; Wang & Zheng, 2015], and image processing [Hussain *et al.*, 2018; Kornblith *et al.*, 2019; Zhang *et al.*, 2016] have witnessed significant improvements using transfer learning methods. In this thesis, all the ASR systems designed for transcribing the audio recordings collected for dementia detection are transferred from the pre-trained ASR system. In the ADReSSo Challenge, the papers with the best result on the classification task [Pan *et al.*, 2021a; Pappagari *et al.*, 2021; Syed *et al.*, 2021] are all based on the pre-trained BERT and transferred learned for linguistic feature extraction. Similarly, in the ADReSSo Challenge, pre-trained wav2vec 2.0 [Baevski *et al.*, 2020] has been used by four papers [Balagopalan & Novikova, 2021; Gauder *et al.*, 2021; Pan *et al.*, 2021a; Zhu *et al.*, 2021] for extracting the acoustic embedding. Using the transfer learning based on the pre-trained model for making up the limitation of the dementia detection dataset should be explored further.

10.4 Concluding Remarks

This thesis works on constructing acoustic- and linguistic-based automatic dementia detection systems using deep learning technologies and clinical knowledge in the hope that these techniques may one day be used in clinical practice. In the studies, both the acoustic and linguistic information are extracted from the audio recordings, and automatic transcripts are generated by the ASR system. The various proposed systems are evaluated with both the publicly available datasets (DementiaBank and ADReSS) and the IVA dataset collected by Royal Hallamshire Hospital (Sheffield, UK). The studies in Chapter 5, Chapter 6 and Chapter 8 successfully used the state-of-the-art deep learning technologies for constructing the systems. In addition, the studies in Chapter 5 and Chapter 7 aim to include some clinical knowledge for designing the system or extracting higher-performing features.

There is also some research proposed recently related to the RQ2 in this thesis. In 2020, BERT [Devlin *et al.*, 2018], which is marked as one of the significant breakthroughs of the decade in the NLP field, started to be used for dementia detection. In 2021, wav2vec2.0 [Baevski *et al.*, 2020], a self-supervised end-to-end ASR system, began to be used for extracting the acoustic features for dementia detection. These are both end-to-end structures though these systems are still being used as a black box, and why they can perform excellently on dementia detection-related tasks has not been analysed, but this will hopefully be the focus of future research.

With a view to being used in clinical practice, three classification tasks are designed similarly in Chapter 8. Based on the designed scenarios, the designed framework is expected to be used in clinical practice because the people living with memory problems can select the framework trained with different scenarios according to their needs. In Chapter 9, the effect of age and cognitive decline on speech are considered in the study for improving the prediction of MMSE scores. The experiments show that speech-based automatic dementia detection systems is a promising area of research that can hopefully one day assist clinical as a low-cost, repeatable, non-invasive, and less stressful toolkit.

References

Ackermann, H. & Ziegler, W. (1991). Articulatory deficits in parkinsonian dysarthria: an acoustic analysis. *Journal of Neurology, Neurosurgery & Psychiatry*, **54**, 1093–1098. 21

Ackermann, H., Hertrich, I., Daum, I., Scharf, G. & Spieker, S. (1997). Kinematic analysis of articulatory movements in central motor disorders. *Movement disorders*, **12**, 1019–1027. 21

Adhikari, S., Thapa, S., Singh, P., Huo, H., Bharathy, G. & Prasad, M. (2021). A comparative study of machine learning and nlp techniques for uses of stop words by patients in diagnosis of alzheimer's disease. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8, IEEE. 90

Agarap, A.F. (2018). Deep learning using rectified linear units (ReLU). *arXiv preprint arXiv:1803.08375*. 80

Albert, M.S., Heller, H.S. & Milberg, W. (1988). Changes in naming ability with age. *Psychology and aging*, **3**, 173. 29

Alhanai, T., Au, R. & Glass, J. (2017). Spoken language biomarkers for detecting cognitive impairment. In *Proc. of ASRU 2017*, IEEE. 101, 108, 113, 173

Almor, A., Kempler, D., MacDonald, M.C., Andersen, E.S. & Tyler, L.K. (1999). Why do alzheimer patients have difficulty with pronouns? working memory, semantics, and reference in comprehension and production in alzheimer's disease. *Brain and language*, **67**, 202–227. 93

Alzheimer's Association (2022). Mild cognitive impairment. https://www.alz.org/alzheimers-dementia/what-is-dementia/related_conditions/mild-cognitive-impairment. 20

Alzheimer's Disease International (2022). Dementia statistics. https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/ [Online]. 2

Alzheimer's Society (2018). What is mixed dementia? https://www.alzheimers.org.uk/blog/what-is-mixed-dementia [Online; accessed 17 August 2018]. 17

Alzheimer's Society (2020). Symptoms of dementia. https://www.nhs.uk/conditions/dementia/symptoms/ [Online; accessed 21 June 2020]. 19

Alzheimer's Society (2022). What is vascular dementia. https://www.alzheimers.org.uk/about-dementia/types-dementia/vascular-dementia [Online; accessed 21 June 2022]. 18

Ammar, R.B. & Ayed, Y.B. (2018). Speech processing for early Alzheimer disease diagnosis: Machine learning based approach. In *International Conference on Computer Systems and Applications (AICCSA)*, 1–8, IEEE. 89

Angelopoulou, G., Kasselimis, D., Makrydakis, G., Varkanitsa, M., Roussos, P., Goutsos, D., Evdokimidis, I. & Potagas, C. (2018). Silent pauses in aphasia. *Neuropsychologia*, **114**, 41–49. 122

Arthurs, E., Steele, R.J., Hudson, M., Baron, M., Thombs, B.D. & Group, C.C.S.R. (2012). Are scores on english and french versions of the phq-9 comparable? an assessment of differential item functioning. *PloS one*, **7**, e52028. 29

Ash, S., McMillan, C., Gross, R.G., Cook, P., Gunawardena, D., Morgan, B., Boller, A., Siderowf, A. & Grossman, M. (2012). Impairments of speech fluency in Lewy body spectrum disorder. *Brain and language*, **120**, 290–302. 122

Atal, B.S. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, **64**, 460–475. 47

Avila, R. & Porto, F.H.d.G. (2020). From diagnosis to rehabilitation: Report of a clinical case of Alzheimer's disease: Implementation of person-centered care. *Alzheimer's & Dementia*, **16**, e042866. 126

Baddeley, A.D. (1983). Working memory. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, **302**, 311–324. 21

Baek, M.J., Kim, H.J., Ryu, H.J., Lee, S.H., Han, S.H., Na, H.R., Chang, Y., Chey, J.Y. & Kim, S. (2011). The usefulness of the story recall test in patients with mild cognitive impairment and Alzheimer's disease. *Aging, Neuropsychology, and Cognition*, **18**, 214–229. 31

Baevski, A., Zhou, H., Mohamed, A. & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*. 64, 118, 180, 181

Bahdanau, D., Cho, K. & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*. 44, 45

Balagopalan, A. & Novikova, J. (2021). Comparing acoustic-based approaches for Alzheimer's disease detection. *arXiv preprint arXiv:2106.01555*. 64, 118, 180

Balagopalan, A., Eyre, B., Rudzicz, F. & Novikova, J. (2020). To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection. *arXiv preprint arXiv:2008.01551*. 61, 96

Banovic, S., Zunic, L.J. & Sinanovic, O. (2018). Communication difficulties as a result of dementia. *Materia socio-medica*, **30**, 221. 126

Baum, S.R. & Pell, M.D. (1999). The neural bases of prosody: Insights from lesion studies and neuroimaging. *Aphasiology*, **13**, 581–608. 22

Bayles, K.A., Kaszniak, A.W. & Tomoeda, C.K. (1987). *Communication and cognition in normal aging and dementia.*. College-Hill Press/Little, Brown & Co. 17

Becker, J.T., Boiler, F., Lopez, O.L., Saxton, J. & McGonigle, K.L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, **51**, 585–594. 8, 54, 55, 157

Beltrami, D., Gagliardi, G., Rossini Favretti, R., Ghidoni, E., Tamburini, F. & Calzà, L. (2018). Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline? *Frontiers in aging neuroscience*, **10**, 369. 4, 42

Bengio, Y., Ducharme, R., Vincent, P. & Janvin, C. (2003). A neural probabilistic language model. *The journal of machine learning research*, **3**, 1137–1155. 39, 40

Berisha, V., Wang, S., LaCross, A. & Liss, J. (2015). Tracking discourse complexity preceding Alzheimer's disease diagnosis: a case study comparing the press conferences of presidents ronald reagan and george herbert walker bush. *Journal of Alzheimer's Disease*, **45**, 959–963. 3, 23

Bier, J.C., Ventura, M., Donckels, V., Van Eyll, E., Claes, T., Slama, H., Fery, P., Vokaer, M. & Pandolfo, M. (2004). Is the addenbrooke's cognitive examination effective to detect frontotemporal dementia? *Journal of neurology*, **251**, 428–431. 28

Bird, H., Ralph, M.A.L., Patterson, K. & Hodges, J.R. (2000). The rise and fall of frequency and imageability: Noun and verb production in semantic dementia. *Brain and language*, **73**, 17–49. 3, 23

Blackburn, D.J., Wakefield, S., Shanks, M.F., Harkness, K., Reuber, M. & Venneri, A. (2014). Memory difficulties are not always a sign of incipient dementia: a review of the possible causes of loss of memory efficiency. *British medical bulletin*, **112**, 71–81. 19

Boersma, P. (2011). Praat: doing phonetics by computer [computer program]. *http://www. praat. org/*. 122

Boersma, P. & Weenink, D. (1996). Praat, a system for doing phonetics by computer, version 3.4. *Institute of Phonetic sciences of the University of Amsterdam, Report*, **132**, 182. 41

Bolt, R.H., Cooper, F.S., David Jr, E.E., Denes, P.B., Pickett, J.M. & Stevens, K.N. (1970). Speaker identification by speech spectrograms: a scientists' view of its reliability for legal purposes. *The Journal of the Acoustical Society of America*, **47**, 597–612. 47

Borod, J.C., Goodglass, H. & Kaplan, E. (1980). Normative data on the boston diagnostic aphasia examination, parietal lobe battery, and the boston naming test. *Journal of Clinical and Experimental Neuropsychology*, **2**, 209–215. 29

Bschor, T., Kühl, K.P. & Reischies, F.M. (2001). Spontaneous speech of patients with dementia of the Alzheimer type and mild cognitive impairment. *International psychogeriatrics*, **13**, 289. 23

Budhkar, A. & Rudzicz, F. (2018). Augmenting word2vec with latent dirichlet allocation within a clinical application. *arXiv preprint arXiv:1808.03967*. 89

Burke, D.M. & Shafto, M.A. (2011). Language and aging. In *The handbook of aging and cognition*, 381–451, Psychology Press. 23

Caccappolo-Van Vliet, E., Manly, J., Tang, M.X., Marder, K., Bell, K. & Stern, Y. (2003). The neuropsychological profiles of mild Alzheimer's disease and questionable dementia as compared to age-related cognitive decline. *Journal of the International Neuropsychological Society*, **9**, 720–732. 30

Campbell, E.L., Docío-Fernández, L., Raboso, J.J. & García-Mateo, C. (2020). Alzheimer's dementia detection from audio and text modalities. *arXiv preprint arXiv:2008.04617*. 5, 37, 121

Carlozzi, N.E., Grech, J. & Tulsky, D.S. (2013). Memory functioning in individuals with traumatic brain injury: An examination of the wechsler memory scale–fourth

edition (WMS–IV). *Journal of clinical and experimental neuropsychology*, **35**, 906–914. 29

Central, D.C. (2020). Stages of Alzheimer's & dementia: Durations scales used to measure progression (gds, fast cdr). https://www.dementiacarecentral.com/aboutdementia/facts/stages/ [Online; accessed 21 June 2020]. 20, 24

Cerhan, J.H., Ivnik, R.J., Smith, G.E., Tangalos, E.C., Petersen, R.C. & Boeve, B.F. (2002). Diagnostic utility of letter fluency, category fluency, and fluency difference scores in Alzheimer's disease. *The Clinical Neuropsychologist*, **16**, 35–42. 30

Chatwin, J. (2014). Conversation analysis as a method for investigating interaction in care home environments. *Dementia*, **13**, 737–746. 30, 31

Chen, J., Ye, J., Tang, F. & Zhou, J. (2021). Automatic detection of Alzheimer's disease using spontaneous speech only. In *Proc. INTERSPEECH 2021*, 3830–3834, ISCA. 64, 96

Chen, L., Tao, J., Ghaffarzadegan, S. & Qian, Y. (2018). End-to-end neural network based automated speech scoring. In *Proc. ICASSP 2018*, 6234–6238, IEEE. 102

Chen, S.F. & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, **13**, 359–394. 82

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*. 43, 44, 45

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K. & Bengio, Y. (2015). Attention-based models for speech recognition. *arXiv preprint arXiv:1506.07503*. 45

Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*. 42, 80, 104

Clark, C.L. (2005). *LabVIEW digital signal processing*. Tata McGraw-Hill Education. 104

Cohan, M. (2013). Stages of dementia: An overview. *End-Stage Dementia Care*, 23–32. 19

Collobert, R. & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, 160–167. 156

Colombo, L., Brivio, C., Benaglio, I., Siri, S. & Cappa, S. (2000). Alzheimer patients' ability to read words with irregular stress. *Cortex*, **36**, 703–714. 22

Colombo, L., Fonti, C. & Cappa, S. (2004). The impact of lexical-semantic impairment and of executive dysfunction on the word reading performance of patients with probable Alzheimer dementia. *Neuropsychologia*, **42**, 1192–1202. 22

Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., Doumouro, C., Gisselbrecht, T., Caltagirone, F., Lavril, T. *et al.* (2018). Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*. 121

Covington, M.A., He, C., Brown, C., Naci, L. & Brown, J. (2006). How complex is that sentence? a proposed revision of the rosenberg and abbeduto d-level scale. 4, 6

Crichton, R.G. & Fallside, F. (1974). Linear prediction model of speech production with applications to deaf speech training. In *Proceedings of the Institution of Electrical Engineers*, 8, 865–873, IET. 47

Croisile, B., Ska, B., Brabant, M.J., Duchene, A., Lepage, Y., Aimard, G. & Trillet, M. (1996). Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain and language*, **53**, 1–19. 23

Cummins, N., Pan, Y., Ren, Z., Fritsch, J., Nallanthighal, V.S., Christensen, H., Blackburn, D., Schuller, B.W., Doss, M.M., Strik, H. *et al.*

(2020). A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition. In *Proc. INTERSPEECH 2020*, 2182–2186, ISCA. 8, 61, 64, 96, 163

Curriculum, N.H. (2022). Patient health questionnaire-9 (PHQ-9). https://www.hiv.uw.edu/page/mental-health-screening/phq-9, 2022-06-30. 29

Davis, S. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, **28**, 357–366. 41, 47, 101

De Cheveigné, A. & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, **111**, 1917–1930. 41

de la Fuente Garcia, S., Ritchie, C. & Luz, S. (2020). Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. *Journal of Alzheimer's Disease*, 1–27. 37

Deb, S., Hare, M. & Prior, L. (2007). Symptoms of dementia among adults with down's syndrome: a qualitative study. *Journal of Intellectual Disability Research*, **51**, 726–739. 126

Degottex, G., Kane, J., Drugman, T., Raitio, T. & Scherer, S. (2014). COVAREP—a collaborative voice analysis repository for speech technologies. In *Proc. ICASSP 2014*, 960–964, IEEE. 64

Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P. & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**, 788–798. 46, 47, 61, 156, 176

Deng, L., Hinton, G. & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *Proc. ICASSP 2013*, 8599–8603, IEEE. 157

Deterding, D. (2001). The measurement of rhythm: A comparison of singapore and british english. *Journal of phonetics*, **29**, 217–230. 22

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. 39, 45, 96, 181

D'haeseleer, E., Meerschman, I., Claeys, S., Leyns, C., Daelman, J. & Van Lierde, K. (2017). Vocal quality in theater actors. *Journal of Voice*, **31**, 510–e7. 22

Dobry, G., Hecht, R.M., Avigal, M. & Zigel, Y. (2011). Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**, 1975–1985. 156

Donnell, A.J., Pliskin, N., Holdnack, J., Axelrod, B. & Randolph, C. (2007). Rapidly-administered short forms of the wechsler adult intelligence scale—3rd edition. *Archives of Clinical Neuropsychology*, **22**, 917–924. 29

Edwards, E., Dognin, C., Bollepalli, B., Singh, M. & Analytics, V. (2020). Multiscale system for Alzheimer's dementia recognition through spontaneous speech. In *Proc. INTERSPEECH 2020*, 2197–2201, ISCA. 61, 106

Egas López, J.V., Tóth, L., Hoffmann, I., Kálmán, J., Pákáski, M. & Gosztolya, G. (2019). Assessing alzheimer's disease from speech using the i-vector approach. In *International Conference on Speech and Computer*, 289–298, Springer. 46

Ehernberger Hamilton, H. (1994). Conversations with an Alzheimer's patient: An interactional sociolonguistic study. 20

Elsey, C., Drew, P., Jones, D., Blackburn, D., Wakefield, S., Harkness, K., Venneri, A. & Reuber, M. (2015). Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics. *Patient Education and Counseling*, **98**, 1071–1077. 2, 66

Englund, B., Brun, A., Gustafson, L., Passant, U., Mann, D., Neary, D. & Snowden, J. (1994). Clinical and neuropathological criteria for frontotemporal dementia. *J Neurol Neurosurg Psychiatry*, **57**, 416–8. 18

Eskenazi, L., Childers, D.G. & Hicks, D.M. (1990). Acoustic correlates of vocal quality. *Journal of Speech, Language, and Hearing Research*, **33**, 298–306. 22

Eyben, F., Wöllmer, M. & Schuller, B. (2010). OpenSMILE: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, 1459–1462. 44, 60

Eyben, F., Weninger, F., Gross, F. & Schuller, B. (2013). Recent developments in OpenSMILE, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, 835–838. 41, 64, 108, 130

Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S. *et al.* (2015). The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, **7**, 190–202. 62, 64

Farzana, S. & Parde, N. (2020). Exploring MMSE score prediction using verbal and non-verbal cues. In *Proc. INTERSPEECH 2020*, 2207–2211, ISCA. 39, 61, 96

Fayek, H.M., Lech, M. & Cavedon, L. (2015). Towards real-time speech emotion recognition using deep neural networks. In *2015 international conference on signal processing and communication systems (ICSPCS)*, 1–5, IEEE. 5, 47, 48

Fearnley, J.M. & Lees, A.J. (1991). Ageing and parkinson's disease: substantia nigra regional selectivity. *Brain*, **114**, 2283–2301. 18

Fedorova, A., Glembek, O., Kinnunen, T. & Matějka, P. (2015). Exploring ANN back-ends for i-vector based speaker age estimation. In *Sixteenth Annual Conference of the International Speech Communication Association*. 156

Feldman, H.H., Jacova, C., Robillard, A., Garcia, A., Chow, T., Borrie, M., Schipper, H.M., Blair, M., Kertesz, A. & Chertkow, H. (2008). Diagnosis and treatment of dementia: 2. diagnosis. *Cmaj*, **178**, 825–836. 24, 25, 26

Finkel, S.I. & Woodson, C. (1997). History and physical examination of elderly patients with dementia. *International psychogeriatrics*, **9**, 71–75. 26

Folstein, M.F., Folstein, S.E. & McHugh, P.R. (1975). "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, **12**, 189–198. 28

Forbes, K.E., Venneri, A. & Shanks, M.F. (2002). Distinct patterns of spontaneous speech deterioration: an early predictor of Alzheimer's disease. *Brain and Cognition*, **48**, 356–361. 21

Forbes, K.E., Shanks, M.F. & Venneri, A. (2004). The evolution of dysgraphia in Alzheimer's disease. *Brain research bulletin*, **63**, 19–24. 21

Forbes-McKay, K.E. & Venneri, A. (2005). Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurological sciences*, **26**, 243–254. 75

Forbes-McKay, K.E., Ellis, A.W., Shanks, M.F. & Venneri, A. (2005). The age of acquisition of words produced in a semantic fluency task can reliably differentiate normal from pathological age related cognitive decline. *Neuropsychologia*, **43**, 1625–1632. 21

Fraser, K.C., Meltzer, J.A. & Rudzicz, F. (2016). Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, **49**, 407–422. 3, 4, 6, 57, 58, 89, 101, 133

Fraser, K.C., Fors, K.L. & Kokkinakis, D. (2019). Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer speech & language*, **53**, 121–139. 57

Fritsch, J., Bergler, C., Wankerl, S. & Nöth, E. (2018). Automatic diagnosis of Alzheimer's disease using neural networks language models. In *Proc. INTERSPEECH 2018*, 5841–5845, ISCA. 40, 57, 58, 59, 75, 139, 173

Fu, Z., Haider, F. & Luz, S. (2020). Predicting mini-mental status examination scores through paralinguistic acoustic features of spontaneous speech. In *Annual International*

Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 5548–5552, IEEE. 7, 155, 170, 176

García-Caballero, A., García-Lado, I., González-Hermida, J., Recimil, M., Area, R., Manes, F., Lamas, S. & Berrios, G. (2006). Validation of the spanish version of the addenbrooke's cognitive examination in a rural community in spain. *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, **21**, 239–245. 28

Gauder, L., Pepino, L., Ferrer, L. & Riera, P. (2021). Alzheimer disease recognition using speech-based embeddings from pre-trained models. In *Proc. INTERSPEECH 2021*, 3795–3799, ISCA. 64, 118, 180

Gauthier, S. (2001). *Management of dementia*. CRC Press. 18

Gayraud, F., Lee, H.R. & Barkat-Defradas, M. (2011). Syntactic and lexical context of pauses and hesitations in the discourse of alzheimer patients and healthy elderly subjects. *Clinical linguistics & phonetics*, **25**, 198–209. 124

Ghahabi, O., Bonafonte, A., Hernando, J. & Moreno, A. (2016). Deep neural networks for i-vector language identification of short utterances in cars. In *Proc. INTERSPEECH 2016*, 367–371, ISCA. 46, 47

Ghahremani, P., Nidadavolu, P.S., Chen, N., Villalba, J., Povey, D., Khudanpur, S. & Dehak, N. (2018). End-to-end deep neural network age estimation. In *Proc. INTERSPEECH 2018*, 277–281, ISCA. 156

Giles, E., Patterson, K. & Hodges, J.R. (1996). Performance on the boston cookie theft picture description task in patients with early dementia of the Alzheimer's type: missing information. *Aphasiology*, **10**, 395–408. 23

Gilks, W.R., Richardson, S. & Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press. 40

Glosser, G. & Deser, T. (1991). Patterns of discourse production among neurological patients with fluent language disorders. *Brain and language*, **40**, 67–88. 24

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, **10**, 1–309. 39, 79

Gómez-Vilda, P., Rodellar-Biarge, V., Nieto-Lluis, V., de Ipiña, K.L., Álvarez-Marquina, A., Martínez-Olalla, R., Ecay-Torres, M. & Martinez-Lage, P. (2015). Phonation biomechanic analysis of Alzheimer's disease cases. *Neurocomputing*, **167**, 83–93. 22

Gónzalez Atienza, M., González López, J.A., Peinado, A.M. *et al.* (2021). An automatic system for dementia detection using acoustic and linguistic features. In *Proc. INTERSPEECH 2021*, ISCA. 121

Goodglass, H. (1980). Naming disorders in aphasia and aging. *Language and Comunication in the Ederly*, 37–45. 29

Goodglass, H. & Kaplan, E. (1972). *The assessment of aphasia and related disorders*. Lea & Febiger. 31

Goodglass, H., Kaplan, E. & Weintraub, S. (1983). *Boston naming test*. Lea & Febiger. 28

Graham, K.S., Patterson, K., Pratt, K.H. & Hodges, J.R. (2001). Can repeated exposure to "forgotten" vocabulary help alleviate word-finding difficulties in semantic dementia? an illustrative case study. *Neuropsychological Rehabilitation*, **11**, 429–454. 94

Green, P. & Kramar, K. (1983). Auditory comprehension test (renamed story recall test). 31

Groves-Wright, K., Neils-Strunjas, J., Burnett, R. & O'Neill, M.J. (2004). A comparison of verbal and written language in Alzheimer's disease. *Journal of Communication Disorders*, **37**, 109–130. 23

Guinn, C.I. & Habash, A. (2012). Language analysis of speakers with dementia of the Alzheimer's type. In *AAAI Fall Symposium Series*. 158

Haider, F., De La Fuente, S. & Luz, S. (2019). An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, **14**, 272–281. 41, 57, 58, 155, 156

Haider, F., de la Fuente, S., Albert, P. & Luz, S. (2020). Affective speech for alzheimer's dementia recognition. In *LREC 2020 Language Resources and Evaluation Conference*, 67. 42, 124

Hamilton, H.E. (2005). *Conversations with an Alzheimer's patient: An interactional sociolinguistic study*. Cambridge University Press. 17

Han, K., Yu, D. & Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth annual conference of the international speech communication association*. 47, 48

Hermansky, H. (1990). Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, **87**, 1738–1752. 41

Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G. & Roche-Bergua, A. (2018). Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, **10**, 260–268. 57, 58, 101

Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B. *et al.* (2017). CNN architectures for large-scale audio classification. In *Proc. ICASSP 2017*, 131–135, IEEE. 61, 64, 118

Hier, D.B., Hagenlocker, K. & Shindler, A.G. (1985). Language disintegration in dementia: Effects of etiology and severity. *Brain and language*, **25**, 117–133. 23, 24

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N. *et al.* (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, **29**, 82–97. 42

Hinton, G.E. & Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. *science*, **313**, 504–507. 102

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **6**, 107–116. 43

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, **9**, 1735–1780. 43

Hoffmann, I., Nemeth, D., Dye, C.D., Pákáski, M., Irinyi, T. & Kálmán, J. (2010). Temporal parameters of spontaneous speech in Alzheimer's disease. *International journal of speech-language pathology*, **12**, 29–34. 4, 41

Holston, E.C. (2005). Stigmatization in Alzheimer's disease research on african american elders. *Issues in mental health nursing*, **26**, 1103–1127. 18

Horii, Y. (1979). Fundamental frequency perturbation observed in sustained phonation. *Journal of Speech, Language, and Hearing Research*, **22**, 5–19. 3, 22

Horii, Y. (1980). Vocal shimmer in sustained phonation. *Journal of Speech, Language, and Hearing Research*, **23**, 202–209. 22

Horley, K., Reid, A. & Burnham, D. (2010). Emotional prosody perception and production in dementia of the Alzheimer's type. *Journal of Speech, Language, and Hearing Research*. 5, 41

Horwitz-Martin, R.L., Quatieri, T.F., Lammert, A.C., Williamson, J.R., Yunusova, Y., Godoy, E., Mehta, D.D. & Green, J.R. (2016). Relation of automatically extracted formant trajectories with intelligibility loss and speaking rate decline in amyotrophic lateral sclerosis. In *Proc. INTERSPEECH 2016*, 1205–1209, ISCA. 22

Howard, J. & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*. 180

Huang, C.W. & Narayanan, S.S. (2016). Attention assisted discovery of sub-utterance structure in speech emotion recognition. In *Proc. INTERSPEECH 2016*, 1387–1391, ISCA. 5, 47, 48, 49

Huang, R. & Ma, C. (2006). Toward a speaker-independent real-time affect detection system. In *18th International Conference on Pattern Recognition (ICPR)*, vol. 1, 1204–1207, IEEE. 47, 48

Hubel, D.H. & Wiesel, T.N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, **148**, 574. 44

Hubel, D.H. & Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, **160**, 106. 44

Hussain, M., Bird, J.J. & Faria, D.R. (2018). A study on CNN transfer learning for image classification. In *UK Workshop on computational Intelligence*, 191–202, Springer. 180

Inoue, T., Tanaka, T., Nakagawa, S., Nakato, Y., Kameyama, R., Boku, S., Toda, H., Kurita, T. & Koyama, T. (2012). Utility and limitations of PHQ-9 in a clinic specializing in psychiatric care. *BMC psychiatry*, **12**, 73. 27

Izawa, Y., Urakami, K., Kojima, T. & Ohama, E. (2009). Wechsler adult intelligence scale, (WAIS-III): Usefulness in the early detection of Alzheimer's disease. *Yonago Acta Medica*, **52**, 11–20. 29

Jack, C.R., Wiste, H.J., Weigand, S.D., Knopman, D.S., Lowe, V., Vemuri, P., Mielke, M.M., Jones, D.T., Senjem, M.L., Gunter, J.L. *et al.* (2013). Amyloid-first and neurodegeneration-first profiles characterize incident amyloid PET positivity. *Neurology*, **81**, 1732–1740. 19

Jack Jr, C.R., Knopman, D.S., Jagust, W.J., Petersen, R.C., Weiner, M.W., Aisen, P.S., Shaw, L.M., Vemuri, P., Wiste, H.J., Weigand, S.D. *et al.* (2013). Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, **12**, 207–216. 19

Jack Jr, C.R., Bennett, D.A., Blennow, K., Carrillo, M.C., Dunn, B., Haeberlein, S.B., Holtzman, D.M., Jagust, W., Jessen, F., Karlawish, J. *et al.* (2018). Nia-aa research framework: toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, **14**, 535–562. 18

Jadoul, Y., Thompson, B. & De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, **71**, 1–15. 157

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M.L. & Ogar, J. (2014). Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 27–37. 4, 5, 6, 37, 38, 122

Jellinger, K.A. & Korczyn, A.D. (2018). Are dementia with lewy bodies and Parkinson's disease dementia the same disease? *BMC medicine*, **16**, 1–16. 19

Kalluri, S.B., Vijayasenan, D. & Ganapathy, S. (2019). A deep neural network based end to end model for joint height and age estimation from short duration speech. In *Proc. ICASSP 2019*, 6580–6584, IEEE. 156

Karlekar, S., Niu, T. & Bansal, M. (2018). Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models. *arXiv preprint arXiv:1804.06440*. 89

Karlik, B. & Olgac, A.V. (2011). Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, **1**, 111–122. 43

Kelly, F. & Harte, N. (2011). Effects of long-term ageing on speaker verification. In *European Workshop on Biometrics and Identity Management*, 113–124, Springer. 7, 155

Kelly, F., Saeidi, R., Harte, N. & Leeuwen, D.A.v. (2014). Effect of long-term ageing on i-vector speaker verification. In *Proc. INTERSPEECH 2014*, ISCA. 157, 163

Kemper, S., Greiner, L.H., Marquis, J.G., Prenovost, K. & Mitzner, T.L. (2001). Language decline across the life span: Findings from the nun study. *Psychology and aging*, **16**, 227. 23

Khodabakhsh, A., Yesil, F., Guner, E. & Demiroglu, C. (2015). Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in turkish conversational speech. *EURASIP Journal on Audio, Speech, and Music Processing*, **2015**, 9. 3, 4, 5, 6, 37, 38

King, D.A., Caine, E.D. & Cox, C. (1993). Influence of depression and age on selected cognitive functions. *The Clinical Neuropsychologist*, **7**, 443–453. 29

Kingma, D.P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 86, 130

Klimova, B., Maresova, P., Valis, M., Hort, J. & Kuca, K. (2015). Alzheimer's disease and language impairments: social intervention and medical treatment. *Clinical interventions in aging*, **10**, 1401. 19, 20, 21

Kneser, R. & Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proc. ICASSP 1995*, 181–184, IEEE. 40

Kong, W., Jang, H., Carenini, G. & Field, T. (2019). A neural model for predicting dementia from language. In *Machine Learning for Healthcare Conference*, 270–286, PMLR. 59

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P.H. *et al.* (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, **1**, 112–124. 5, 41, 125

Koo, J., Lee, J.H., Pyo, J., Jo, Y. & Lee, K. (2020). Exploiting multi-modal features from pre-trained networks for Alzheimer's dementia recognition. *arXiv preprint arXiv:2009.04070*. 39, 61, 96, 118

Kornblith, S., Shlens, J. & Le, Q.V. (2019). Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2661–2671. 180

Kroenke, K. & Spitzer, R.L. (2002). The PHQ-9: a new depression diagnostic and severity measure. *Psychiatric annals*, **32**, 509–515. 29

Kwon, O.W., Chan, K., Hao, J. & Lee, T.W. (2003). Emotion recognition by speech signals. In *Eighth European conference on speech communication and technology*. 47, 48

Kye, S.M., Jung, Y., Lee, H.B., Hwang, S.J. & Kim, H. (2020). Meta-learning for short utterance speaker recognition with imbalance length pairs. *arXiv preprint arXiv:2004.02863*. 139

Laine, M., Laakso, M., Vuorinen, E. & Rinne, J. (1998). Coherence and informativeness of discourse in two dementia types. *Journal of Neurolinguistics*, **11**, 79–87. 24

Landauer, T.K. & Dumais, S.T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, **104**, 211. 39

Landspítalinn, T.J., Snædal, J. & Jónsson, J. (1993). A progression in the neuropsychological decline of icelandic patients with probable or possible dementia of the Alzheimer's type: A longitudinal study. *Aging Clinical and Experimental Research*, **5**, 217–228. 20

Lanza, C., Knoerzer, O., Weber, M. & Riepe, M.W. (2014). Autonomous spatial orientation in patients with mild to moderate Alzheimer's disease by using mobile assistive devices: a pilot study. *Journal of Alzheimer's disease*, **42**, 879–884. 20

Le, X., Lancashire, I., Hirst, G. & Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three british novelists. *Literary and linguistic computing*, **26**, 435–461. 75

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W. & Jackel, L.D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, **1**, 541–551. 44

LeCun, Y., Haffner, P., Bottou, L. & Bengio, Y. (1999). Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, 319–345, Springer. 104

Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S. & Narayanan, S. (2004). Emotion recognition based on phoneme classes. In *Eighth International Conference on Spoken Language Processing*. 47, 48

Lehiste, I. (1970). Suprasegmentals, cambridge, massachusetts & london, uk. 22

Li, J., Yu, J., Ye, Z., Wong, S., Mak, M., Mak, B., Liu, X. & Meng, H. (2021). A comparative study of acoustic and linguistic features classification for Alzheimer's disease detection. In *Proc. ICASSP 2021*, 6423–6427, IEEE. 57, 59

Lin, D. (1996). On the structural complexity of natural language sentences. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*. 38

Linville, S.E. (1996). The sound of senescence. *Journal of voice*, **10**, 190–200. 155

López, J.V.E., Tóth, L., Hoffmann, I., Kálmán, J., Pákáski, M. & Gosztolya, G. (2019). Assessing Alzheimer's disease from speech using the i-vector approach. In *International Conference on Speech and Computer*, 289–298, Springer. 61, 156

López-de Ipiña, K., Alonso, J.B., Travieso, C.M., Solé-Casals, J., Egiraun, H., Faundez-Zanuy, M., Ezeiza, A., Barroso, N., Ecay-Torres, M., Martinez-Lage, P. *et al.* (2013). On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. *Sensors*, **13**, 6730–6745. 41

Lopez-de Ipiña, K., Alonso, J., Solé-Casals, J., Barroso, N., Henriquez, P., Faundez-Zanuy, M., Travieso, C., Ecay-Torres, M., Martinez-Lage,

P. & Eguiraun, H. (2015). On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature. *Cognitive Computation*, **7**, 44–55. 101

López-de Ipiña, K., Solé-Casals, J., Eguiraun, H., Alonso, J.B., Travieso, C.M., Ezeiza, A., Barroso, N., Ecay-Torres, M., Martinez-Lage, P. & Beitia, B. (2015). Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach. *Computer Speech & Language*, **30**, 43–60. 101

Luck, J.E. (1969). Automatic speaker verification using cepstral measurements. *The Journal of the Acoustical Society of America*, **46**, 1026–1032. 47

Lugger, M. & Yang, B. (2007). The relevance of voice quality features in speaker independent emotion recognition. In *Proc. ICASSP 2007*, 17–20, IEEE. 46, 48

Luong, M.T., Pham, H. & Manning, C. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*. 45

Luz, S. (2009). Locating case discussion segments in recorded medical team meetings. In *Proceedings of the third workshop on Searching spontaneous conversational speech*, 21–30. 41

Luz, S. (2017). Longitudinal monitoring and detection of Alzheimer's type dementia from spontaneous speech data. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, 45–46, IEEE. 57, 58, 101, 133

Luz, S., de la Fuente, S. & Albert, P. (2018). A method for analysis of patient speech in dialogue for dementia detection. *arXiv preprint arXiv:1811.09919*. 41, 106, 158

Luz, S., Haider, F., de la Fuente, S., Fromm, D. & MacWhinney, B. (2020a). Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. *arXiv preprint arXiv:2004.06833*. 10, 60, 61, 62, 108, 122, 124, 133, 169

Luz, S., Haider, F., de la Fuente, S., Fromm, D. & MacWhinney, B. (2020b). Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. In *Proc. INTERSPEECH 2020*, ISCA. 54, 59, 70

Luz, S., Haider, F., de la Fuente, S., Fromm, D. & MacWhinney, B. (2021). Detecting cognitive decline using speech only: The ADReSSo challenge. *medRxiv*. 10, 54, 61, 63

Ma, J., Sethu, V., Ambikairajah, E. & Lee, K.A. (2016a). Twin model G-PLDA for duration mismatch compensation in text-independent speaker verification. In *Proc. INTERSPEECH 2016*, 1853–1857, ISCA. 139, 173

Ma, X., Yang, H., Chen, Q., Huang, D. & Wang, Y. (2016b). Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 35–42, ACM. 43

Maas, A.L., Hannun, A.Y. & Ng, A.Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICM*. 107, 145

MacKay, A., Connor, L.T. & Storandt, M. (2005). Dementia does not explain correlation between age and scores on boston naming test. *Archives of Clinical Neuropsychology*, **20**, 129–133. 29

MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press. 56

MacWhinney, B. (2017). Tools for analyzing talk part 2: The CLAN program. *Pittsburgh, PA: Carnegie Mellon University. Retrieved from http://talkbank. org/manuals/-CLAN. pdf*. 62

Magerman, D.M. (1995). Statistical decision-tree models for parsing. *arXiv preprint cmp-lg/9504030*. 38

Mahmoodi, D., Marvi, H., Taghizadeh, M., Soleimani, A., Razzazi, F. & Mahmoodi, M. (2011). Age estimation based on speech features and support vector

machine. In *Third Computer Science and Electronic Engineering Conference (CEEC)*, 60–64, IEEE, Athens, Greece. 155

Manohar, V., Povey, D. & Khudanpur, S. (2017). Jhu kaldi system for Arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning. In *Proc. of ASRU 2017*, 346–352, IEEE. 129, 146

Martínez-Sánchez, F., JJ, G.M., Pérez, E., Carro, J. & Arana, J.M. (2012). Expressive prosodic patterns in individuals with Alzheimer's disease. *Psicothema*, **24**, 16–21. 5, 41, 42

Martínez-Sánchez, F., Meilán, J.J., Vera-Ferrandiz, J.A., Carro, J., Pujante-Valverde, I.M., Ivanova, O. & Carcavilla, N. (2017). Speech rhythm alterations in spanish-speaking individuals with Alzheimer's disease. *Aging, Neuropsychology, and Cognition*, **24**, 418–434. 115, 122

Masrani, V. (2018). *Detecting dementia from written and spoken language*. Ph.D. thesis, University of British Columbia. 58, 59

Mathuranath, P., Nestor, P., Berrios, G., Rakowicz, W. & Hodges, J. (2000). A brief cognitive test battery to differentiate alzheimer's disease and frontotemporal dementia. *Neurology*, **55**, 1613–1620. 28

Mathuranath, P., Hodges, J.R., Mathew, R., Cherian, P.J., George, A. & Bak, T.H. (2004). Adaptation of the ace for a malayalam speaking population in southern india. *International journal of geriatric psychiatry*, **19**, 1188–1194. 28

Maxim, J. & Bryan, K. (1994). *Language of the elderly: A clinical perspective*. Whurr Pub Ltd. 23

McCulloch, W.S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, **5**, 115–133. 42

McKeith, I. (2007). Dementia with lewy bodies. *Handbook of clinical Neurology*, **84**, 531–548. 19

McKeith, I.G., Boeve, B.F., Dickson, D.W., Halliday, G., Taylor, J.P., Weintraub, D., Aarsland, D., Galvin, J., Attems, J., Ballard, C.G. *et al.* (2017). Diagnosis and management of dementia with lewy bodies: Fourth consensus report of the dlb consortium. *Neurology*, **89**, 88–100. 19

McNamara, P., Obler, L.K., Au, R., Durso, R. & Albert, M.L. (1992). Speech monitoring skills in Alzheimer's disease, Parkinson's disease, and normal aging. *Brain and language*, **42**, 38–51. 75, 76, 138

McRae, P.A., Tjaden, K. & Schoonings, B. (2002). Acoustic and perceptual consequences of articulatory rate change in Parkinson disease. *Journal of Speech, Language, and Hearing Research*, **45**, 35–50. 21

Meilan, J.J., Martinez-Sanchez, F., Carro, J., Carcavilla, N. & Ivanova, O. (2018). Voice markers of lexical access in mild cognitive impairment and Alzheimer's disease. *Current Alzheimer Research*, **15**, 111–119. 5, 42

Meilán, J.J., Martínez-Sánchez, F., Martínez-Nicolás, I., Llorente, T.E. & Carro, J. (2020). Changes in the rhythm of speech difference between people with nondegenerative mild cognitive impairment and with preclinical dementia. *Behavioural neurology*, **2020**. 5, 19, 22, 42

Meilán, J.J.G., Martínez-Sánchez, F., Carro, J., López, D.E., Millian-Morell, L. & Arana, J.M. (2014). Speech in Alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders*, **37**, 327–334. 5, 41, 101, 113, 155, 173, 176

Metropolis, N. & Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, **44**, 335–341. 45

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. 39, 79

Mirheidari, B. (2018). *Detecting early signs of dementia in conversation*. Ph.D. thesis, University of Sheffield. 5, 39, 42, 54, 58, 66, 79, 82, 105, 129, 227

Mirheidari, B., Blackburn, D., Reuber, M., Walker, T. & Christensen, H. (2016). Diagnosing people with dementia using automatic conversation analysis. In *Proc. INTERSPEECH 2016*, 1220–1224, ISCA. 42

Mirheidari, B., Blackburn, D., Harkness, K., Walker, T., Venneri, A., Reuber, M. & Christensen, H. (2017). Toward the automation of diagnostic conversation analysis in patients with memory complaints. *Journal of Alzheimer's Disease*, **58**, 373–387. 4, 141

Mirheidari, B., Blackburn, D., O'Malley, R., Walker, T., Venneri, A., Reuber, M. & Christensen, H. (2019a). Computational cognitive assessment: Investigating the use of an intelligent virtual agent for the detection of early signs of dementia. In *Proc. ICASSP 2019*, 2732–2736, IEEE. 41, 66, 82, 101, 105, 129

Mirheidari, B., Blackburn, D., Walker, T., Reuber, M. & Christensen, H. (2019b). Dementia detection using automatic analysis of conversations. *Computer Speech & Language*, **53**, 65–79. 3, 64, 66, 69, 105

Mirheidari, B., Blackburn, D., O'Malley, R., Venneri, A., Walker, T., Reuber, M. & Christensen, H. (2020). Improving cognitive impairment classification by generative neural network-based feature augmentation. In *Proc. INTERSPEECH 2020*, 2527–2531, ISCA. 129, 146, 148, 149, 150, 152, 174, 176

Mirsamadi, S., Barsoum, E. & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *Proc. ICASSP 2017*, 2227–2231, IEEE. 5, 47, 48, 49

Mittal, A., Sahoo, S., Datar, A., Kadiwala, J., Shalu, H. & Mathew, J. (2020). Multi-modal detection of Alzheimer's disease from speech and text. *arXiv preprint arXiv:2012.00096*. 57, 58, 59

Monsch, A.U., Bondi, M.W., Butters, N., Salmon, D.P., Katzman, R. & Thal, L.J. (1992). Comparisons of verbal fluency tasks in the detection of dementia of the Alzheimer type. *Archives of neurology*, **49**, 1253–1258. 30

Morris, R.G. (1987). Articulatory rehearsal in Alzheimer type dementia. *Brain and Language*, **30**, 351–362. 21

Mueller, K.D., Hermann, B., Mecollari, J. & Turkstra, L.S. (2018). Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *Journal of clinical and experimental neuropsychology*, **40**, 917–939. 3, 30, 37, 75

Mundt, J. & King, M. (2003). Predicting early treatment drop out using interactive voice response (IVR). *Alcoholism: Clinical and Experimental Research*, **27**, 28A. 49

Muriel Lezak, D., Bigler, E. & Tranel, D. (2012). Neuropsychological assessment. new york. 30

Nabers, A., Ollesch, J., Schartner, J., Kötting, C., Genius, J., Haussmann, U., Klafki, H., Wiltfang, J. & Gerwert, K. (2016a). An infrared sensor analysing label-free the secondary structure of the abeta peptide in presence of complex fluids. *Journal of biophotonics*, **9**, 224–234. 26

Nabers, A., Ollesch, J., Schartner, J., Kotting, C., Genius, J., Hafermann, H., Klafki, H., Gerwert, K. & Wiltfang, J. (2016b). Amyloid-β-secondary structure distribution in cerebrospinal fluid and blood measured by an immuno-infrared-sensor: a biomarker candidate for alzheimer's disease. *Analytical chemistry*, **88**, 2755–2762. 26

Nair, V. & Hinton, G.E. (2010). Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning 2010*. 43

Nakatsu, R., Nicholson, J. & Tosa, N. (1999). Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, 343–351. 47, 48

Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L. & Chertkow, H. (2005). The montreal

cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, **53**, 695–699. 28

Nasrolahzadeh, M., Mohammadpoori, Z. & Haddadnia, J. (2016a). Analysis of mean square error surface and its corresponding contour plots of spontaneous speech signals in Alzheimer's disease with adaptive wiener filter. *Computers in Human Behavior*, **61**, 364–371. 42

Nasrolahzadeh, M., Mohammadpoory, Z. & Haddadnia, J. (2016b). A novel method for early diagnosis of Alzheimer's disease based on higher-order spectral estimation of spontaneous speech signals. *Cognitive neurodynamics*, **10**, 495–503. 42

Nicholas, M., Obler, L., Albert, M. & Goodglass, H. (1985a). Lexical retrieval in healthy aging. *Cortex*, **21**, 595–606. 29

Nicholas, M., Obler, L.K., Albert, M.L. & Helm-Estabrooks, N. (1985b). Empty speech in Alzheimer's disease and fluent aphasia. *Journal of Speech, Language, and Hearing Research*, **28**, 405–410. 138

Nooteboom, S. *et al.* (1997). The prosody of speech: melody and rhythm. *The handbook of phonetic sciences*, **5**, 640–673. 22

Ober, B.A., Dronkers, N.F., Koss, E., Delis, D.C. & Friedland, R.P. (1986). Retrieval from semantic memory in Alzheimer-type dementia. *Journal of clinical and experimental neuropsychology*, **8**, 75–92. 30

Oechsle, K., Goerth, K., Bokemeyer, C. & Mehnert, A. (2013). Anxiety and depression in caregivers of terminally ill cancer patients: impact on their perspective of the patients' symptom burden. *Journal of palliative medicine*, **16**, 1095–1101. 30

Olson, R.A., Chhanabhai, T. & McKenzie, M. (2008). Feasibility study of the montreal cognitive assessment (moca) in patients with brain metastases. *Supportive care in cancer*, **16**, 1273–1278. 28

O'Malley, R.P.D., Mirheidari, B., Harkness, K., Reuber, M., Venneri, A., Walker, T., Christensen, H. & Blackburn, D. (2021). Fully automated cognitive screening tool based on assessment of speech and language. *Journal of Neurology, Neurosurgery & Psychiatry*, **92**, 12–15. 69

Orimaye, S.O., Wong, J.S.M. & Fernandez, J.S.G. (2016). Deep-deep neural network language models for predicting mild cognitive impairment. In *BAI@ IJCAI*, 14–20. 43

Orimaye, S.O., Wong, J.S., Golden, K.J., Wong, C.P. & Soyiri, I.N. (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC bioinformatics*, **18**, 34. 3, 4, 5, 6, 38, 40, 57, 58, 101

Orimaye, S.O., Wong, J.S.M. & Wong, C.P. (2018). Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia. *PloS one*, **13**, e0205636. 57

Östberg, P., Bogdanović, N. & Wahlund, L.O. (2009). Articulatory agility in cognitive decline. *Folia Phoniatrica et Logopaedica*, **61**, 269–274. 21

Pakhomov, S., Chacon, D., Wicklund, M. & Gundel, J. (2011). Computerized assessment of syntactic complexity in Alzheimer's disease: A case study of iris murdoch's writing. *Behavior research methods*, **43**, 136–144. 3

Pan, Y., Zheng, T. & Chen, C. (2017). I-vector kullback-leibler divisive normalization for PLDA speaker verification. In *Global Conference on Signal and Information Processing (GlobalSIP)*, 56–60, IEEE. 47

Pan, Y., Mirheidari, B., Reuber, M., Venneri, A., Blackburn, D. & Christensen, H. (2019). Automatic hierarchical attention neural network for detecting AD. In *Proc. INTERSPEECH 2019*, 4105–4109, ISCA. 8, 11

Pan, Y., Mirheidari, B., Reuber, M., Venneri, A., Blackburn, D. & Christensen, H. (2020a). Improving detection of Alzheimer's disease using automatic speech

recognition to identify high-quality segments for more robust feature extraction. In *Proc. INTERSPEECH 2020*, 4961–4965, ISCA. 9, 12

Pan, Y., Mirheidari, B., Tu, Z., O'Malley, R., Walker, T., Venneri, A., Reuber, M., Blackburn, D. & Christensen, H. (2020b). Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification. In *Proc. INTERSPEECH 2020*, 4806–4810, ISCA. 9, 12

Pan, Y., Mirheidari, B., Harris, J.M., Thompson, J.C., Jones, M., Snowden, J.S., Blackburn, D. & Christensen, H. (2021a). Using the outputs of different automatic speech recognition paradigms for acoustic-and BERT-based Alzheimer's dementia detection through spontaneous speech. In *Proc. INTERSPEECH 2021*, 3810–3814, ISCA. 63, 64, 96, 118, 179, 180

Pan, Y., Nallanthighal, V.S., Blackburn, D., Christensen, H. & Härmä, A. (2021b). Multi-task estimation of age and cognitive decline from speech. In *Proc. ICASSP 2021*, 7258–7262, IEEE. 10, 12

Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *Proc. ICASSP 2010*, 5206–5210, IEEE. 146

Pappagari, R., Cho, J., Moro-Velazquez, L. & Dehak, N. (2020). Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity. In *Proc. INTERSPEECH 2020*, 2177–2181, ISCA. 61

Pappagari, R., Cho, J., Joshi, S., Moro-Velazquez, L., Zelasko, P., Villalba, J. & Dehak, N. (2021). Automatic detection and assessment of Alzheimer disease using speech and language technologies in low-resource scenarios. In *Proc. INTERSPEECH 2021*, ISCA. 63, 64, 96, 180

Pasquier, F. (1999). Early diagnosis of dementia: neuropsychology. *Journal of neurology*, **246**, 6–15. 3

Pendlebury, S.T., Cuthbertson, F.C., Welch, S.J., Mehta, Z. & Rothwell, P.M. (2010). Underestimation of cognitive impairment by mini-mental state examination versus the montreal cognitive assessment in patients with transient ischemic attack and stroke: a population-based study. *Stroke*, **41**, 1290–1293. 28

Pennebaker, J.W., Francis, M.E. & Booth, R.J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, **71**, 2001. 37

Pennington, J., Socher, R. & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543. 39, 79, 85

Pérez-Toro, P., Bayerl, S., Arias-Vergara, T., Vásquez-Correa, J., Klumpp, P., Schuster, M., Nöth, E., Orozco-Arroyave, J. & Riedhammer, K. (2021). Influence of the interviewer on the automatic assessment of Alzheimer's disease in the context of the ADReSSo challenge. In *Proc. INTERSPEECH 2021*, 3785–3789, ISCA. 64, 96

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*. 180

Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G. & Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Archives of neurology*, **56**, 303–308. 2, 20

Petti, U., Baker, S. & Korhonen, A. (2020). A systematic literature review of automatic alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association*, **27**, 1784–1797. 37

Pistono, A., Jucla, M., Barbeau, E.J., Saint-Aubert, L., Lemesle, B., Calvet, B., Koepke, B., Puel, M. & Pariente, J. (2016). Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimer's disease. *Journal of Alzheimer's Disease*, **50**, 687–698. 122, 124

Politis, M., Wu, K., Molloy, S., G. Bain, P., Chaudhuri, K.R. & Piccini, P. (2010). Parkinson's disease symptoms: the patient's perspective. *Movement Disorders*, **25**, 1646–1651. 19

Pompili, A., Abad, A., de Matos, D.M. & Martins, I.P. (2020a). Pragmatic aspects of discourse production for the automatic identification of Alzheimer's disease. *IEEE Journal of Selected Topics in Signal Processing*, **14**, 261–271. 57, 58, 59

Pompili, A., Rolland, T. & Abad, A. (2020b). The INESC-ID multi-modal system for the ADReSS 2020 challenge. *arXiv preprint arXiv:2005.14646*. 39, 61, 96

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. *et al.* (2011). The kaldi speech recognition toolkit. Tech. rep., IEEE Signal Processing Society. 82, 129

Prosek, R.A., Montgomery, A.A., Walden, B.E. & Hawkins, D.B. (1987). An evaluation of residue features as correlates of voice disorders. *Journal of communication disorders*, **20**, 105–117. 22

Qiao, Y., Yin, X., Wiechmann, D. & Kerz, E. (2021). Alzheimer's disease detection from spontaneous speech through combining linguistic complexity and (dis) fluency features with pretrained language models. *arXiv preprint arXiv:2106.08689*. 64, 96

Quintas, S., Mauclair, J., Woisard, V. & Pinquier, J. (2020). Automatic Prediction of Speech Intelligibility based on X-vectors in the context of Head and Neck Cancer. In *Proc. INTERSPEECH 2020*, ISCA. 156, 165

Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. (2018). Improving language understanding by generative pre-training. 180

Ravanelli, M. & Bengio, Y. (2018a). Interpretable convolutional filters with sincnet. *arXiv preprint arXiv:1811.09725*. 101, 102, 112

Ravanelli, M. & Bengio, Y. (2018b). Speaker recognition from raw waveform with sincnet. In *IEEE Spoken Language Technology Workshop (SLT)*, 1021–1028, IEEE. 5, 9, 47, 102, 107, 173

Ravanelli, M. & Bengio, Y. (2018c). Speech and speaker recognition from raw waveform with sincnet. *arXiv preprint arXiv:1812.05920*. 47

Rentoumi, V., Paliouras, G., Danasi, E., Arfani, D., Fragkopoulou, K., Varlokosta, S. & Papadatos, S. (2017). Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: A computational linguistics analysis. In *International Conference on Cognitive Infocommunications 2017*, 33–38, IEEE. 4, 5, 6, 37, 38

Reubold, U., Harrington, J. & Kleber, F. (2010). Vocal aging effects on f0 and the first formant: A longitudinal analysis in adult speakers. *Speech Communication*, **52**, 638–651. 155

Ringman, J., Younkin, S., Pratico, D., Seltzer, W., Cole, G., Geschwind, D., Rodriguez-Agudelo, Y., Schaffer, B., Fein, J., Sokolow, S. *et al.* (2008). Biochemical markers in persons with preclinical familial Alzheimer disease. *Neurology*, **71**, 85–92. 19

Roark, B., Mitchell, M. & Hollingshead, K. (2007). Syntactic complexity measures for detecting mild cognitive impairment. In *Biological, translational, and clinical language processing*, 1–8. 5, 38

Roark, B., Mitchell, M., Hosom, J.P., Hollingshead, K. & Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *transactions on audio, speech, and language processing*, **19**, 2081–2090. 5, 38, 41, 124, 125

Rohanian, M., Hough, J. & Purver, M. (2021). Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. *arXiv preprint arXiv:2106.15684*. 64, 96, 125

Rohdin, J., Silnova, A., Diez, M., Plchot, O., Matějka, P. & Burget, L. (2018). End-to-end DNN based speaker recognition inspired by i-vector and PLDA. In *Proc. ICASSP 2018*, 4874–4878, IEEE. 5, 46, 47

Rohdin, J., Silnova, A., Diez, M., Plchot, O., Matějka, P., Burget, L. & Glembek, O. (2020). End-to-end DNN based text-independent speaker recognition for long and short utterances. *Computer Speech & Language*, **59**, 22–35. 5, 46, 47

Rosenberg, S. & Abbeduto, L. (1987). Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults 1. *Applied Psycholinguistics*, **8**, 19–32. 6

Ross, E.D., Edmondson, J.A., Seibert, G.B. & Homan, R.W. (1988). Acoustic analysis of affective prosody during right-sided wada test: A within-subjects verification of the right hemisphere's role in language. *Brain and Language*, **33**, 128–145. 23

Ross, G.W., Cummings, J.L. & Benson, D.F. (1990). Speech and language alterations in dementia syndromes: Characteristics and treatment. *Aphasiology*, **4**, 339–352. 3

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*. 156

Rumelhart, D.E., Hinton, G.E., Williams, R.J. *et al.* (1988). Learning representations by back-propagating errors. *Cognitive modeling*, **5**, 1. 43

Sadjadi, S.O., Ganapathy, S. & Pelecanos, J.W. (2016). Speaker age estimation on conversational telephone speech using senone posterior based i-vectors. In *Proc. ICASSP 2016*, 5040–5044, IEEE. 156

Sainath, T.N., Vinyals, O., Senior, A. & Sak, H. (2015a). Convolutional, long short-term memory, fully connected deep neural networks. In *Proc. ICASSP 2015*, IEEE. 102, 173

Sainath, T.N., Weiss, R.J., Senior, A., Wilson, K.W. & Vinyals, O. (2015b). Learning the speech front-end with raw waveform CLDNNs. In *Sixteenth Annual Conference of the International Speech Communication Association*. 102

Sandoval, S., Berisha, V., Utianski, R.L., Liss, J.M. & Spanias, A. (2013). Automatic assessment of vowel space area. *The Journal of the Acoustical Society of America*, **134**, EL477–EL483. 22

Sarawgi, U., Zulfikar, W., Soliman, N. & Maes, P. (2020). Multimodal inductive transfer learning for detection of Alzheimer's dementia and its severity. *arXiv preprint arXiv:2009.00700*. 57, 58, 59

Sarkar, A.K., Matrouf, D., Bousquet, P.M. & Bonastre, J.F. (2012). Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification. In *Thirteenth Annual Conference of the International Speech Communication Association*. 139

Satt, A., Sorin, A., Toledo-Ronen, O., Barkan, O., Kompatsiaris, I., Kokonozi, A. & Tsolaki, M. (2013). Evaluation of speech-based protocol for detection of early-stage dementia. In *Proc. INTERSPEECH 2013*, 1692–1696, ISCA. 106, 122

Satt, A., Hoory, R., König, A., Aalten, P. & Robert, P.H. (2014). Speech-based automatic and robust detection of very early dementia. In *Fifteenth Annual Conference of the International Speech Communication Association*. 106, 122

Schmidtke, K., Pohlmann, S. & Metternich, B. (2008). The syndrome of functional memory disorder: definition, etiology, and natural course. *The American Journal of Geriatric Psychiatry*, **16**, 981–988. 137

Schuller, B., Müller, R., Lang, M. & Rigoll, G. (2005). Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensemble. In *Proc. INTERSPEECH 2005*, ISCA. 47, 48

Schuller, B., Steidl, S. & Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. In *Proc. INTERSPEECH 2009*, ISCA. 148

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. & Narayanan, S.S. (2010). The INTERSPEECH 2010 paralinguistic challenge. In *Proc. INTERSPEECH 2010*, 2794–2797, ISCA. 41, 108

Schuster, M. & Paliwal, K.K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, **45**, 2673–2681. 43

Searle, T., Ibrahim, Z. & Dobson, R. (2020). Comparing natural language processing techniques for Alzheimer's dementia prediction in spontaneous speech. *arXiv preprint arXiv:2006.07358*. 39, 61, 96

Selkoe, D.J. & Hardy, J. (2016). The amyloid hypothesis of alzheimer's disease at 25 years. *EMBO molecular medicine*, **8**, 595–608. 26

Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S. & Zhang, C. (2018). Disan: Directional self-attention network for RNN/CNN-free language understanding. In *Thirty-Second AAAI Conference on Artificial Intelligence*. 44

Sherod, M.G., Griffith, H.R., Copeland, J., Belue, K., Krzywanski, S., Zamrini, E.Y., Harrell, L.E., Clark, D.G., Brockington, J.C., Powers, R.E. *et al.* (2009). Neurocognitive predictors of financial capacity across the dementia spectrum: Normal aging, mild cognitive impairment, and Alzheimer's disease. *Journal of the International Neuropsychological Society*, **15**, 258–267. 30

Shimomura, T., Mori, E., Yamashita, H., Imamura, T., Hirono, N., Hashimoto, M., Tanimukai, S., Kazui, H. & Hanihara, T. (1998). Cognitive loss in dementia with lewy bodies and Alzheimer disease. *Archives of Neurology*, **55**, 1547–1552. 29

Shivakumar, P.G. & Georgiou, P. (2020). Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. *Computer speech & language*, **63**, 101077. 180

Singh, S., Bucks, R.S. & Cuerden, J.M. (2001). Evaluation of an objective technique for analysing temporal variables in dat spontaneous speech. *Aphasiology*, **15**, 571–583. 124

Skodda, S. & Schlegel, U. (2008). Speech rate and rhythm in Parkinson's disease. *Movement disorders: official journal of the Movement Disorder Society*, **23**, 985–992. 122

Smith, S.R., Chenery, H.J. & Murdoch, B.E. (1989). Semantic abilities in dementia of the Alzheimer type. ii. grammatical semantics. *Brain and Language*, **36**, 533–542. 3, 23, 24

Snowdon, D.A. (2003). Healthy aging and dementia: findings from the nun study. *Annals of internal medicine*, **139**, 450–454. 3

Snyder, D., Garcia-Romero, D., Povey, D. & Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. In *Proc. INTERSPEECH 2017*, 999–1003, ISCA. 5, 46, 47, 64

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In *Proc. ICASSP 2018*, 5329–5333, IEEE. 5, 42, 46, 47, 61, 156, 164, 176

Spiegl, W., Stemmer, G., Lasarcyk, E., Kolhatkar, V., Cassidy, A., Potard, B., Shum, S., Song, Y.C., Xu, P., Beyerlein, P. *et al.* (2009). Analyzing features for automatic age estimation on cross-sectional data. In *Tenth Annual Conference of the International Speech Communication Association*, Brighton, United Kingdom. 156

Spitzer, R.L., Kroenke, K., Williams, J.B. & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, **166**, 1092–1097. 29, 30

Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G. & Schuller, B. (2011). Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *Proc. ICASSP 2011*, 5688–5691, IEEE. 5, 47, 48

Sukhbaatar, S., Weston, J., Fergus, R. *et al.* (2015). End-to-end memory networks. In *Advances in neural information processing systems*, 2440–2448. 80

Sundberg, J. & Sataloff, R. (2005). *Vocal tract resonance*. Plural Publishing San Diego, California. 22

Sutskever, I., Vinyals, O. & Le, Q.V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112. 44, 45

Syed, M.S.S., Syed, Z.S., Lech, M. & Pirogova, E. (2020). Automated screening for Alzheimer's dementia through spontaneous speech. In *Proc. INTERSPEECH 2020*, ISCA. 39, 61, 96, 169

Syed, Z.S., Syed, M.S.S., Lech, M. & Pirogova, E. (2021). Tackling the ADReSSo challenge 2021: The MUET-RMIT system for Alzheimer's dementia recognition from spontaneous speech. In *Proc. INTERSPEECH 2021*, 3815–3819, ISCA. 63, 64, 96, 180

Tieleman, T. & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, **4**, 26–31. 107, 145

Tilk, O. & Alumäe, T. (2016). Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Proc. INTERSPEECH 2016*, 3047–3051, ISCA. 83, 96, 130

Toledo, C.M., Aluísio, S.M., Dos Santos, L.B., Brucki, S.M.D., Trés, E.S., de Oliveira, M.O. & Mansur, L.L. (2018). Analysis of macrolinguistic aspects of narratives from individuals with Alzheimer's disease, mild cognitive impairment, and no cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, **10**, 31–40. 5, 37

Tomashenko, N., Caubrière, A. & Estève, Y. (2019). Investigating adaptation and transfer learning for end-to-end spoken language understanding from speech. In *Proc. INTERSPEECH 2019*, 824–828, ISCA. 180

Tomoeda, C.K. & Bayles, K.A. (1993). Longitudinal effects of Alzheimer disease on discourse production. *Alzheimer Disease and Associated Disorders*. 23

Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatlóczki, G., Biró, E., Zsura, F., Pákáski, M. & Kálmán, J. (2015). Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In *Sixteenth Annual Conference of the International Speech Communication Association*. 122

Triapthi, A., Chakraborty, R. & Kopparapu, S.K. (2021). Dementia classification using acoustic descriptors derived from subsampled signals. In *European Signal Processing Conference (EUSIPCO)*, 91–95, IEEE. 57, 58

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B. & Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proc. ICASSP 2016*, 5200–5204, IEEE. 5, 47, 48, 101

Tseng, B.H., Shen, S.S., Lee, H.Y. & Lee, L.S. (2016). Towards machine comprehension of spoken content: Initial TOEFL listening comprehension test by machine. *arXiv preprint arXiv:1608.06378*. 76

Turner, G.S., Tjaden, K. & Weismer, G. (1995). The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, **38**, 1001–1013. 126

Tzirakis, P., Zhang, J. & Schuller, B.W. (2018). End-to-end speech emotion recognition using deep neural networks. In *Proc. ICASSP 2018*, 5089–5093, IEEE. 101

Ujiro, T., Tanaka, H., Adachi, H., Kazui, H., Ikeda, M., Kudo, T. & Nakamura, S. (2018). Detection of dementia from responses to atypical questions asked by embodied conversational agents. In *Proc. INTERSPEECH 2018*, 1691–1695, ISCA. 5

Variani, E., Lei, X., McDermott, E., Moreno, I.L. & Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *Proc. ICASSP 2014*, 4052–4056, IEEE. 42, 46, 47

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*. 39, 45

Videnovic, A., Bernard, B., Fan, W., Jaglin, J., Leurgans, S. & Shannon, K.M. (2010). The montreal cognitive assessment as a screening tool for cognitive dysfunction in huntington's disease. *Movement disorders*, **25**, 401–404. 28

Vieira, R.T., Caixeta, L., Machado, S., Silva, A.C., Nardi, A.E., Arias-Carrión, O. & Carta, M.G. (2013). Epidemiology of early-onset dementia: a review of the literature. *Clinical practice and epidemiology in mental health: CP & EMH*, **9**, 88. 18

Vincze, V., Gosztolya, G., Tóth, L., Hoffmann, I., Szatlóczki, G., Bánréti, Z., Pákáski, M. & Kálmán, J. (2016). Detecting mild cognitive impairment by exploiting linguistic information from transcripts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 181–187. 5, 38

Vuorinen, E., Laine, M. & Rinne, J. (2000). Common pattern of language impairment in vascular dementia and in Alzheimer disease. *Alzheimer Disease & Associated Disorders*, **14**, 81–86. 21

Walker, T., Christensen, H., Mirheidari, B., Swainston, T., Rutten, C., Mayer, I., Blackburn, D. & Reuber, M. (2020). Developing an intelligent virtual agent to stratify people with cognitive complaints: A comparison of human–patient and intelligent virtual agent–patient interaction. *Dementia*, **19**, 1173–1188. 66

Wang, D. & Zheng, T.F. (2015). Transfer learning for speech and language processing. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 1225–1237, IEEE. 180

Wang, N., Chen, M. & Subbalakshmi, K.P. (2020). Explainable cnn-attention networks (c-attention network) for automated detection of Alzheimer's disease. *arXiv preprint arXiv:2006.14135*. 57

Wang, N., Cao, Y., Hao, S., Shao, Z. & Subbalakshmi, K. (2021). Modular multi-modal attention network for Alzheimer's disease detection using patient audio and language data. In *Proc. INTERSPEECH 2021*, 3835–3839, ISCA. 64, 118

Warnita, T., Inoue, N. & Shinoda, K. (2018). Detecting Alzheimer's disease using gated convolutional neural network from audio data. *arXiv preprint arXiv:1803.11344*. 3, 40, 44, 75, 105, 108, 122, 124, 127

Wechsler, D. (1945). Wechsler memory scale. 29

Wechsler, D. (2008). Wechsler adult intelligence scale–fourth edition (WAIS–IV). *San Antonio, TX: NCS Pearson*, **22**, 498. 29

Weiner, J. & Schultz, T. (2018). Selecting features for automatic screening for dementia based on speech. In *International Conference on Speech and Computer*, 747–756, Springer. 61, 101

Weiner, J., Engelbart, M. & Schultz, T. (2017). Manual and automatic transcriptions in dementia detection from speech. In *Proc. INTERSPEECH 2017*, 3117–3121, ISCA. 121, 122

Weiner, J., Angrick, M., Umesh, S. & Schultz, T. (2018). Investigating the effect of audio duration on dementia detection using acoustic features. In *Proc. INTERSPEECH 2018*, 2324–2328, ISCA. 3

Wild, B., Eckl, A., Herzog, W., Niehoff, D., Lechner, S., Maatouk, I., Schellberg, D., Brenner, H., Müller, H. & Löwe, B. (2014). Assessing generalized anxiety disorder in elderly people using the gad-7 and gad-2 scales: results of a validation study. *The American journal of geriatric psychiatry*, **22**, 1029–1038. 30

Wilhelms-Tricarico, R. (1995). Physiological modeling of speech production: Methods for modeling soft-tissue articulators. *The Journal of the Acoustical Society of America*, **97**, 3085–3098. 21

Wold, S., Esbensen, K. & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, **2**, 37–52. 140

Wyss-Coray, T. & Rogers, J. (2012). Inflammation in Alzheimer disease—a brief review of the basic science and clinical literature. *Cold Spring Harbor perspectives in medicine*, **2**, a006346. 17

Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C. & Schuller, B. (2019). Speech emotion classification using attention-based LSTM. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**, 1675–1685. 5, 48, 49

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057. 45, 46

Yancheva, M. & Rudzicz, F. (2016). Vector-space topic models for detecting Alzheimer's disease. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2337–2346. 5, 39, 57, 58, 59, 133

Yancheva, M., Fraser, K.C. & Rudzicz, F. (2015). Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 134–139. 7, 57, 58, 101, 180

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489. 45, 76, 80, 85

Ye, C., Liu, J., Chen, C., Song, M. & Bu, J. (2008). Speech emotion classification on a riemannian manifold. In *Pacific-Rim Conference on Multimedia*, 61–69, Springer. 47, 48

Yngve, V.H. (1960). A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, **104**, 444–466. 38

Yu, D., Wang, S., Li, J. & Deng, L. (2010). Word confidence calibration using a maximum entropy model with constraints on confidence and word distributions. In *Proc. ICASSP 2010*, 4446–4449, IEEE. 121

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z. & Church, K. (2020). Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease. In *Proc. INTERSPEECH 2020*, 2162–2166, ISCA. 39, 61, 88, 96, 122

Yumoto, E., Gould, W.J. & Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *The journal of the Acoustical Society of America*, **71**, 1544–1550. 22

Zadikoff, C., Fox, S.H., Tang-Wai, D.F., Thomsen, T., De Bie, R.M., Wadia, P., Miyasaki, J., Duff-Canning, S., Lang, A.E. & Marras, C. (2008). A comparison of the mini mental state exam to the montreal cognitive assessment in identifying cognitive deficits in parkinson's disease. *Movement disorders*, **23**, 297–299. 28

Zargarbashi, S. & Babaali, B. (2019). A multi-modal feature embedding approach to diagnose Alzheimer's disease from spoken language. *arXiv preprint arXiv:1910.00330*. 61

Zhang, S. (2008). Emotion recognition in chinese natural speech by combining prosody and voice quality features. In *International Symposium on Neural Networks*, 457–464, Springer. 46, 48

Zhang, W., Li, R., Zeng, T., Sun, Q., Kumar, S., Ye, J. & Ji, S. (2016). Deep model based transfer and multi-task learning for biological image analysis. *IEEE transactions on Big Data*, **6**, 322–333. 180

Zhang, Y., Du, J., Wang, Z., Zhang, J. & Tu, Y. (2018). Attention based fully convolutional network for speech emotion recognition. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 1771–1775, IEEE. 48, 49

Zhu, Y., Obyat, A., Liang, X., Batsis, J.A. & Roth, R.M. (2021). WavBERT: Exploiting semantic and non-semantic speech using wav2vec and BERT for dementia detection. In *Proc. INTERSPEECH 2021*, 3790–3794, ISCA. 63, 64, 96, 118, 179, 180

Appendix A

Traditional Features used in Chapter 8

In Section 8.3.2, 20-dimension features inspired by conversation analysis and 7-dimension linguistic features are used as the extra input of the twin-CCLA system for extracting the TR-2 feature. The 20-dimension conversation analysis features are shown in Table A.1. These features were proposed based on the IVA_{60} dataset. Section 3.2.2 in Mirheidari [2018] provides a more detailed description of these features. The audio recordings from the IVA dataset include the speech from a neurologist, a participant and accompanying person(s). The features are designed according to the speech from different roles. Specifically, "APs" is the abbreviation for accompanying person(s), "Pat" is the abbreviation for the participant, and "AV" is the abbreviation for average. For example, APsNoOfTurns means "the number of turns from the accompanying person(s)".

The 7-dimension linguistic features are achieved by feature dimension reduction on the 600-dimension word vectors. As proposed in Mirheidari [2018], the pre-trained 300-dimension Glove word embedding is used for generating the 300-dimension word vector for each word in the transcripts. Then, to get the representation of each transcript, the average and variance of the word vectors belonging to a transcript are calculated and concatenated into a 600-dimension word vector. Finally, PCA is used for reducing the 600-dimension feature into a 7-dimension feature.

Table A.1: *The traditional features used as the extra input of the designed twin-CCLA system for extracting TR-2 feature.*

Type	**Features**
Acoustic(8)	APsNoOfTurns PatNoOfTurns NeuNoOfTurns APsAVTurnLength PatAVTurnLength PatFailureExampleAVPauses NeuAVTurnLength PatAVPauses
Lexical(4)	PatAVUniqueWords NeuAVUniqueWords APsAVUniqueWords PatAVAllWords
Semantic(8)	PatMeForWhoConcerns PatFailureExampleEmptyWords PatFailureExampleAllTime PatDontKnowForExpectation PatAVFillers PatAVEmptyWords AVNoOfRepeatedQuestions AVNoOfTopicsChanged