Detecting early signs of dementia in conversation

Bahman Mirheidari
Supervisor: Dr. Heidi Christensen

Department of Computer Science

University of Sheffield

This dissertation is submitted for the degree of

Doctor of Philosophy

March 2019

to my best friend and half of me, my lovely wife Zahra.

Declaration

I hereby declare that this dissertation is of my own work, except where I specifically referred to the works done by the other authors in the text. The contents of the study are original and have not been submitted for any other awards, qualifications, or degrees in universities. Parts of the findings of this study have already been published as the journal or the conference papers.

Bahman Mirheidari
March 2019

Acknowledgements

First and foremost, enormous thanks are due to my supervisor Dr. Heidi Christensen. Without her endless support, encouragement and guidance this study would not have been completed. I have learned a lot from her and I consider myself extremely lucky to have worked this PhD under her supervision. I am really grateful to my PhD panel members: Prof. Phil Green, Dr. Mark Stevenson and Dr. Stuart Cunningham. Their useful feedback and suggestions have been always constructive and invaluable. I would like to thank the Department of Computer Science (especially the head of department Prof. Guy Brown and the director of admissions Prof. Jon Barker) for providing the financial support throughout the study and access to the grid and the department servers. I would like to thank my thesis examiners Prof. Tanja Schultz and Prof. Roger Moore for their very helpful suggestions and corrections to the thesis.

Special thanks to Prof. Markus Reuber, Dr. Daniel Blackburn, Dr. Traci Walker, Dr. Kirsty Harkness and Prof. Annalena Venneri for their collaborations in publishing the papers and providing the medical dataset to work with. I also thank the MSc students and colleagues from the department of Neurology for collecting data (Casey Rutten, Imke Mayer, Thomas Swainston, Ronan O'Malley and Eric Brook) and all the participants who have been involved in the Intelligent Virtual Agent (IVA) study.

I wish to thank the previous colleagues from the Speech and Hearing (SpandH) group for their friendship, guidance and help especially in my first year of study (Dr. Ricardo Marxer, Dr. Mortaza Doulaty, Dr. Saeid Mokaram , Dr. Erfan Loweimi, Dr. Iigo Casanueva, Dr. Rosanna Milner, Dr. Yulan Liu, Dr. Sara Al-Shareef, Dr. Amy Beeston, Dr. Maryam Al Dabel, Dr. Oscar Saz Torralba) as well the current colleagues (Mauro Nicolao, Sadeen Alharbi, Mashael AlSaleh, Rabab Algadhy, Lubna Alhinti, Bader Matar Alotaibi, Gerardo Roa Dabike, Harry Jackson, Asif Jalal, Jisi Zhang, Feifei Xiong, Dr. Ning Ma, Dr. Madina Hasan, Dr. Salil Deena, Yilin Pan). My appreciation goes to Ronan O'Malley for kindly helping me in editing the medical background of the thesis.

Last but not least, I would like to thank my parents for their support and love throughout my life, especially my mother who passed away right in the middle of my PhD and I miss her a lot. I particularly would like to thank

my wife and my son, for their support, encouragement, tolerance and help in completion of this study.

Abstract

Dementia can affect a person's speech, language and conversational interaction capabilities. The early diagnosis of dementia is of great clinical importance. Recent studies using the qualitative methodology of Conversation Analysis (CA) demonstrated that communication problems may be picked up during conversations between patients and neurologists and that this can be used to differentiate between patients with Neuro-degenerative Disorders (ND) and those with non-progressive Functional Memory Disorder (FMD). However, conducting manual CA is expensive and difficult to scale up for routine clinical use.

This study introduces an automatic approach for processing such conversations which can help in identifying the early signs of dementia and distinguishing them from the other clinical categories (FMD, Mild Cognitive Impairment (MCI), and Healthy Control (HC)). The dementia detection system starts with a speaker diarisation module to segment an input audio file (determining who talks when). Then the segmented files are passed to an automatic speech recogniser (ASR) to transcribe the utterances of each speaker. Next, the feature extraction unit extracts a number of features (CA-inspired, acoustic, lexical and word vector) from the transcripts and audio files. Finally, a classifier is trained by the features to determine the clinical category of the input conversation.

Moreover, we investigate replacing the role of a neurologist in the conversation with an Intelligent Virtual Agent (IVA) (asking similar questions). We show that despite differences between the IVA-led and the neurologist-led conversations, the results achieved by the IVA are as good as those gained by the neurologists. Furthermore, the IVA can be used for administering more standard cognitive tests, like the verbal fluency tests and produce automatic scores, which then can boost the performance of the classifier.

The final blind evaluation of the system shows that the classifier can identify early signs of dementia with an acceptable level of accuracy and robustness (considering both sensitivity and specificity).

List of Acronyms and Abbreviations

AACD Aging-Associated Cognitive Decline

ACE Addenbrooke Cognitive Examination

ACE-III Addenbrooke Cognitive Examination-III

ACE-R Addenbrooke Cognitive Examination-Revised

AdaBoost Adaptive Boost

AD Alzheimer's Disease

AHC Agglomerative Hierarchical Clustering

AM Acoustic Model

AMI Augmented Multi-party Interaction

ANN Artificial Neural Network

AoA Age of Acquisition

AP Accompanying Person

ASAT Automatic Speech Attribute Transcription

ASR Automatic Speech Recognition

AUC Area Under Curve

BIC Bayesian Information Criterion

BLSTM Bidirectional Long Short-Term Memory

BMVS Bing Mobile Voice Search

BNT Boston Naming Test

BoW Bag of Words

bvFTD behavioural variant Fronto-Temporal Dementia

CA Conversation Analysis

CANDy Conversational Assessment of Neurocognitive dysfunction

CBoW Continuous Bag of Words

CDR Clinical Dementia Rating

CH Call Home

CHiME Computational Hearing in Multi-source Environments

CLR Cross Likelihood Ratio

CMS Cepstral Mean Subtraction

CMU Carnegie Mellon University

CNN Convolutional Neural Network

CPU Central Processing Unit

CRF Conditional Random Fields

CSF CerebroSpinal Fluid

CTC Connectionist Temporal Classification

CT Computed Tomography

CVN Cepstral Variance Normalisation

DBN Deep Belief Net

DCT Discrete Cosine Transform

DER Diarisation Error Rate

DiarTK Diarisation Toolkit

Diar Diarisation

DPE Diarisation Posterior Entropy

DLB Dementia with Lewy Bodies

DNN Deep Neural Network

Doc2vec Document to Vector

DoH Department of Health

DPD Depressive Pseudo-Dementia

DTI Diffusion Tensor Imaging

EEG Electroencephalogram

EM Expectation-Maximisation

FFT Fast Fourier Transform

FLD Frontal Lobe Dementia

FMD Functional Memory Disorder

fMLLR feature-space Maximum Likelihood Linear Regression

fMPE feature-space Minimum Phone Error

Fsh Fisher

FTD Fronto-Temporal Dementia

GAD-7 Generalised Anxiety Disorder assessment-7

GloVe Global Vector

GMM Gaussian Mixture Model

GP General Practitioner

GPU Graphics Processing Unit

GT Good Turing

Hal Hallamshire

HC Healthy Control

HMM Hidden Markov Model

IB Information Bottleneck

iBT internet-Based Test

ICR Information Change Rate

ICSI International Computer Science Institute

ILSE Interdisciplinary Longitudinal Study on adult development and aging

IMDB Internet Movie Database

IVA Intelligent Virtual Agent

JFA Joint Factor Analysis

KLD KullbackLeibler Divergence

KN Kneser-Ney

LDA Linear Discriminant Analysis

LIUM Laboratoire d'Informatique de l'Universit du Mans

LIWC Linguistic Inquiry and Word Count

LLR Log Likelihood Ratio

LM Language Model

LPC Linear Predictive Coding

LR Logistic Regression

LSA Latent Semantic Analysis

LSTM Long Short-Term Memory

LVCSR Large Vocabulary Continuous Speech Recognition

MAP Maximum A Posterior

MCE Minimising Classification Error

MCI Mild Cognitive Impairment

MEG Magnetoencephalogram

MFCC Mel Frequency Cepstral Coefficient

MGB Multi-Genre Broadcast

MLE Maximum Likelihood Criteria Estimation

Ml Machine Learning

ML Maximum Likelihood

MLP Multi Layer Perceptron

MMI Maximum Mutual Information

MMSE Mini Mental Status Examination

MOCA Montreal Cognitive Assessment

MPE Minimum Phone Error

MRI Magnetic Resonance Imaging

MRSI Magnetic Resonance Spectroscopic Imaging

MWE Maximum Word Error

NCLR Normalised Cross Likelihood Ratio

Neu Neurologist

nfPPA non-fluent Primary Progressive Aphasia

ND Neuro-degenerative Disorders

NHS National Health Services

NIST National Institute of Standard and Technology

NLP Natural Language Processing

NLTK Natural Language Toolkit

NoG No Gaps

NoV No Overlaps

NumSp Number of Speakers

Orc Oracle

OOV Out Of Vocabulary

Pat Patient

PCA Principle Component Analysis

PD Parkinson' Disease

PER Phone Error Rate

PET Positron Emission Tomography

PHQ-9 Patient Health Questionnaire-9

PLDA Probabilistic Linear Discriminant Analysis

PLP Perceptual Linear Prediction

POS Part Of Speech

PRI Perceptual Reasoning Index

PSI Processing Speed Index

QE Quantile Equalisation

RBF Radial Basis Function

RBM Restricted Boltzmann Machine

RFE Recursive Feature Elimination

RF Random Forest

RMT Rich Meeting Transcription

RMS Root Mean Squared

RNN Recurrent Neural Network

ROC Receiver Operating Characteristic

RT Rich Transcription

SAD Speech Activity Detection

SAT Speaker Adaptation Training

SWB Switchboard

SD Semantic Dementia

SCD Speaker Change-point Detection

SCd Subjective Cognitive Decline

SDS Spoken Dialogue System

Sez Seizure

SGD Stochastic Gradient Descent

SHoUT Speech Recognition Research at the University of Twente

SigRNN Sigmoid-unit-type Recurrent Neural Network

SI Speaker Independent

sMBR state-level Minimum Bayes Risk

SpandH Speech and Hearing

SPECT Single Photon Emission Computed Tomography

SPLICE Stereo Piece-wise Linear Compensation for Environment

SRE Speaker Recognition Evaluation

SVM Support Vector Machine

TCD Transcranial Doppler

TDNN Time Delay Neural Network

TF-IDF Term Frequency-Inverse Document Frequency

TIMIT Texas Instruments-Massachusetts Institute of Technology

TTS Text To Speech

TV Total Variability

UAR Unweighted Average Recall

UBM Universal Background Model

UK United Kingdom

VAD Voice Activity Detection

VB Variational Bayes

VCI Verbal Comprehension Index

VD Vascular Dementia

VTLN Vocal Tract Length Normalisation

W2vec Word to Vector

WAIS-IV Wechsler Adult Intelligence Scale-IV

WAIS Wechsler Adult Intelligence Scale

WDER Word Diarisation Error Rate

WER Word Error Rate

WFST Weighted Finite-State Transducer

WMI Working Memory Index

WMS Wechsler Memory Scale

WMS-IV Wechsler Memory Scale-IV

WSJ Wall Street Journal

WSJCam Wall Street Journal Cambridge

Contents

Contents xv

List of Figures xix

List of Tables xxiii

1 Introduction 1

1.1 Motivation . . . 3
1.2 Focus of study . . . 4
1.3 Thesis contributions . . . 6
1.4 Thesis structure . . . 9
1.5 List of publications . . . 10

2 Dementia 13

2.1 What is dementia? . . . 15
2.2 Stages of dementia and effect on communication . . . 16
2.3 Other causes of memory difficulties . . . 18
2.4 Current diagnosis processes . . . 20
2.4.1 Cognitive tools . . . 20
2.4.1.1 Minimal Mental Status Examination . . . 21
2.4.1.2 Montreal Cognitive Assessment . . . 21
2.4.1.3 Addenbrooke Cognitive Examination . . . 22
2.4.1.4 Boston Naming Test . . . 22
2.4.1.5 Wechsler Adult Intelligence Scale . . . 22
2.4.1.6 Wechsler Memory Scale . . . 23
2.4.1.7 Patient Health Questionnaire-9 . . . 23
2.4.1.8 Generalised Anxiety Assessment-7 . . . 23
2.4.2 Assessing conversational ability . . . 23
2.4.2.1 Conversation Analysis . . . 24
2.4.2.2 Analysis of conversation of people with dementia . . . 26

2.4.3 Advantages of an automatic screening tool . . . 29
2.5 Summary . . . 30

3 Automatic dementia detection using analysis of conversation 33
3.1 Literature review . . . 35
3.2 Dementia detection system . . . 40
3.2.1 Hallamshire data . . . 41
3.2.2 Extracted features . . . 42
3.3 Baseline results . . . 47
3.3.1 Feature selection . . . 49
3.3.2 Feature type importance . . . 54
3.4 Discussion . . . 55
3.5 Summary . . . 58

4 Automatic speech recognition 61
4.1 Spontaneous speech recognition . . . 64
4.1.1 Front-end processing . . . 65
4.1.2 Acoustic modelling . . . 66
4.1.3 Language modelling . . . 68
4.1.4 Search (decoding) . . . 69
4.1.5 LVCSR challenges . . . 70
4.2 Deep neural networks . . . 72
4.2.1 RNN/LSTM and TDNN . . . 75
4.2.2 End-to-end ASR . . . 78
4.3 Semi-supervised learning . . . 80
4.4 State-of-the-art . . . 82
4.5 Automated transcription . . . 86
4.5.1 Baseline ASR . . . 87
4.5.2 Adding additional data . . . 88
4.5.3 Improving the Acoustic Model (AM) . . . 90
4.5.4 Improving the Language Model (LM) . . . 91
4.5.5 Word Error Rate (WER) per speaker group . . . 93
4.6 Discussion . . . 95
4.7 Summary . . . 98

5 Speaker diarisation 99
5.1 Diarisation . . . 101
5.1.1 Diarisation architecture . . . 102

5.1.1.1 Speech activity detection . . . 103
5.1.1.2 Speaker change detection or speaker segmentation . . . 103
5.1.1.3 Speaker clustering . . . 104
5.1.1.4 Re-segmentation . . . 104
5.1.2 Diarisation toolkits . . . 104
5.1.3 State-of-the-art . . . 105
5.2 Baseline diarisation for dementia detection . . . 110
5.2.1 Effect of overlapping speech and within-turn gaps . . . 111
5.2.2 Word diarisation error rate . . . 113
5.3 I-vector based diarisation . . . 114
5.4 Discussion . . . 115
5.5 Summary . . . 118

6 Feature extraction 119
6.1 Introduction . . . 121
6.1.1 Extended acoustic features . . . 121
6.1.2 Extended lexical features . . . 123
6.1.3 Word vector features . . . 125
6.2 Classification results . . . 129
6.2.1 Effect of different feature types . . . 129
6.2.2 Confusion matrix . . . 130
6.2.3 Feature selection . . . 131
6.2.4 The Receiver Operating Characteristic (ROC) curve . . . 134
6.3 Discussion . . . 134
6.4 Summary . . . 138

7 Intelligent Virtual Agent 139
7.1 Introduction . . . 141
7.2 Verbal fluency tests . . . 141
7.3 Using an IVA to elicit conversation . . . 142
7.3.1 IVA datasets . . . 144
7.4 Results . . . 147
7.4.1 Confusion matrix . . . 148
7.4.2 Feature selection . . . 149
7.4.3 The Receiver Operating Characteristic curve . . . 152
7.4.4 Comparing neurologist-led to IVA-led conversations . . . 152
7.5 Discussion . . . 154
7.6 Summary . . . 157

8 Final evaluation **159**
8.1 Introduction . . . 161
8.2 Final results . . . 162
8.2.1 Effect of adding the healthy control group . . . 162
8.2.2 Processing the verbal fluency tests . . . 164
8.2.3 Combining the conversations with the verbal fluency tests . . . 165
8.2.4 Combining all the IVA datasets . . . 167
8.2.5 Feature selection (Recursive Feature Elimination (RFE)) . . . 169
8.2.6 Feature selection (statistically significant) . . . 171
8.2.7 F1-measure . . . 173
8.3 Discussion . . . 175
8.4 Summary . . . 179

9 Conclusions and further work **181**
9.1 Conclusions . . . 182
9.1.1 Feasibility of developing an automatic system to identify dementia . 183
9.1.2 Techniques and methodologies required for the system . . . 184
9.1.3 Providing diagnostic information for neurologists . . . 185
9.1.4 Keeping track of dementia over time . . . 186
9.1.5 Final evaluation on a real clinical setting . . . 186
9.2 Future work . . . 187
9.2.1 Improving the components of the system . . . 187
9.2.2 Dealing with other challenges of conversations . . . 187
9.2.3 Extracting other types of features . . . 188
9.2.4 Investigating other cognitive tests . . . 188
9.2.5 Improving the IVA . . . 188
9.2.6 Longitudinal applications . . . 189
9.3 Concluding remarks . . . 189

References **191**

A Conversation Analysis Symbols **223**

B General guidelines of the Hallamshire study **225**

List of Figures

1.1 *Automatic dementia detection system.* 4

2.1 *A sample CA of a conversation (from Lerner [2004]) between two speakers, Dean and Nixon. The numbers in parentheses indicate gap in tenth of second, the arrows shows pitch change; degree signs indicates softness; brackets shows overlapping time; underscore displays stress, etc.* 25

3.1 *Automatic dementia detection system.* 40
3.2 *Comparison of accuracy rates using individual classifiers based on all features and the most significant (10) features.* 52
3.3 *Comparison of accuracy rates using individual classifiers based on all features and the 10 most statistically significant features.* 53
3.4 *Classification accuracy rates for different types of features acoustic, lexical, semantic and visual-conceptual, as well as all features.* 54

4.1 *General architecture of a conventional Large Vocabulary Continuous Speech Recognition (LVCSR) system.* 64
4.2 *Example of an Hidden Markov Model (HMM) with 6 states for observation sequence o_1 to o_6, output probabilities for each state $b_j(o_t)$, and transition probabilities a_{ij}'s [Young et al., 2006].* 67
4.3 *Multi-layer neural network with feed-forward way from the input x, through the hidden layers h^j, to the network output h^4 [Bengio, 2009].* 72
4.4 *A simple Recurrent Neural Network (RNN) with one input, one output and one recurrent hidden unit [Lipton et al., 2015].* 76
4.5 *Long Short-Term Memory (LSTM) with a recurrent projection and an optional non-recurrent projection layer [Sak et al., 2014].* 77
4.6 *Sub-sampling technique to improve computation cost of LSTMs. Sub-sampled connections (red lines), original LSTMs connections (blue and red lines) [Peddinti et al., 2015].* 79
4.7 *Block diagram of the automatic segmentation system.* 89

4.8 *WER per speaker for different patient groups: ND, FMD. Pat=patient, Neu=neurologist, Aps=accompanying person(s), All spks=all speakers.* . . . 95

5.1 *General diarisation system. (a) Alternative clustering schemes. (b) General speaker diarisation architecture [Miro et al., 2012].* 103

5.2 *Example of removing the within-turn gaps. After eliminating the gaps (here, three gaps), four shorter segments of speaker 1 are merged together to form a longer segment.* 112

6.1 *Confusion matrix for the classifier (Diarisation (Diar)+Automatic Speech Recognition (ASR)) using all 99 combined features.* 131

6.2 *The classifier accuracy for the three levels of transcription automation (manual transcript (hatched red), ASR (grey), Diar+ASR (hatched green)) using the top n features from the combined features.* $n = 1, ..., 10$. *The lowest number of features to achieve a 100% accuracy, for the three levels of transcription automation are marked by arrows.* 132

6.3 *ROC curve for the classifier of Diar+ASR using the 99 features (k-fold* ($k = 15$) *cross validation, accuracy of the classifier: 90%).* 135

6.4 *ROC curve for the classifier of Diar+ASR using the top 3 features (k-fold* ($k = 15$) *cross validation, accuracy of the classifier: 100%).* 135

7.1 *The IVA setup: a laptop was located on a table displaying the IVA to the participants, while a hue camera as well as the laptop's built-in web-cam was video recording the session. Distant microphones on the table were used for audio recording.* 143

7.2 *The IVA acting as a neurologist. The web page plays a question and the patient can listen again by pressing the 'repeat' button (or 'space bar' key) or pressing the 'next' button (or 'enter' key) to move to the next question.* . . 143

7.3 *Confusion matrix for the 3-way classifier (Auto transcript+segmentation).* 149

7.4 *The ROC curve for the FMD/ND/MCI classifier using the 72 features for Diar+ASR.* 151

7.5 *The ROC curve for the FMD/ND/MCI classifier using the top 5 features for Diar+ASR.* 151

7.6 (a) Distribution of the average turn length (in seconds). (b) Distribution of the average silence (in seconds). (c) Distribution of the average overall duration (in seconds). 153

8.1 *Accuracy and ROC-Area Under Curve (AUC) of the 11 Logistic Regression (LR) classifiers for the conversations of the IVA2017/18 datasets with error bars (red lines: chance levels).* 163
8.2 *Accuracy and ROC-AUC of the 11 LR classifiers using the verbal fluency tests' features of the IVA2017/18 datasets (red lines: chance levels).* 166
8.3 *Accuracy and ROC-AUC of the 11 LR classifiers using both the conversations and the verbal fluency tests of the IVA2017/18 datasets (red lines: chance levels).* 166
8.4 *Accuracy and ROC-AUC of the 11 LR classifiers for the IVA2016/17/18 datasets (78 features) (red lines: chance levels).* 168
8.5 *Confusion matrix for the four-way classifier (IVA2016/17/18 datasets).* . . 169
8.6 *Accuracy and ROC-AUC of the 11 LR classifiers for the IVA2016/17/18 datasets (the most significant features (22)) (red lines: chance levels).* . . . 170
8.7 *Confusion matrix for the four-way classifier (22 most significant features using the RFE approach (IVA2016/17/18 datasets)).* 171
8.8 *Confusion matrix for the four-way classifier (24 most statistically significant features (IVA2016/17/18 datasets)).* 174
8.9 *The top 50 words with the highest insertion errors.* 177
8.10 *The top 50 words with the highest deletion errors.* 178
8.11 *The top 50 words with the highest substitution errors.* 178

List of Tables

2.1 *Common language deficits among people with dementia [Tang-Wai and Graham, 2008].* . . . 18

2.2 *Summary of the key qualitative features extracted via CA by Elsey et al. [Elsey et al., 2015].* . . . 28

3.1 *Demographic information of the participants.* . . . 43

3.2 *Hallamshire data set information.* . . . 43

3.3 *Qualitative features from Table 2.2, and corresponding features extracted from the transcripts by automatic analysis of conversation. Prefixes: Patient (Pat)=patient, Neurologist (Neu)=neurologist, and Accompanying Person (AP)s=accompanying person(s).* . . . 44

3.4 *Types of extracted features: acoustic, lexical, semantic and visual-conceptual.* . . . 47

3.5 *Classification accuracy rate using all 22 extracted features (10-fold cross validation).* . . . 48

3.6 *P-values of significance test (five repeats of 2-fold cross validation) for the pairs of the six classifiers. P-value less than 0.05 indicates a significant difference.* . . . 48

3.7 *The most significant (top 10) features with the highest contributions for the classification between the ND and the FMD patients using the RFE on the train set.* . . . 51

3.8 *The 10 most statistically significant features using the normality tests (Shapiro-Wilk and D'Agostino) and then parametric (Student's t-test) for normal (norm.) features and non-parametric (Mann-Whitney U test) for non-normal (non-norm.) features. Feature marked with star were in the top 10 features in Table 3.7.* . . . 53

4.1 *Baseline speech recognition results: the average WER with the standard deviation in brackets.* . . . 88

4.2 *Improved speech recognition results: the average WER with the standard deviation in brackets.* . . . 90

4.3 *Evaluating metrics for the six LMs. For training the LMs we used the leave-one-out cross validation approach (i.e. for each model 30 LMs). Oracle (Orc): oracle, LM: LM, Fisher (Fsh): Fisher dataset, Switchboard (SWB): Switchboard dataset. PPL: average perplexities, #vocab: vocabulary size, #Out Of Vocabulary (OOV): average number of out of vocabulary words, #3-g: average number of 3-grams, 3-g cov.: average coverage of 3-grams, #4-g: average number of 4-grams 4-g cov.: average coverage of 4-grams.* . . . 93

4.4 *Speech recognition results for different LMs.* . . . 94

5.1 *Diarisation error (consisting of the missing speaker error: E_{MISS}, false alarm error: E_{FA}, and speaker error: E_{SPKR}) for the Hallamshire (Hal)30 data using the Speech Recognition Research at the University of Twente (SHoUT) and Laboratoire d'Informatique de l'Universit du Mans (LIUM) toolkits.* . . . 111

5.2 *Diarisation Error Rate (DER) for Hal30 data using the SHoUT toolkit after removal of overlapping segments (no-overlaps: No Overlaps (NoV)), within-turn gaps (no-gaps: No Gaps (NoG), and both (no-overlaps/no-gaps: NoV_NoG).* . . . 112

5.3 *Word diarisation error for the baseline diarisation systems.* . . . 113

5.4 *DER for the Hal30 data using the Kaldi diarisation trained by the switchboard (SWB) or seizure data mixed with half of the Hal data (held-out approach)(Seizure (Sez)Hal), with or without knowing the number of speakers (Number of Speakers (NumSp)).* . . . 115

5.5 *Word diarisation error for the i-vector based diarisation systems.* . . . 115

6.1 *20 CA-inspired features: acoustic, lexical, semantic. Note: two visual-conceptual features from Table 3.3 were removed.* . . . 122

6.2 *List of the extended acoustic features.* . . . 123

6.3 *List of the extended lexical features.* . . . 124

6.4 *Accuracy of the LR classifier to classify between ND and FMD patients using different feature types, the three levels of transcription automation and segmentation. Inside the brackets is the number of features. CA: CA-inspired features, E-LX: extended lexical features, E-AC: extended acoustic features, WV: word vector features, *: due to the errors caused by the diarisation systems, 3 out of 30 word vector based results were missing. Combined features: final column shows results for all feature combined together.* 129

6.5 *Top 10 features with the highest contributions in classification between ND and FMD patients for the three levels of transcription automation (Manual transcript, ASR, and Diar+ASR). WV_col1: The word vector features column 1. *: were seen in Table 3.7. AC: acoustic features, LX: lexical features, SM: semantic features, and WV: word vector features.* 133

7.1 *Conversational questions/verbal fluency tests asked of the participants in three summers: 2016, 2017, and 2018.* 145

7.2 *The number of participants in the IVA2016/2017/2018 datasets with the diagnostic classes. FMD: Functional Memory Disorder, ND: Neurodegenerative Disorder, MCI: Mild Cognitive Impairment, HC: Healthy Control.* 146

7.3 *Demographic information of the participants in the IVA2016 dataset.* . . . 146

7.4 *Demographic information of the participants in the IVA2017/2018 datasets.* 147

7.5 *Accuracy of the LR classifier to classify between different patient groups: FMD/ND/MCI, FMD/ND, FMD/MCI, and ND/MCI using the 72 combined features.* 147

7.6 *Accuracy of the LR classifier to classify between different patient groups: FMD/ND/MCI, FMD/ND, FMD/MCI, and ND/MCI using top features for the three systems. NT: Number of top features.* 149

7.7 *Top features the RFE for the manual transcript and the Auto transcript+segmentation. WV_col1: The word vector features column 1. AC: acoustic features, LX: lexical features, and WV: word vector features.* 150

7.8 *Accuracy of the FMD/ND classifier for the human-lead (HUM) and the IVA-led (IVA) conversations.* 152

8.1 *verbal fluency tests' features.* 165

8.2 *The 22 significant features using the RFE approach. Features in bold were also in Table 3.7 (top 10 features on Hal dataset using only CA-inspired features).* 172

8.3 *The 24 statistically significant features using the normality tests (Shapiro-Wilk and D'Agostino) and then parametric (Student's t-test) for normal (norm.) features and non-parametric (Mann-Whitney U test) for non-normal (non-norm.) features. Features in bold were also in Table 3.8 (top 10 features on Hal dataset using the statistic tests. Features with '*' were in Table 8.2.* 173

8.4 *Unweighted average precision, recall and F1-measure for the four-way classifiers. conv.: conversations, fl.tst.: fluency tests* 174

A.1 *Some common symbols of Conversation Analysis (CA) (Lerner [2004]).* . . 223

A.2 *CA symbols continue..* 224

Chapter 1

Introduction

Contents

1.1 Motivation 3

1.2 Focus of study 4

1.3 Thesis contributions 6

1.4 Thesis structure 9

1.5 List of publications 10

Dementia is an umbrella term covering a broad category of memory disorders, mostly observed amongst the elderly, even though it is not a natural consequence of the human ageing process. Dementia normally starts with subtle word finding difficulties and a decline in thinking or memorising ability, however, it aggravates over time and interferes with almost all aspects of daily functioning, and ultimately leads to loss of communication ability. People with severe dementia become passive and often unaware of the presence of others. They may totally forget about their basic living needs such as eating, drinking and taking rest. Thus they need 24-hour care and monitoring.

The number of people suffering from dementia in the United Kingdom (UK) has increased significantly in recent years and the economic impact of dementia on the society is huge. It is now one of the major concerns of the UK National Health Services (NHS). According to a recent update from the Department of Health (DoH) in January 2018 [Department of Health, 2018], there are around 850,000 people in the UK living with dementia, and this is estimated to rise to 1 million by 2025 and continue to reach 2 million by 2050. The costs of dementia for the society is over £26 billion a year, from which £11.6 billion is for unpaid care. Every 3 seconds a person develops dementia in the world and it was estimated that there were around 50 million people with dementia worldwide in 2017, which is going to be almost doubled each 20 years, soaring to 152 million by 2050 [Alzheimer's Disease International, 2018]. The main causes of dementia in the UK are Alzheimer's Disease (AD) (covering over 60% of all cases) and Vascular Dementia (VD) (20%) [Department of Health, 2018].

Normally, speech and language are influenced early on in dementia. Dementia can affect the prosodic features of speech such as pitch, intonation and loudness. People with dementia may also lose the ability to remember words. In the early stage of developing dementia they might find their own strategies to cope with the problem. For instance, they might refer to the names by words such as "thing" or "thingy", or try to find substitutions (e.g. "car" instead of "truck"). However, over time, the issue become worse. Therefore, loss of vocabulary, impoverished (or simplified) syntax/semantics, and overuse of semantically empty words are commonly found in the language of people with dementia [Appell et al., 1982; Bayles and Kaszniak, 1987; Hamilton, 1994; Tang-Wai and Graham, 2008].

1.1 Motivation

Generally, there is no cure for dementia, but, there are drugs and treatments (e.g. cognitive stimulation therapy and cognitive rehabilitation) which can help with reducing dementia symptoms. However, the treatments are most effective in the early stage of the disease before dementia has developed and irreversible brain damage has occurred. Therefore, early detection of signs of dementia is highly desirable.

Currently, there are no known or reliable bio-markers for dementia and the process of diagnosing this disorder is very complex, mostly due to overlapping symptoms with normal ageing and low accuracy of existing cognitive screening tools.

Current tests capable of identifying people at high risk of developing dementia are expensive and invasive: Positron Emission Tomography (PET) scans expose people to radiation, and amyloid analysis of the CerebroSpinal Fluid (CSF) involves a lumbar puncture. The currently available tests for stratifying (screening) people with cognitive complaints, based on pen-and-paper testing, lack sensitivity or specificity especially early in the disease process. Most non-invasive tests also suffer from learning effects which prohibit frequent re-testing.

Therefore, it is highly desirable to build a cheap and reliable non-invasive automatic screening tool to identify people at risk of developing dementia. People found at high risk of developing dementia can be quickly referred to specialist clinics for treatments, while at the same time, people at low risk of developing dementia can be reassured much quicker.

An automatic screening tool assessing a person's language and communication skill, can be easily used for re-testing people as their conditions change, and it ideally would be used without presence of an examiner in a person's own home.

Recent advances in speech technology and machine learning has opened up a wide range of applications including medical aids and helps in diagnosis. This project investigates a solution for automatic detection and a screening tool, based on analysis of a person's speech and language.

1.2 Focus of study

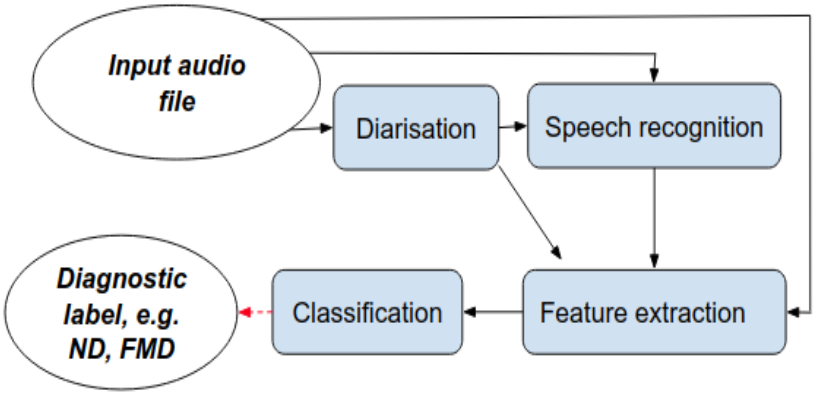


Figure 1.1: *Automatic dementia detection system.*

There have been a number of studies based on social interaction and communication ability of people with dementia using **CA**[1] [Elsey et al., 2015; Jones, 2015; Jones et al., 2015; Kindell et al., 2013; Oba et al., 2018; Perkins et al., 1998]. It is evident from these studies that patient interactions (communicating with others, e.g. doctors, caregivers and family members) can contribute to finding several features that could be used in identifying dementia. These approaches require audio and/or video recording of the conversations with the patient, and the recordings are subsequently transcribed before being qualitatively analysed by an expert (conversation analyst). The process is carried out manually, which is time-consuming and relatively expensive and not applicable for large-scale use. One alternative is to develop a system for doing the automatic analysis of the conversations[2] where dedicated speech technology is used to analyse the audio-recorded interactions.

Automatic analysis of conversation is an emerging and challenging area of research involving numerous disciplines to automate all the steps of the manual CA including Automatic Speech Recognition (ASR), speaker diarisation and classifier. Furthermore, in order to develop an automatic system to identify dementia, it is necessary to utilise

[1]It was originally introduced by Sacks et al. [1974] in sociology but then expanded to different fields of studies. Refer to Lerner [2004]; Sidnell and Stivers [2012] for more information.

[2]Please note that the phrase Conversation Analysis (CA) is used for a specific methodological approach. To avoid confusion, we will use 'analysis of conversation' to cover the general analysis of conversations which can be done through different approaches including CA, analysing acoustics, etc.

the latest techniques and tools developed for Natural Language Processing (NLP) and Machine Learning (Ml). Figure 1.1 shows the block diagram of an automatic dementia detection system, which consists of a diarisation toolkit, followed by a speech recognition, feature extraction and classification units [1].

Each of the technologies listed above have their own limitations and issues and in addition, natural human conversation is mostly unstructured and spontaneous with extra hidden complexities such as handling turn-taking, overlapping talk, repair, coping with disfluencies or hesitations and non-linguistic information (e.g. emotions).

This study is an attempt towards developing a screening tool based on analysing conversations that could be used in differentiating people with dementia from other disorders and also monitoring the signs of dementia in patients who have already developed the disorder.

Especially, this work will try to answer the following questions:

1. Is it feasible to develop an automatic tool to help doctors in detecting dementia?

2. What kind of speech, text and machine learning technologies and tools can be used for developing such a system?

3. How to generate more detailed diagnostic analysis of the conversations for the doctors?

4. How to collect data and keep track of the signs of dementia in the patients' speech and language over time?

The methodology that we will use to develop our system is based on Prototype Model[2] systems development. We will start with building an initial system to automatically extract features inspired by the qualitative interactions findings of Elsey et al. [2015] study (i.e. try to find the automatic equivalents of their features), and use those features to train classifiers to detect dementia. Therefore, the focus of the initial system will be on feature

[1] For more details refer to **Section 3.2**

[2] The Prototype Model approach develops an estimated sample of the final system, which is built earlier to give an idea of the functionality of the system in advance, and collect feedback from the clients.

extraction and classification tasks. This simplification would be possible by directly extracting features from the manual transcripts of the conversations between doctors and patients. Building the prototype model of the system helps us to find an answer to the first research question, proof-of-concept and feasibility of developing automatic dementia detection system. Later on, we will add more automation to the system, i.e. speaker diarisation and ASR (answering question 2).

We also investigate extracting different types of features and try to provide more diagnostic information for doctors (answering question 3).

Finally, we develop an Intelligent Virtual Agent (IVA) to act as a neurologist leading conversations with patients. This will ensure that our dementia detection system doesn't require speciality of an expert and the automatic tool can be easily used everywhere (e.g. in GP or patient's home). The IVA also let us to collect more data from patients during the study (answering question 4) and keeping the information over time to observe the changes.

1.3 Thesis contributions

1. An automatic pipeline for a dementia detection system based on conversations: Most work on automatic dementia detection is not based on analysis of conversation (between doctor and patient), but rather they implement systems for automatically processing existing cognitive tools such as describing a picture, naming animals. If there were interviews between patients and examiner, the speech segments relating to the examiner were manually removed from the audio recording. A number of studies have used the manual transcripts of interviews and a few authors have attempted to use ASRs for converting the input audio to transcriptions. It is very rare that studies have included a full, automatic pipeline consisting of speaker diarisation, ASR, feature extraction and classification (Weiner et al. [2018] worked in parallel using a similar pipeline to our system). The pipeline is introduced in **Chapter 3**, and the system components in **Chapter 4**, **5** and **6**). We first introduced the pipeline system in our two published papers [Mirheidari et al., 2017b, 2016] and gradually completed

the components of the system throughout the study.

2. Data collection using an IVA: The main dataset that we started our work with, was originally collected by a number of neurologists in the Hallamshire Hospital with the aim of doing manual conversation analysis of interactions between doctors and patients, i.e., not intended for automatic processing. Therefore, the recording are not similar to the standard recordings in convenient corpora for speech recognition. For instance, the microphones were not close to the patients and the accompanying persons, doors were opening/closing during the interviews, you could clearly hear the other background noises, and even noise made of the pen and paper used by the neurologists. Due to the high cost of replicating similar settings (especially recruiting neurologists) to collect more data, we developed an IVA to act as a neurologist and ask a number of questions from the participants. Using the IVA, a number of master students (from medical background) as part of their study collected data from a number of participants in summer 2016, 2017 and 2018 (IVA: **Chapter 7**). Some results from comparing doctor-patient interaction and the IVA-patient conversation are published in two Interspeech papers [Mirheidari et al., 2017a, 2018a].

3. Developing a cognitive test tool: Introducing the IVA also allowed us to not only elicit conversations but also administer more standard cognitive tests and investigate methods for automatically scoring them. We hypothesised that the automatic scores of the cognitive tests are useful in identifying dementia. Then, we showed that they can boost the performance of the automatic dementia detection system, when they are combined with the conversation-based features [Mirheidari et al., 2018b] (see **Chapter 8**).

4. Novel types of features for identifying dementia: During the study we investigated different types of features for identifying early signs of dementia. The types of features are conversation analysis-inspired features [Mirheidari et al., 2017b], acoustic features/lexical features [Mirheidari et al., 2017a] and word vector representation features [Mirheidari et al., 2018a]. Although a few of these features can be found in other studies (e.g. average length of pauses, average number of nouns), the collection of the

features we explored are unique (features: **Chapter 8**).

5. Evaluation in a clinical settings: To evaluate our tool and our automatic methods findings from this study we worked closely with a team of neurologists, neuroscientists and neuropsychologists which enabled us to test out system in the local memory clinic. All participants were recruited by clinicians in the Hallamshire Hospital and the gold standard for their diagnosis were based on their medical history, screening and cognitive tests and a team of experts and medical consultants (final evaluation: **Chapter 8**).

1.4 Thesis structure

The remainder of the thesis is organised as follows:

Chapter 2 reviews the literature regarding dementia and its impact on language and communication skill, as well as the current diagnosis processes for people with dementia.

Chapter 3 begins by reviewing other studies focused on developing an automatic dementia detection system and the challenges of developing such a system. Then we demonstrate our pipeline system consisting of a speaker diarisation tool to segment the audio stream and identify the speakers, followed by an ASR, and a feature extraction unit as well as a classifier trained to identify dementia conversations. Initially we focus on the feature extraction and classification tasks to build a baseline system inspired by human conversation analysis, and gradually, in the following chapters we automate other components of the pipeline.

Chapter 4 introduces the IVA component of the dementia detection system. The chapter briefly describes the challenges of spontaneous speech recognition, the architecture of an IVA system and outlines how neural networks have been applied to acoustic and language modelling of IVA. For our dementia detection system, we start with training a baseline IVA based on HMM-GMMs. Then we add an extra dataset to our dataset which enables us to train a state-of-the-art IVA, based on neural network models.

Chapter 5 describes the speaker diarisation module of the system which segments the input audio streams and identifies the speakers of the segments. We also select a final classifier of the dementia detection system.

Chapter 6 includes the feature extraction component of the system. We explore different types of features which are useful in identifying the early signs of dementia.

Chapter 7 introduces the IVA, who acts as a neurologist and asks questions from patients. We compare the IVA-led conversations with the neurologist-led conversations.

Chapter 8 includes the final evaluation of the system based on both the accuracy and the robustness of the final classifier. Finally, the chapter focuses on the automatic scoring of two cognitive tests collected by the IVA and how they can improve the performance of the overall system.

Chapter 9 contains the summary, conclusions and further work.

1.5 List of publications

1. Mirheidari, B., Blackburn, D., Reuber, M., Walker, T., and Christensen, H. (2016). Diagnosing people with dementia using automatic conversation analysis. In *Proceedings of Interspeech*, pages 1220–1224. ISCA [Mirheidari et al., 2016].

2. Mirheidari, B., Blackburn, D., Harkness, K., Venneri, A., Reuber, M., Walker, T., and Christensen, H. (2017a). An avatar-based system for identifying individuals likely to develop dementia. In *Proceedings of Interspeech*. ISCA [Mirheidari et al., 2017a].

3. Mirheidari, B., Blackburn, D., Harkness, K., Walker, T., Venneri, A., Reuber, M., and Christensen, H. (2017b). Toward the automation of diagnostic conversation analysis inpatients with memory complaints. *Journal of Alzheimer's Disease*, (Preprint):1–15 [Mirheidari et al., 2017b].

4. Blackburn, D., Mirheidari, B., Rutten, C., Mayer, I., Walker, T., Christensen, H., and Rueber, M. (2017a). Po029 an avatar aid in memory clinic [Blackburn et al., 2017a].

5. Blackburn, D., Reuber, M., Christensen, H., Mayer, I., Rutten, C., Venneri, A., and Mirheidari, B. (2017b). An avatar to screen for cognitive impairment. *Journal of the Neurological Sciences*, 381:319 [Blackburn et al., 2017b].

6. Mirheidari, B., Blackburn, D., Walker, T., Venneri, A., Reuber, M., and Christensen, (2018a). Detecting signs of dementia using word vector representations. *Proc. Interspeech 2018*, pages 1893–1897 [Mirheidari et al., 2018a].

7. Walker, T., Christensen, H., Mirheidari, B., Swainston, T., Rutten, C., Mayer, I., Blackburn, D., and Reuber, M. (2018). Developing an intelligent virtual agent to stratify people with cognitive complaints: A comparison of humanpatient and intelligent virtual agent–patient interaction. *Dementia*, page 1471301218795238 [Walker et al., 2018].

8. Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., and Christensen, H. (2019). Dementia detection using automatic analysis of conversations. *Computer Speech & Language*, 53:65–79 [Mirheidari et al., 2019].

9. Al-Hameed, S., Benaissa, M., Christensen, H., Mirheidari, B., Blackburn, D., and Reuber, M. (2018). Using acoustic measures to assess cognitive interactional capability in patients presenting with memory problems. *Submitted to PLOS ONE* [Al-Hameed et al., 2018].

10. Mirheidari, B., Daniel, B., OMalley, R., Walker, T., Venneri, A., Reuber, M., and Christensen, H. (2018b). Computational cognitive assessment: investigating the use of an intelligent virtual agent for the detection of early signs of dementia. *Submitted to ICASSP 2019* [Mirheidari et al., 2018b].

Chapter 2

Dementia

Contents

2.1 What is dementia? **15**

2.2 Stages of dementia and effect on communication **16**

2.3 Other causes of memory difficulties **18**

2.4 Current diagnosis processes **20**

2.4.1 Cognitive tools 20

2.4.2 Assessing conversational ability 23

2.4.3 Advantages of an automatic screening tool 29

2.5 Summary **30**

This study aims to identify early signs of dementia through analysis of conversation. It is, therefore, intuitive to begin the thesis with an introduction to dementia. Dementia is a state of sustained impairment of multiple cognitive domains leading to an impairment in function. In particular, we will focus on the effects of dementia on language and communication skill and how we can find clues for identifying dementia through analysis of people's conversation. This chapter is structured as follows:

Section 2.1 includes the definition of dementia and explores the common causes.

Section 2.2 describes the different stages of dementia in terms of severity and how this disorder affects language and communication skill of people.

Section 2.3 explores other important memory issues with similar symptoms to dementia. These include non-neuro-degenerative causes, which have the potential to improve and present challenges to the diagnosis of dementia.

Section 2.4 briefly introduces different scans and screening tests which can help in diagnosing dementia. Then we will list the benefits of developing an automatic tool to help doctors in diagnosing dementia.

Finally, **Section 2.5** summarises the key points of this chapter.

2.1 What is dementia?

Dementia is not a natural consequence of human ageing, but rather it is a syndrome[1] of progressive deterioration in intellect, personality and communicative functioning [Bayles and Kaszniak, 1987, ch: 1 p: 1] e.g. loss of reasoning and maintaining attention, orientation, and learning skills [Hamilton, 1994, ch: 1 p: 6]. Dementia is associated with multi-domain cognitive impairments (with over 50 causes [Bayles and Kaszniak, 1987, ch: 1 p: 1]) including at least memory impairment and one other deterioration such as aphasia (inability to produce speech or/and comprehend), apraxia (inability to perform a planned task), agnosia (inability to process sensory information), and executive dysfunction [Bayles and Kaszniak, 1987, ch: 1], [Quinn, 2014, ch: 1 p: 1]. Irreversible dementia, caused by brain diseases such as AD and Fronto-Temporal Dementia (FTD), are referred to as Neuro-degenerative Disorders (ND)s. These are characterised by the progressive loss of neuronal cells (up to 10% of the brain) [Bayles and Kaszniak, 1987, ch: 1 p: 8].

The most common cause of dementia is AD, which itself accounts for approximately 60% of all dementia cases [Alzheimer's Society, 2015a; Department of Health, 2018]. AD typically presents with an insidious progressive memory impairment gradually affecting executive and visuospatial functioning as well as the patient's behaviour. Other important causes of dementia are VD, Dementia with Lewy Bodies (DLB) and FTD [Quinn, 2014, ch: 1 p: 7]. In VD (the second most common cause of dementia [Alzheimer's Society, 2015d; Department of Health, 2018] accounting for 20% of dementia in the UK) vascular events (e.g. reduced blood supply) are responsible for cognitive decline in presence of strokes [Quinn, 2014, ch: 1 p: 8]. DLB shares clinical features with both AD and Parkinson' Disease (PD) including impairments of cognition and movement respectively [Alzheimer's Society, 2015b]. FTD or Frontal Lobe Dementia (FLD) [Alzheimer's Society, 2015c] is another important type of dementia characterised by progressive non-fluent speech (called non-fluent Primary Progressive Aphasia (nfPPA)), or loss of knowledge of object meaning (known as Semantic Dementia (SD)), or progressive changes in personality and behaviour (called behavioural variant Fronto-Temporal Dementia (bvFTD)) [Fernández-Matarrubia

[1] 'constellation of signs and symptoms associated with a morbid process' [Bayles and Kaszniak, 1987, ch: 1 p: 1]

et al., 2014]. FTD typically has a young onset (starts before age 65) causing behavioural features such as dis-inhibition, loss of sympathy or empathy and preservation (repeating same responses) or compulsions [Quinn, 2014, ch: 1 p: 10].

2.2 Stages of dementia and effect on communication

In terms of clinical severity, there are three stages of dementia: early, middle and late [Bayles and Kaszniak, 1987, ch: 1 p: 8] [Hamilton, 1994, ch: 1 p: 11],[Klimova et al., 2015]. In addition to this, there is a pre-clinical stage [Alberdi et al., 2016] characterised by pathological changes in the brain, the blood, and the CSF, but preceding the development of any symptoms. Detecting this stage is very hard and it may start up to 20 years before the symptoms appear.

In the early stage of dementia (known as mild dementia), the patient experiences subtle memory (short- and medium-term [Klimova et al., 2015]) and language issues like object naming difficulties. Initially this may be compensated for with strategies developed by the patient. There may also be signs of disorientation for time, but generally not for place or person. Discourse is somehow 'wordy', 'imprecise' and 'off-topic' and the patient finds it difficult to detect sarcasm or humour [Hamilton, 1994]. She/He may get trouble with everyday activities, especially complex activities or those in a professional context. Behavioural symptoms such as depression or anxiety may emerge as she/he becomes aware of the deterioration [Klimova et al., 2015]. These features can be subtle and may not be easily detectable. Specialist assessment is therefore valuable.

In the middle (moderate) stage, naming issues increase and conversation seems less meaningful, irrelevant and less interactive. The patient may be disoriented to time and place [Hamilton, 1994]. They are typically not able to adequately react to conversations and their attention, mathematics, reading and writing skills deteriorate significantly. In addition, they may develop psychotic symptoms such as delusion and hallucination [Klimova et al., 2015].

In the late (severe) stage, the disease is characterised by disorientation for time, place and person, and lack of communication. The patient may not be aware of the presence of

others, she/he produces limited discourse which is filled with repetition and nonsensical utterances [Hamilton, 1994, ch: 1 p: 11]. They also lose judgement, reasoning and social skills [Klimova et al., 2015].

At the moment there is no cure for dementia, however, there are a number of medicines and treatments to help with dementia symptoms [NHS, 2018]. For instance, 'Acetylcholinesterase inhibitors' (e.g. Donepezil and Rivastigmine) improve communication between nerves by increasing the availability of acetylcholine, an important neurotransmitter that is reduced in AD [Mehta et al., 2012], and 'Memantine' can block the effect of an excessive amount of a chemical known as 'glutamate' in the brain of people with moderate or severe AD [Reisberg et al., 2003]. Conditions such as stroke, depression and high blood pressure can affect symptoms of dementia, and in the late stages of dementia, behavioural and psychological symptoms (such as anxiety, aggression, delusions and hallucinations) can emerge. There are a number of complementary treatments to deal with these conditions, such as cognitive stimulation therapy, cognitive rehabilitation and reminiscence and life story work [NHS, 2018].

Communication is one of the most important and complex human behaviours. Speech, language and the ability to communicate are affected early in the natural history of dementia. Prosodic characteristics of speech of people with dementia (fundamental frequency, intonation, speech rate, etc.) could be affected. People might use more hesitation and pauses in their conversations as they struggle to find an appropriate word [Gonzalez-Moreira et al., 2015; Khodabakhsh et al., 2015]. They may communicate by non-verbal means like posture, facial expressions and eye contacts. However, their intentional, verbal communication ability declines. People communicate in order to share ideas. People with dementia gradually lose their abilities to produce meaningful communication and comprehend ideas [Bayles and Kaszniak, 1987, ch: 2] .

Dementia affects both the expressive and comprehension elements of language. Patients experience both cognitive and behavioural impairment affecting their communication abilities resulting in ineffective communication and inappropriate behaviours [Potkins et al., 2003].

The language degradation normally manifests through object naming difficulties or

Table 2.1: *Common language deficits among people with dementia [Tang-Wai and Graham, 2008].*

Speech and language difficulties	Example
Disfluency (disfluent speech hesitant/-faltering with abnormal prosody and reduced phrase length)	"My speech ... I can't tell the, I can't ... express it."
Word finding (empty speech/lack of meaning, reduced content words, pauses while searching for words, using generic substitutions like 'thing', and use of circumlocution)	Empty speech: "You can see out there and the things are out there.". Circumlocution: "Something that goes up in the air" (to indicate 'helicopter')
Simplifying grammar or grammar errors	"My wife, umm, teacher, umm, full time, umm, umm, children, umm, school."
Paraphasic errors (literal substitution for one sound, or semantic substitution)	Literal substitution: "tricycle" instead of "bicycle", semantic substitution: "car" instead of "truck"

loss of vocabulary, verbal disfluencies, simplified grammar, and overuse of words with empty meaning [Bayles and Kaszniak, 1987, ch: 3], [Hamilton, 1994, ch: 1]. Table 2.1 [Tang-Wai and Graham, 2008] shows some of the common language difficulties amongst people with dementia with a few explanatory examples. For instance, instead of helicopter they may use empty speech and say "something that goes up in the air", and simplify grammar like "My wife, umm, teacher ...", instead of "My wife is a teacher ...".

People with dementia gradually lose their sociolinguistic abilities, i.e. their language worsens in social situation. For instance, when they start a conversation, they may speak inappropriately, too loudly or repeat the same phrases [Klimova et al., 2015].

2.3 Other causes of memory difficulties

People referred to doctors with memory complaints may not have dementia. There are a number of memory disorders that share symptoms with dementia. In contrast to ND, these conditions are often reversible. The overlapping symptoms can cause confusion for specialists and makes the diagnosis of dementia challenging.

Although some specialists may consider the early stage of dementia the same as the 'Mild Cognitive Impairment (MCI)', there are important differences. MCI refers to not only patients with neuro-degenerative pathology but also to patients with depression-related cognitive issues or cognitive impairments due to alcohol/drug use or other co-morbidities [Blackburn et al., 2014]. MCI symptoms may stabilise or even improve over time. Thus, not all patients diagnosed as MCI develop AD (only between 10 and 15%). The reason some people progress to AD and others do not is unknown [Alberdi et al., 2016].

Subjective memory complaints are very common amongst the patients referred to memory clinics. Subjective Cognitive Decline (SCd) is an earlier 'clinical manifestation' of AD than MCI. However, self-awareness of memory difficulties often does not necessarily indicate the presence of dementia. There might be a number of other factors causing subjective memory issues such as psychological, environmental and pathological features [Blackburn et al., 2014].

Functional Memory Disorder (FMD) is one of the major causes of memory problems. FMD is a syndrome, a medical and psychological condition causing failure of memory and concentration in daily life. It is not related to organic factors and presumably caused by distress and psychological factors. In contrast to SCd patients, their problems are not subjective but rather are credible and real [Schmidtke et al., 2008]. Depression can also cause nonorganic cognitive disorder and it can be associated with FMD. FMD can cause and promote depression, but most FMD patients are not depressed.

Although depression can contribute to the presence of dementia, there is a group of patients who are depressed but not demented. Responding 'I don't know' to questions and confusing, for instance, salt with sugar, demonstrates lack of concentration/interest rather than dementia. Inability to perform constructive tasks is common among Depressive Pseudo-Dementia (DPD) patients. Elderly depressed patients are particularly liable to become confused, disoriented and incontinent. Also, lack of intellect, sadness, fatigue and insomnia are common in DPD patients [Kramer, 1982].

2.4 Current diagnosis processes

The process of diagnosing dementia is complicated and typically consists of a number of examinations and screenings including [Quinn, 2014, ch: 1 pp: 3-6] general examination (to identify co-morbid conditions[1] e.g. atrial fibrillation, congestive heart failure), cognitive evaluation (to test attention, orientation, memory, executive function, language, etc.) and neurological examination (such as cranial nerve testing and gait assessment).

Recently, many physiological signals have been evaluated by researchers to help in the diagnosis process, including CSF, blood samples, Computed Tomography (CT) scans, PET molecular scans, Single Photon Emission Computed Tomography (SPECT) (measuring brain activity), structural and functional Magnetic Resonance Imaging (MRI), Magnetic Resonance Spectroscopic Imaging (MRSI), Diffusion Tensor Imaging (DTI) (type of MRI scanning micro-structure of brain), Transcranial Doppler (TCD) ultrasonography, Electroencephalogram (EEG)/Magnetoencephalogram (MEG), and eye movements [Alberdi et al., 2016]. However, these investigations are costly and/or invasive.

In addition to the psychological changes, neurologists consider behavioural changes in patients such as sleeping/walking patterns, and speech and communication ability. There are a wide range of screening tools and cognitive batteries to assess the communication ability of patients.

The assessment of people with memory complaints normally begins with a history taken from the patient and any accompanying person(s) [Elsey et al., 2015] which may be followed by a number of screening tests. The examiner normally uses a pen and paper to take notes during the interview and scores different parts of the tests.

2.4.1 Cognitive tools

Each cognitive battery or tool is comprised of various assessments (e.g. verify orientation to time and place by asking about the current year and where they are), and completing tasks (e.g. retelling stories, describing pictures). Based on the patient's responses and the predefined criteria the examiner scores each task. The total score can then be

[1] presence of one or more disorders simultaneously

compared with the standard threshold(s) (cut-offs) to determine the level of dementia. It is important to note that a good screening tool should have both high 'sensitivity' and high 'specificity'. Sensitivity refers to the ability of a test to identify correctly those with a disease (true positives), while specificity refers to the test's ability to identify those without a disease (true negatives).

There is a wide range of neuropsychological assessment tools that can be used, depending on a patient's conditions. A number of the common tests and batteries are listed below. Each one of these tools has their own limitations and neurologists may use some of them and/or other tests, in addition to interviews with patients, carers or family members to help in diagnosing dementia.

2.4.1.1 Minimal Mental Status Examination

The most common tool for cognitive evaluation is the Mini Mental Status Examination (MMSE) [Folstein et al., 1975]. It is a 30-point questionnaire which assesses orientation to time and place, repeating/remembering words, calculation, naming objects, etc. A cognitively normal person may score over 25. A score less than 10 indicates a severe cognitive impairment, a score between 10 and 20 shows a moderate impairment, and a person scoring between 20 and 25 might be considered to have a mild cognitive issue. Despite its simplicity and popularity, the MMSE has limitations and typically requires additional comprehensive instruments [Quinn, 2014, ch: 1 p: 5].

2.4.1.2 Montreal Cognitive Assessment

The Montreal Cognitive Assessment (MOCA) is another widely used brief screening tool which is specially designed to detect MCI [Nasreddine et al., 2005]. Similar to the MMSE, it has 30 points, although its cutoff score for normal performance is 26. It assesses various skills including visuo-constructional skill (drawing cubes and clocks), naming, repeating, short-term memory, verbal fluency, attention and abstraction. Compared to the MMSE, the MOCA is more sensitive to MCI but with low specificity and the most appropriate cut-off point is not clearly agreed [Coen et al., 2016].

2.4.1.3 Addenbrooke Cognitive Examination

The Addenbrooke Cognitive Examination (ACE), first introduced by Mathuranath et al. [2000], has 100 points and focuses on five cognitive skills: attention/orientation, memory, language, verbal fluency, and visuospatial ability. It takes more time than the MMSE and requires a higher level of familiarity with cognitive disorders (i.e. the examiner of the MMSE can be a General Practitioner (GP), but the ACE normally carries out by a neurologist). Addenbrooke Cognitive Examination-Revised (ACE-R) [Mioshi et al., 2006] is a revised version of the ACE with more clear domain scores and two cut-offs at 88 and 82. A more recent version of the ACE is Addenbrooke Cognitive Examination-III (ACE-III) which adds some similar items from the MMSE to the ACE assessment [Hsieh et al., 2013].

2.4.1.4 Boston Naming Test

The Boston Naming Test (BNT) [Goodglass et al., 1983] is a common Confrontational Naming test designed to assess people with language difficulties such as aphasia and dementia. It consists of 60 pictures that are presented one by one to a patient, who are asked to say what the pictures represent. If the patient struggles to name an item, the examiner gives a hint or performs 'phonemic cuing', i.e. the first phoneme of the word. Despite its popularity, there are some studies questioning whether the BNT is adequately standardised and whether it captures all of the processes known to be involved in a successful naming [Harry and Crowe, 2014].

2.4.1.5 Wechsler Adult Intelligence Scale

The Wechsler Adult Intelligence Scale (WAIS) is the most common intelligence quotient (IQ) test, measuring intelligence and cognitive abilities in adults. The fourth version of the test (Wechsler Adult Intelligence Scale-IV (WAIS-IV)) released in 2008 comprises of 10 core subsets and five supplementary tests. There are four main index scores including Verbal Comprehension Index (VCI), Perceptual Reasoning Index (PRI), Working Memory Index (WMI), and Processing Speed Index (PSI) [Wechsler, 2014].

2.4.1.6 Wechsler Memory Scale

The Wechsler Memory Scale (WMS) [Wechsler, 1945] is designed to assess different memory functioning. The latest version (Wechsler Memory Scale-IV (WMS-IV))[Wechsler, 2014] contains seven sub-tests: logical memory, verbal paired associates, visual reproduction, brief cognitive status exam, designs, spatial addition and symbol span.

2.4.1.7 Patient Health Questionnaire-9

The Patient Health Questionnaire-9 (PHQ-9) is a short self-assessment questionnaire with 9 items designed to detect depression and its severity [Kroenke and Spitzer, 2002]. Each question can be scored between 0 and 3, making 27 score in total. People with scores up to 4 are normal. A score between 5 and 9 shows mild depression, 10 to 14 moderate, 15 to 19 moderately severe and over 20 severe depression. It assesses 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] reported a high sensitivity but a low specificity for PHQ-9.

2.4.1.8 Generalised Anxiety Assessment-7

The Generalised Anxiety Disorder assessment-7 (GAD-7) questionnaire has seven questions to assess generalised anxiety and its severity. Each question has scores between 0 and 3 (21 maximum score). A score of less than 5 shows mild anxiety, between 6 and 14 indicates a moderately severe anxiety, and above 15 severe anxiety [Spitzer et al., 2006].

2.4.2 Assessing conversational ability

As previously mentioned, assessing people with memory complaints starts with history taking. Neurologists may also need different cognitive tests and screening tools to help them in diagnosing dementia. Traditionally, most of the tests were designed based on assessing language solely, reflecting the methodological techniques used for the cognitive batteries, picture describing and similarity tests. For instance, according to the findings

from a study conducted by Blanken et al. [1987], analysing short interviews with dementia and healthy participants, the patients with dementia were able to produce almost the same amount of both simple and complex sentences as the healthy control group, however, the number of nouns produced by people with dementia was considerably lower than those used by the healthy seniors.

Bucks et al. [2000] also measured the linguistic features from the interviews with AD patients and a healthy group. They reported lower rates of noun production amongst the AD patients, but a higher usage of pronouns and verbs in contrast to healthy elderly, although the verbs produced by the AD group were poor in terms of comprehension and lexical richness. Recently, the validity of studies based only on the language produced by people with dementia has been questioned and there has been a significant increase in understanding the social interaction of people with dementia, which tends to show more pragmatic deficits such as inappropriate word selection, topic shift, taking turn, etc [Jones, 2015; Kindell et al., 2013].

2.4.2.1 Conversation Analysis

The CA was originally introduced in sociology [1] around 1967-1968, however, it has rapidly expanded to other disciplines including linguistics, communication, political science, anthropology and psychology [Sidnell and Stivers, 2012]. CA is a qualitative research method designed to investigate the structural organisation of everyday social interaction.

Conversations are built on structures known as adjacency pairs (such as question and answer, greeting and greeting, compliment and down player, request and grant [Jurafsky and Martin, 2008], they take place as a joint activity between two or more interlocutors who exchange discourses in a consecutive manner (turns). Turn-taking behaviour occurs based on a rule identified by Sacks et al. [1974] for the first time. At a transition relevance place, the current speaker might select the next speaker, if not, any other speakers may take the next turn. Occasionally the current speaker has to carry on when no one else takes the next turn. There are other important rules in a conversation such as topic management (carrying on with the current topic or initiating a new one) and repair (e.g. to change a

[1] by three pioneers: Emanuel Schegloff, Harvey Sacks and Gail Jefferson

message due to false starts, mishearing and misunderstanding) [Sidnell and Stivers, 2012].

The process of CA requires a number of steps including audio and/or video recording of a conversation, transcribing the encounters, and finally carrying out a qualitative analysis by a trained expert (conversation analyst). In addition to the words being exchanged in a conversation, the transcription should include extra information such as non-verbal behaviours (e.g. turning to someone, looking at windows, opening a door), pauses between words or speaker turns, the loudness of utterance and overlapping speech. The CA community has developed standard symbols to display such information (for more details refer to [Lerner, 2004]).

Figure 2.1 (from [Lerner, 2004]) shows a part of a conversation between two speakers Dean and Nixon. First Dean starts with the utterance "I don't know the full extent 'v it." and after 0.7 sec gap, continues with "uheh". After another gap of 0.9 sec, Nixon replies "I don'noo 'bout anything else except". Before finishing the end part of the word except, Dean says "I don't either in I w'd also", and the conversation carries on. The left square bracket indicates the overlap between two utterances, gaps with timing in tenth of second are inside parentheses, arrows indicate pitch goes up/down, degree signs shows softness, underscore shows stress, etc. (for more details refer to [Lerner, 2004], **Appendix A** summarises some of the common CA symbols)

Dean: I ↑don't kno:w thė (·) full extent ↓'v it.↓
(0.7)
Dean: °↓Uh:::eh°
(0.9)
Nixon: °I don'noo° 'bout anything else exc^{hh}e[pt
Dean: → [I don't either in I: °w'd (h)als(h)o
→ hhate tuh learn [some a'] these thi]ngs. ·hh·hh·hh·hh
Nixon: [W e l l] y a : h]
(0.2)
Dean: So ↑That's,hhhh that's that situation.

Figure 2.1: *A sample CA of a conversation (from Lerner [2004]) between two speakers, Dean and Nixon. The numbers in parentheses indicate gap in tenth of second, the arrows shows pitch change; degree signs indicates softness; brackets shows overlapping time; underscore displays stress, etc.*

2.4.2.2 Analysis of conversation of people with dementia

Analysis of conversation is a promising approach for understanding and analysing the communication ability and social interaction of people with dementia. Heidi Hamilton, as part of her PhD thesis, for around four and half years recorded conversations with a patient with AD to understand the patient's communicative ability at the discourse level (known as Conversation Discourse Analysis). The results of her work were published as a book [Hamilton, 1994]. She found that initially the patient had slight memory impairments but remained active and able to use strategies to mask her issues. For instance she used alternative words to handle the word-finding difficulties; she was somewhat confused, yet aware of her circumstances and the alternative words she was using were semantically close to the original words. However, gradually the patient's communication abilities deteriorated. In the second stage of disease (moderate dementia), her awareness decreased considerably, but she still managed to deal with her difficulties by using empty or unrelated alternative words. In the third stage (severe dementia), however, the level of her function was reduced markedly and she could barely initiate any conversations. Instead, she would produce excessive inappropriate responses or no response when asked questions. In the end, she entered a totally passive state, where, to communicate with others, the patient could only produce very limited utterances such as “uhhuh”, “mhm”,“mm hm”, “mmm” and “hmm”.

Perkins et al. [1998] worked on analysing the ability to produce different forms of discourse by people with dementia, including picture description, storytelling, procedural discourse and clinical interviews, and especially focused on the role of the conversational partner (caregivers) in interactions. They analysed three key aspects in the interactions: the turn taking, repair and topic management. They noticed that the positive attitude of the conversational partner has an important impact on the success of conversations.

Kindell et al. [2013] used CA approach to examine everyday communication of a family living with dementia: a participant with SD and his wife. They took a few video recordings of the family at home. Their findings revealed that the subject repeated the practice of enactment (speaking out as an actor who is performing an act in a scene and describing

events and using body movements point out somewhere and talking loudly, etc.) as a strategy to enable him to generate a higher level of meaningful communication rather than using a limited vocabulary alone.

Considering the limitations of the current neurological tests, Oba et al. [2018] investigated the feasibility of using a conversation-based test among the residents of care homes (Conversational Assessment of Neurocognitive dysfunction (CANDy)[1]) to assess cognitive functioning of patients. In a conversation between the caregiver and the patients, the test is run, evaluating frequent characteristics of conversation such as "repeatedly asking same question", "vague understanding of conversational partner", "lack of showing interest in conversation" and "not an expansion of conversation content". They found correlations between the results of the test and other standard cognitive tests like MMSE, while their approach was reported to be less invasive (causing less distress to patients) and less intrusive to the relationship between the patients and the examiners.

Elsey et al. [2015]; Jones et al. [2015] applied CA to doctor-patient interactions in a memory clinic. The study revealed several features that could be used to distinguish between patients suffering from FMD (people with memory complaints not due to neurodegenerative aetiology) and patients with ND. Interviews were conducted between neurologists and patients (15 ND and 15 FMD) and audio and video data was recorded. The patients were encouraged in advance to bring someone along to help them throughout the process (Accompanying Person (AP)). The interviews consisted of two parts: history-taking conversation and the formal assessment ACE-R (more details about the study will be covered in the next chapter).

They found several important qualitative features[2] which can differentiate between ND and FMD, such as the role of the AP (ND group could not talk most of the time and the APs answered the questions on their behalf) and the meaning of "I don't know" (for ND it means inability to recall not unsure), ND patients typically cannot recall the last time their memory let them down. Table 2.2 summarises the distinctive qualitative

[1] http://cocolomi.net/candy/en/

[2] Using the CA approach, they found a number of qualitative features which differ in the two patient groups. Since the term feature can cause confusion for the readers, we refer to these features as the "qualitative features".

Table 2.2: *Summary of the key qualitative features extracted via CA by Elsey et al. [Elsey et al., 2015].*

Qualitative feature	Findings in FMD group	Findings in ND group
F1) Accompanying persons	gave a second opinion to confirm the patient	main spokesperson
F2) Response to 'who's most concerned'	mostly patients	both patients and partners, or 'I don't know' answer
F3) Patient recall of recent memory failure	yes, with details	they had difficulties to answer, used filling words, or replied 'all the time'
F4) Inability to answer	marks unsure of response rather than inability to recall	marks inability to recall, nonverbal behaviours like head turning and long gaps
F6) Patients' elaborations and length of turns	answer all parts of the question	answered only a single part of the question
F6) Elaboration of answers	yes	no, despite having a second chance

features found using the CA by Elsey et al. [2015].

2.4.3 Advantages of an automatic screening tool

The process of diagnosing dementia is difficult due to overlapping symptoms between dementia and other memory and depression-related disorders (e.g. SCd, FMD, DPD), and normal ageing. Unfortunately, most of the tests capable of identifying people at high risk of developing dementia are expensive and invasive. They may expose people to risks such as radiation, lumbar puncture, distress, etc. Other standard screening tools are much cheaper and non-invasive. These are mostly based on pen-and-paper and lack sensitivity or specificity. There is also the learning effect which does not allow frequent, repeated use of the tests.

There is, therefore, an urgent medical need for a reliable, repeatable, non-invasive, easy to use, and low-cost tool for identifying people at risk of developing dementia. This would ensure quicker access to specialist assessment and treatment for those found to be at high risk. It would also allow for reassurance for those at low risk of developing dementia. An ideal tool would allow re-testing for those at intermediate risk or those whose performance fluctuates and would be usable without requiring assessor expertise, for example, in people's own homes.

2.5 Summary

Dementia disorders are associated with progressively deteriorating memory, human intelligence, personality and communication ability. ND is a group of irreversible dementias, which cause loss of the neuron cells in the brain. AD is the most common cause of dementia accounting for up to 60% of dementia cases. Other important types of dementia include Lewy body, Frontotemporal, and Vascular dementia. In terms of severity, dementia can be categorised generally into three stages: mild, moderate and severe.

Language and communication is affected early on in dementia and as the disease progresses people with dementia lose their ability to produce meaningful communication. Effects on language includes disfluencies, word finding difficulties and using empty speech or generic substitutions (e.g 'thing'), grammar simplification and literal/semantic substitutions. Dementia also affects acoustics and speech prosodic features, i.e. pitch, intonation, longer pauses in conversation, and it may cause people to use more hesitation words, filler words and pauses like "umm" and "er".

Diagnosing dementia is a challenging task and comprises of several examinations and tests, such as brain scans (MRI, and PET), blood tests, Gait assessment, and many cognitive screening tools like MMSE, MOCA, ACE-R. Although widely used, each of these tests and screening tools has its own limitations in both sensitivity (identifying true positives) and specificity (identifying true negatives). There is, therefore, a need to build a cheap, non-invasive and reliable automatic screening tool to identify people at risk of developing dementia. There are a number of studies on dementia focusing on the analysis of conversation, which reveals more communication ability of people when they interact with doctors, family members or caregivers. In the study carried out by Elsey et al. [2015], they found a number of qualitative features that can be used for distinguishing between people with ND and FMD disorders.

This thesis is about delivering an automatic detection tool for screening and monitoring dementia through analysis of conversation. Developing a low-cost non-invasive and reliable tool for identifying people with a high risk of developing dementia will enable quicker referral to a specialist. It can also bring reassurance for those found at low risk

of developing dementia.

Chapter 3

Automatic dementia detection using analysis of conversation

Contents

3.1 Literature review 35

3.2 Dementia detection system 40

3.2.1 Hallamshire data 41

3.2.2 Extracted features 42

3.3 Baseline results 47

3.3.1 Feature selection 49

3.3.2 Feature type importance 54

3.4 Discussion 55

3.5 Summary 58

This chapter is about exploring and finding an answer to the first research question, i.e. the feasibility of developing an automatic tool to detect dementia via analysis of conversation. Due to the complexity of such a system and influenced by the "prototyping model" software development, we start with a simplified version of the system (assuming other components are ready) to produce results. In the following chapters, we will add more automated components to the system. This chapter mainly focuses on providing a "proof-of-concept" for the development of an automatic dementia detection system. Inspired by Elsey et al. [2015] qualitative study, we extract automatic features from the same conversations to identify dementia. The chapter is organised as follows:

Section 3.1 is a literature review of the recent studies focussing on automatic detection of dementia using speech processing and/or analysis of conversation.

Section 3.2 introduces the pipeline of our system, which consists of a speaker diarisation unit, an ASR, a feature extraction component and a classifier.

Section 3.3 contains the details of a pilot study conducted as a proof-of-concept of the introduced system.

Section 3.4 and 3.5 are the discussion and the summary of the chapter, respectively.

3.1 Literature review

Whilst the automatic analysis of interaction is quite a new field of study [Moore, 2015; Shriberg, 2005], a significant amount of work has been carried out using machine learning techniques to identify signs of dementia in patients' speech and language. Most of these studies are based on features extracted from audio recordings of people in order to detect dementia. Unfortunately, due to lack of standard datasets, each study is based on their own dataset (collected from different people with different medical conditions, and with different recording conditions). Sharing these datasets is not available due to ethical issues. The authors investigated different features from their datasets. However, due to high variances in recording conditions and the participants, it is very difficult to compare between the results gained by different studies.

Some researchers worked on extracting only acoustic features from the speech of people with dementia. Lopez de Ipina et al. [2013] investigated a number of acoustic features including durations (e.g. voice/unvoiced segments), time domain (short time energy), frequency domain (spectral centroid), and the fractal dimension from the AZTIAHO database of multilingual recordings of the spontaneous speech of 50 healthy adults and 20 Alzheimer patients. They used a Multi Layer Perceptron (MLP) classifier to distinguish between the patients with dementia and the healthy control group. Later Lopez de Ipina et al. [2015], introduced different fractal dimensional algorithms as a quantitative measurement capturing dynamic systems in multidimensional space. The accuracy of their classifier was around 80%.

Roark et al. [2011] extracted a number of speech-and language-related features from a recall task of the Clinical Dementia Rating (CDR) procedure, to distinguish between 37 healthy people and 37 people with MCI. They investigated the use of both manually annotated time alignments as well as an automatic approach (forced alignment of the ASR and automatic parsers) to identify different sets of features. They found that combining the features identified using the automated approach with the neuropsychological test scores outperformed other feature combinations.

Toth et al. [2015] found other acoustic and lexical features (such as the number of

phones per second, length of utterance and pauses) very useful in identifying patients with MCI. They have trained their ASR using the BEA Hungarian Spoken Language Database (spontaneous speech of people with MCI) focusing only on the phoneme recognition.

The same group later [Gosztolya et al., 2016] expanded the initial feature set from 27 features to 84 'extended' features included descriptors for silence pauses, filled pauses and some particular phones, and 708 'overcomplete' features redundant version of features with different descriptors for 57 phones, pauses, breathing noises, laughter and coughs. They investigated applying a number of different feature selection algorithms to identify the most informative and significant features for classification. The results revealed that training a classifier with a few features obtained by an efficient feature selection algorithm can outperform classifiers trained on all the features of the initial feature sets, the extended or overcompleted ones. They also suggested a new technique for the feature selection ('correlation-based' method) which can be as accurate as forward feature selection algorithm, yet, more efficient and faster. Recently, they used ASR outputs to extract the features [Toth et al., 2018]. They had 38 healthy controls as well as 48 patients with MCI. Using the most important features for classification, they gained a 75% accuracy rate (F1 measure 78.8%) by a Random Forest (RF) classifier.

Lehr et al. [2012] developed an automatic system to asses the MCI patient's memory as well as the healthy controls in retelling a story (part of the Wechsler Logical Memory (WLM) test) (35 MCI and 37 Healthy Control (HC)). They trained an ASR to recognise 25 elements from the story (keywords) both in immediate re-telling (patient should rephrase what they heard from the story without any delay) and delayed re-telling of the story (50 features to train an Support Vector Machine (SVM) classifier). Despite the high Word Error Rate (WER) (between 23% and 43%) of the ASR, the classification accuracy was close to the classification achieved by the manual transcripts with an 81.5% accuracy rate.

Jarrold et al. [2014] combined half of their ASR outputs with half of their human transcriptions of spontaneous speech to extract acoustic and lexical features and classified 48 participants into a group of patients with different types of dementia and a healthy control group. The classification accuracy amongst all types of subjects was 61%, while

the binary classification accuracy between AD and healthy controls rose to 88%.

Thomas et al. [2005] extracted several lexical and semantic features to achieve 95% accuracy in a binary classification task differentiating between patients with severe dementia and normal controls. The performance of such automated diagnostic approaches, however, dropped to around 75% when they attempted to differentiate between patients with mild dementia and healthy elders. When four classes of cognitive performance were introduced (severe dementia, moderate dementia, mild dementia, and a normal group), the classification accuracy decreased drastically to around 50%.

Satt et al. [2013] carried out a study on 89 subjects (43 with MCI, 27 with AD, and 19 healthy adults). The subjects were asked to complete tasks such as verbally describing a picture while looking at it, looking once at a picture and describing it from memory and repeating a sentence given by the interviewer. They extracted a number of vocal features for each task (e.g. total speech duration, the standard deviation of pause duration, the average verbal reaction time). Then they used N-lowest p-value approach to select a number of features and train classifiers. They gained 80% accuracy for a binary classifier between MCI and AD. Later on, modifying some of the features, they reported a classifier accuracy between HC and MCI of 79% and between HC and AD of 87% [König et al., 2015].

In a study on the data collected from the Interdisciplinary Longitudinal Study on adult development and aging (ILSE)(a German dataset consisting of 1000 participants' spontaneous speech in their middle adulthood and later life spanning, over 10000 hours recordings), Weiner et al. [2016] focused on extracting a number of acoustic and qualitative features (e.g. silence duration, silence to speech ratio, word rate, phoneme rate) to train a classifier to distinguish between three categories: AD (5 patients), Aging-Associated Cognitive Decline (AACD) (13 patients) and HC (80). They emphasised that the ratio of their selection of the different patient groups reflected a very similar real-life situation. They have used the manual transcriptions for the lexical features and trained a Voice Activity Detection (VAD) to calculate the acoustic features. Using a Linear Discriminant Analysis (LDA) classifier, they have obtained 85.7% accuracy amongst the three participant groups with 0.66 Unweighted Average Recall (UAR). While differentiating between the

healthy and AD was successful, the classifier was not capable of categorising the healthy group from the AACD patients. Then they trained an ASR to transcribe the conversations and extract a number of different features (acoustic, lexical richness, perplexity features, etc.) from both the manual transcriptions and the outputs of the ASR [Weiner et al., 2017]. Despite having a high mean WER (59.2%) of the ASR, the automatic within-speaker perplexity features (not the manual) achieved the best UAR of 0.623 among all features. They also trained an speaker diarisation and used it in their pipeline of automatic system to segment the audio files of the interviews [Weiner et al., 2018]. Using only the acoustic features and a Gaussian classifier, they gained 0.493 UAR for their original transcribed dataset. However, on an un-transcribed data of 218 subjects (in 241 interviews), the classifier achieved 0.645 UAR. They also found out that using only 12.5 mins of the interviews was fairly enough to gain the best results by the classifier.

Fraser et al. [2015] and Yancheva et al. [2015] used the Dementia Bank corpus (containing speech of patients with AD, vascular dementia, MCI and healthy controls describing the Boston 'Cookie Theft' picture[1]) to predict changes in patients' MMSE scores over time. The researchers extracted a wide range of features (over 477 lexico-syntactic, acoustic, and semantic) and selected the 40 most informative, reporting an accuracy of over 92% in terms of the distinction of AD patients from HC.

Their relatively high accuracy comes from the human transcriptions of the audio files, however, in their next study [Zhou et al., 2016] they used ASR to produce automatic transcriptions. Their best ASR had a 38.24% WER. Ignoring the prosodic and the acoustic features, this time, they have only extracted lexical features to train an SVM classifier in order to differentiate between the HC and AD patients. The accuracy of the classifier drops significantly as the ASR WER increases. Poor quality of the recordings and challenges of people's voice (with a high level of breathiness, jitter, shimmer, and slower rate) by ASR were reported as the main challenges for the ASR. What is more, the distinction between AD and HC represents much less of a diagnostic challenge in clinical practice than the differentiation of MCI and age-matched adults without cognitive complaints - or even age-matched adults with non-progressive memory complaints. The dataset contains

[1]'Cookie Theft' picture is a part of the Boston Diagnostic Aphasia Examination in which the examiner shows the picture to the patient asking them to tell about everything that is going on in the picture.

240 audio recordings of people with either "possible AD" or "probable AD" from 167 participants and 233 additional recordings of HC group from 97 speakers.

Yancheva and Rudzicz [2016] used Global Vector (GloVe) word vectors to make 10 common clusters of the words (only nouns and verbs) in the training set of the dementia group as well as 10 common cluster of words in the HC (each cluster representing a topic or related words, e.g. C0: window, floor, curtains, plate, kitchen, D0: cookie, cookies, cake, baking, apples). Then using the average scaled distance between the words of a given transcript from the test set and the created clusters, they extracted 20 semantic features. In addition they calculated the 'idea density' as the number of topics mentioned in the transcript divided by the number of words, and the 'idea efficiency' as the number of topics mentioned in the transcript divided by the length of recording. They gained 80% accuracy and 80% F1 when they combined all their features with the lexicosyntactic and acoustic features from Fraser and Hirst [2016].

Working on the same dataset, Al-Hameed et al. [2016] achieved a 94% classification accuracy using acoustic-only features, thereby avoiding the need to use ASR. They also predicted the MMSE scores from the acoustic features as well as adding an MCI group for classification [Al-Hameed et al., 2017]. They gained an accuracy rate between 89.2% and 92.4% for the pairwise classifications.

Asgari et al. [2017] used the Linguistic Inquiry and Word Count (LIWC) software to categorise the words of the transcripts of interviews between interviewers and people with MCI and healthy group (14 MCI and 27 HC). The subcategories included 68 different categories like positive emotions and negations, fillers, home, sport, job, nonfluencies ('um','er'). They passed these 68 dimensions for each conversation to train an SVM classifier with an Radial Basis Function (RBF) kernel. They gained a 76.2% accuracy rate for the classification.

Other modalities aside from voice have also been investigated and found to be good predictors for cognitive decline and dementia including eye movement [Parsons et al., 2017; Zhang et al., 2016], olfactory sense [Karunanayaka et al., 2017; Lafaille-Magnan et al., 2015] and even hand dexterity [Stringer et al., 2018].

In general, the distinction between AD and HC represents much less of a diagnos-

tic challenge than the differentiation of MCI and age-matched adults without cognitive complaints, or even age-matched adults with non-progressive memory complaints.

In brief, recent research has demonstrated that automatic audio and speech technology may provide diagnostic markers that can aid the classification between e.g. HC and people with AD or MCI. However, most studies have focused on providing a supplementary, automatic method based on existing test procedures currently used in the clinical settings like picture description. In addition, many research studies have used manual transcription, thereby side-stepping the known challenges associated with the automated analysis of spontaneous, conversational speech.

3.2 Dementia detection system

In this section we introduce the pipeline of our dementia detection system as well as our original dataset for the study, the "Hallamshire data". Our initial focus was on replicating the findings of Elsey et al. [2015] with human transcriptions. Therefore, we started with the last component of the system (feature extraction and classification). In order to prove the feasibility of the approach, a pilot study was conducted and the results published in Mirheidari et al. [2017b]. The rest of this chapter includes details of the study.

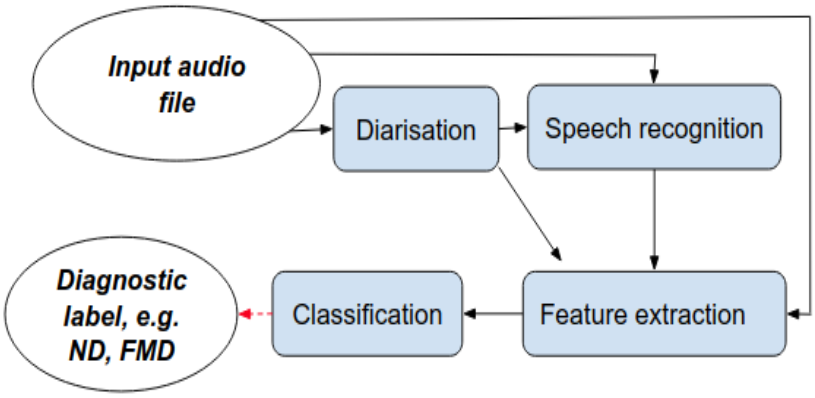


Figure 3.1: *Automatic dementia detection system.*

A fully automatic dementia detection system would comprise of several units including the ASR, the speaker diarisation, the feature extraction and the machine learning

classifier (Figure 3.1)(It is worth mentioning that Weiner et al. has started working on a similar pipeline since 2016. In their recent paper [Weiner et al., 2018], they included the diarisation unit, similar to our dementia detection system). First, an audio file containing a recording of the conversations is passed to a diarisation tool to identify the speech portions of the input audio stream, as well as the speaker identification of each speech segment. Diarisation techniques first identify the speech and non-speech (silence, music, background noise, etc.) portions of the input audio stream, then, processing the speech parts of the input streams, they identify the speaker of each segment. This information is then passed to an ASR system. The ASR is given both the input audio file and the output produced by the diarisation tool to generate a string of words spoken by each speaker (patient, neurologist and AP). Next, the output of the diarisation tool and the ASR are given to the feature extraction unit to extract a number of features. Some features may rely on techniques such as signal-, text- and natural language processing as well as spoken language understanding. A number of acoustic features can be extracted directly from the audio recording. Finally, the extracted features are sent to a machine learning classifier to decide which category the whole conversation belongs to; e.g. ND and FMD. Note that, there were one or two accompanying persons in the conversations. Therefore we considered all of the other speakers as the APs.

The diarisation, ASR and classification are further described in the next chapters.

3.2.1 Hallamshire data

The first step for assessing people with memory difficulties is the history-taking from the patient and accompanying person(s). It is believed that the patient's history is a key to appropriate diagnose and treatments. Therefore, the interaction between doctor, AP and patient during the history-taking is assumed to play an important role for the diagnosis process. The initial motivation was to apply CA to find diagnostic clues from the conversations in different patient groups.

Recruiting the participants for Hallamshire study took place between October 2012 and October 2014 at the Royal Hallamshire Hospital in Sheffield, United Kingdom. The participants were routinely encouraged to bring someone along to the memory clinic ap-

pointment if possible (AP). All participants underwent MRI brain imaging and cognitive screening using the ACE-R. Participants underwent detailed neuropsychological testing with a neuropsychological battery which included the MMSE [Folstein et al., 1975], tests of short and long term memory (verbal and non-verbal) [Wechsler, 1997], tests of abstract reasoning [Raven, 1995; Rey, 1964], tests of attention and executive function [Stroop, 1935], language comprehension, naming by confrontation, category and letter fluency [De Renzi and Faglioni, 1978].

The participating doctors were encouraged to adhere to a communication guide (for the guidelines of the study refer to Appendix B), which had been developed in close co-operation with these clinicians and was based on their routine practice. Neurologists were guided to start their history-taking with an open enquiry, not explicitly directing patients to talk about their memory problems. They were encouraged to maximise patients' opportunities to produce an account of their own concerns and to minimise interruptions. After this open beginning, neurologists were asked to prompt further extended talk from patients by encouraging them to give an example of when their memory let them down. Finally, the communication guide listed some specific enquiries (such as who was more concerned about the memory difficulties, the patient or others). The ACE-R was carried out after the history-taking and not recorded or analysed. Demographic information of the participants for this study as well as information about the cognitive test scores are presented in Table 3.1.

There were 30 audio recordings of the interviews[1] between the neurologists, patients and APs with an average recording length of 16 minutes and an average utterance length of 4.6 seconds for each interview (see Table 3.2 for more information about the Hallamshire data set).

3.2.2 Extracted features

In the process of translating qualitative features from the earlier CA study [Elsey et al., 2015; Jones et al., 2015], it became apparent that, in most cases, several complementary programmable features had to be combined to generate a reasonably close translation of a

[1] we only had access to the audio files of the interview

Table 3.1: *Demographic information of the participants.*

	FMD (n=15)	ND (n=15)	Mean	Cut Off	Score range	P-value
Age	57.8(±2.02)	63.73(±2.29)	*N/A*	*N/A*	*N/A*	$p = 0.06$
Female	60%	53%				ns^*
ACE-R	93.0(±1.4)	58.27(±5.21)		88	0 – 100	$p < 0.0001$
MMSE	28.87(±0.19)	18.79(±1.97)	28.88(1.28)	26.32	0 – 30	$p < 0.0001$
PHQ-9	5.6(±1.02)	5.25(±2.04)		5	0 – 27	ns
GAD-7	4.73(±1.23)	4.75(±1.52)		5	0 – 21	ns
CF	19.8(±0.11)	17.15(±0.93)	19.65(0.63)	18.39	0 – 20	$p = 0.0052$
VPA	16.87(±0.74)	5.85(±0.94)	14.81(3.76)	7.29	0 – 24	$p < 0.0001$
P&PT	51.13(±0.19)	44.50(±2.49)	51.23(0.82)	49.59	0 – 52	$p = 0.0063$
Rey's CF	34.0(±0.44)	21.42(±3.02)	33.70(2.30)	29.1	0 – 36	$p < 0.0001$
SF	52.73(±2.91)	23.77(±4.03)	59.81(13.17)	33.47	N/A^{**}	$p < 0.0001$
PF	41.2(±3.02)	19.15(±3.69)	45.58(12.05)	21.48	N/A^{**}	$p < 0.0001$
DS	6.73(±0.33)	4.54(±0.48)	6.76(1.48)	3.8	0 – 9	$p = 0.0007$
VCA	13.2(±0.2)	10.08(±0.97)	13.77(0.51)	12.75	0 – 14	$p = 0.0023$
TT	34.97(±0.27)	26.50(±1.89)	34.67	1.03	0 – 36	$p < 0.0001$
PM	15.07(±0.92)	5.25(±1.1)	12.37	2.08	0 – 25	$p < 0.0001$

Legends: ACE-R: Addenbrooke's Cognitive Examination - Revised; MMSE: Mini Mental State Examination; PHQ-9: Patient Health Questionnaire-9; GAD-7: Generalised Anxiety Assessment 7; CF: Confrontational Naming; VPA: Verbal Paired Associates; P&PT-Pyramid & Palm Trees; Rey's CF: Rey's Complex Figure; SF: Semantic Fluency; PF: Phonemic Fluency; DS: Digit Span; VCA: Visuoconstructive Apraxia; TT: Token task; PM: Prose Memory. trials.
*: not significant
**: For the CF and PF tests there is no maximum score as it depends on individuals' word production speed within the time limit of three one minute trials. We have included all the maximum scores on the cognitive tests apart from GAD-7 and PHQ-9 where we have included the minimum to reflect a score if no depression or anxiety were present.

Table 3.2: *Hallamshire data set information.*

Number of audio recordings	30 (FMD: 15, ND: 15)
Number of speakers	81 (Patient: 30, Neurologist: 30, AP: 21)
Number of utterances	6266
Total length of recordings	8 hours
Average recording length	16 minutes
Average utterance length	4.6 seconds

Table 3.3: *Qualitative features from Table 2.2, and corresponding features extracted from the transcripts by automatic analysis of conversation. Prefixes: Pat=patient, Neu=neurologist, and APs=accompanying person(s).*

Qualitative feature [Elsey et al., 2015]	**Corresponding extracted feature(s)**
F1. Role of accompanying persons (APs)	Number of turns (**APsNoOfTurns** , **PatNoOfTurns**); Average length of turn (**APsAVTurnLength**, **PatAVTurnLength**); Average unique words (**APsAVUniqueWords**, **PatAVUniqueWords**)
F2."Who's most concerned?"	Patient answered "me" (**PatMeForWhoConcerns**)
F3. Recall of recent memory failure	Number of empty words (**PatFailureExampleEmptyWords**); average length of pauses (**PatFailureExampleAVPauses**); used "all the time" (**PatFailureExampleAllTime**)
F4. Inability to answer	Patient replies "I don't know" to the question about their expectations of the memory clinic appointment (**PatDontKnowForExpectation**); Frequency of "don't know" responses in combination with turning to APs (**PatAVNoOfDontKnow**); Average instances of head shakes (**PatAVNoOfShakesHead**); Average number of filler words (**PatAVFillers**); Average number of empty words (**PatAVEmptyWords**); Average number of common words (**PatAVAllWords**)
F5. Responding to compound questions	Average number of repeated questions (**AVNoOfRepeatedQuestions**)
F6. Patients' elaborations	Average unique words (**PatAVUniqueWords**, Average turn length **PatAVTurnLength**)
Role of the neurologist (Not investigated in original study)	Number of turns (**NeuNoOfTurns**); Average length of turns([sec]) (**NeuAVTurnLength**); Average number of unique words (**NeuAVUniqueWords**); Average number of topics discussed (**AVNoOfTopicsChanged**); Average length of pauses by patient (**PatAVPauses**)

qualitative observation (see Table 2.2). Table 3.3 shows how the six key features described qualitatively were translated into 17 features suitable for conversation based analysis. Note that [Elsey et al., 2015] refereed to these qualitative features simply as "features", but to not cause confusion for the readers we refer to them as qualitative features. In addition, we defined five potentially diagnostic features suited for conversation based analysis which focused on the interactional contributions of the neurologist, but which were not based on any previous qualitative findings (see Table 3.3).

To extract some of those features, a common NLP approach known as the Bag of Words (BoW) model [Salton, 1983] was used. This technique underpins many search

engines (like Google) and is supported by numerous NLP packages (e.g. Natural Language Toolkit (NLTK) [Bird et al., 2009]). The BoW ignores the order of words, punctuation, commonly used words in English (such as 'the', 'a', 'this'), and trims verbs to their stems. For instance, for the clause 'He wanted to get a new job', the BoW would contain the words: 'want', 'new', and 'job'.

F1.(Role of the APs): The detection of most of the features depends on an automatic way of identifying turns, i.e., splitting the conversation into questions and answers. This is a relatively hard task. However, this study is based on the automated analysis of a small number of highly structured conversations (30 conversations in total), in which new topics are almost exclusively initiated by the clinician. This means we were able to use a far simpler topic detection[1] method relying on the detection of particular words or phrases in a turn. This facilitates the extraction of features aiding the identification of the role of the APs (F1). Features such as the number of turns, the average length of turn and the average number of unique words produced by the patient and the APs can be used individually to determine whether the patient or the AP talks more. A total of six features are defined to select the dominant speaker: the number of turns in the conversation (PatNoOfTurns and APsNoOfTurns), the average length of the turns (PatAVTurnLength and APsAVTurnLength), and the average unique number of words in the whole conversation (PatAVUniqueWords and APsAVUniqueWords).

F2.(Who's most concerned?): To extract information related to who is the most concerned about the patient's condition (F2), the topic detection approach described above is used first to identify the question, and subsequently assess the associated answer to determine whether the patient has replied that they are the most concerned (in effect answering "me" or similar words (I, myself, etc.)) or not (PatMeForWhoConcerns). Since not all the patients were asked this question, the feature actually had three possible values: "yes", "no", and "not available".

F3.(Recall of recent memory failure:) It relates to the question when patients last noticed a problem with their memory. Patients with ND were found to give three different types of answer to this question: providing mostly empty words, answering with

[1]detecting the topic of a particular question, i.e. what they are talking about.

a lot of hesitation or gaps in the speech, or answering something to the effect of 'all the time'. Therefore, three features were defined to capture the answer to this question: number of empty words in the response (PatFailureExampleEmptyWords), the average length of silences within the utterances (PatAVPauses), pause for failure example (PatFailureExampleAVPauses), and replying "all the time" (PatFailureExampleAllTime).

F4.(Inability to answer): In order to extract the feature "inability to answer" (F4), five different features were defined. The feature PatDontKnowForExpectation indicates that either the patient has replied "I don't know" or used a similar phrase in response to the question about what expectations they had when they came to the clinic. Elsey et al. [2015] also described "don't know" responses at other points of the interaction as diagnostically meaningful, although they differentiated between different types of this particular response: contextualised "don't knows" in which the speaker provides appropriate information addressing parts of a question but identifies particular aspects s/he is unable to answer, or non-contextualised "don't know" responses in which no attempt is made to provide a more detailed reply to any aspect of a question. To improve the diagnostic contribution of "don't know" statements, we, therefore, did not only count these utterances (PatAVNoOfDontKnow), we also coded additional information sometimes associated with these words (such as patient turns head to the AP encouraging them to answer the question instead of the patient). Similarly, we coded head shaking (translated into the feature PatAVNoOfShakesHead). Other important features, which may be helpful in determining the meaning of "don't know" statements, are the average number of filler words like "I mean", "I see" (PatAVFillers), the average number of empty words such as "er", "em" (PatAVEmptyWords) and the average number of words in a turn (ignoring very common words such as "a", "the", "that", PatAVAllWords).

F5.(Responding to compound questions): In their responses to compound (multi-part) questions (F5), ND patients typically failed to answer all parts of the question so the neurologist had to repeat the question in the following turn. This is captured by feature AVNoOfRepeatedQuestions which takes into account parts of compound questions which were not answered by the patient straight away.

F6.(Patient's elaborations) The lack of elaboration of answers by patients with

Table 3.4: *Types of extracted features: acoustic, lexical, semantic and visual-conceptual.*

Type	**Features**
Acoustic	APsNoOfTurns PatNoOfTurns NeuNoOfTurns APsAVTurnLength PatAVTurnLength PatFailureExampleAVPauses NeuAVTurnLength PatAVPauses
Lexical	PatAVUniqueWords NeuAVUniqueWords APsAVUniqueWords PatAVAllWords
Semantic	PatMeForWhoConcerns PatFailureExampleEmptyWords PatFailureExampleAllTime PatDontKnowForExpectation PatAVFillers PatAVEmptyWords AVNoOfRepeatedQuestions AVNoOfTopicsChanged
Visual-conceptual	PatAVNoOfShakesHead PatAVNoOfDontKnow

ND was captured by the features PatNoOfTurns, PatAVTurnLength, and PatAVUniqueWords.

In addition, we also extracted three extra features based on the contributions of neurologists (NeuNoOfTurns, NeuAVTurnLength and NeuAVUniqueWords). Although the differential diagnostic value of the neurologists' contribution has not been studied explicitly by Elsey et al. [2015], it has been identified as a conversational observation of potential value by others [Hamilton, 1994; Perkins et al., 1998]. Similar to the APs and patient features, NeuNoOfTurns, NeuAVTurnLength, and NeuAVUniqueWords were identified. Finally, the feature AVNoOfTopicsChanged takes into account the average number of different topics discussed by the neurologist and patient throughout the conversation.

The extracted features can be divided into four different types: acoustic, lexical, semantic and visual (non-verbal). Table 3.4 lists all features.

3.3 Baseline results

There are several standard machine learning classifiers, however, choosing the best classifier for a given dataset is a challenging task, because each one has advantages and

Table 3.5: *Classification accuracy rate using all 22 extracted features (10-fold cross validation).*

Classifier	**Accuracy**(%)
Linear SVM	90
Random Forest	92.5
AdaBoost	95
Perceptron	90
Logistic Regression	92.5
Linear via SGD	90
AVG (STD)	*91.7(2.04)*

Table 3.6: *P-values of significance test (five repeats of 2-fold cross validation) for the pairs of the six classifiers. P-value less than 0.05 indicates a significant difference.*

-	**SVM**	**RF**	**AdaBoost**	**Perceptron**	**LR**	**SGD**
SVM	-	0.2484	0.9394	1.0000	**0.0318**	0.1829
RF	0.2484	-	0.5119	0.0582	0.5157	1.0000
AdaBoost	0.9394	0.5119	-	0.9469	0.7644	0.5367
Perceptron	1.0000	0.0582	0.9469	-	0.6338	0.4868
LR	**0.0318**	0.5157	0.7644	0.6338	-	0.3741
SGD	0.1829	1.0000	0.5367	0.4868	0.3741	-

disadvantages, depending on factors such as the number of samples of training and testing data, and the variances of the different features in the data. Therefore, a very common methodology is to try several classifiers and use a validation approach to find the best classifier for a particular dataset.

The focus of this study was the differentiation between patients with ND and FMD, so a binary machine learning classifier was used. The "Scikit-learn" [Pedregosa and Varoquaux, 2011] is a Python library with a wide range of machine learning classifiers. From this library, six standard machine learning classifiers were chosen: SVM with linear kernel, RF, Adaptive Boost (AdaBoost), Perceptron, Logistic Regression (LR), and Stochastic Gradient Descent (SGD).

In this study, we used a common evaluation technique, the 10-fold cross-validation approach. In this approach, data is divided into 10 groups. For each group, the data of the group is held out as a test set and the remaining groups are used for training. The average accuracy over all the test sets determines the accuracy of the classifier. Table 3.5 displays the overall accuracy in percentage for the six evaluated classifiers using all 22

features extracted from the transcripts using 10-fold cross validation. The best accuracy rate was achieved by the AdaBoost classifier with 95%, while the minimum accuracy was achieved by three classifiers: linear SVM , linear via SGD and Perceptron classifiers all with 90%. The mean classification accuracy of all classifiers was 91.7% (with a standard deviation of 2.04%).

McNemar's test (McNemar [1947]) could not identify any significant differences between the accuracy rates of the six classifiers. We also applied another machine learning classifier's significant test suggested by Dietterich [1998] in which the classification should be repeated 5 times using a 2-fold cross validation. This method only showed a significant difference between the LR and the linear SVM classifiers with a p-value of 0.0318. Table 3.6 shows all the p-values calculated for each pair of the six classifiers. Note that the p-value of comparing the Perceptron classifier with the RF classifier was the second minimum p-value in the table, but it was slightly bigger than the 0.05 threshold (0.0582), hence does not constitute a significant difference.

Thus, based on the significance test, the LR classifier accuracy was significantly better than the SVM classifier, while there were not significant differences between the remaining classifier. In the subsequent chapters only one classifier was chosen (LR) as the nominated classifier.

3.3.1 Feature selection

Generally, the best features to use in automated classification approaches are complementary and highly discriminative for the task at hand. In practice though, it is common for two types of features to exhibit a high degree of interdependence, and a process of feature selection is often beneficial. This makes the machine learning model simpler (fewer features need to be extracted), regulates the variance amongst the extracted features and, more importantly, reduces the risk of overfitting (many features do not necessarily yield better classification, but rather make the final prediction too dependent on a specific dataset [Guyon and Elisseeff, 2003]).

One approach is to consider the input data (disregarding the output classes) with the aim of identifying those features with the greatest variance and diversities using statistical

tests such as t-tests. Another approach considers the outputs of the classification task in order to find the most discriminative features. Feature selection in this way depends on the amount a particular feature contributes to the classification. Some classifiers, such as those that are based on trees, automatically use feature contribution for the classification task. Therefore, they have a built-in ranking which can show the importance of features.

For linear classifiers, Recursive Feature Elimination (RFE) (see [Pedregosa and Varoquaux, 2011] for more details) is a common approach to selecting the most significant features. RFE finds the most important features by examining how eliminating each feature from the feature set affects the classification accuracy. One by one, the feature making the smallest contribution is eliminated and the accuracy of the remaining features is evaluated. Elimination continues until all features have been eliminated. Reverse order elimination shows the importance of the features in the classification task.

For the tree-based classifiers (AdaBoost and RF) the built-in ranking was employed and for the other linear classifiers, the RFE technique (on the train set) was used to identify the best features. The most significant (top 10) features overall were selected by combining the feature rankings of five classifiers (the six classifiers except the Perceptron). Table 3.7 lists the most important features contributing to the classification. The most significant five features were the average number of unique words used by the neurologist (rephrasing the questions somewhat differently depending on the patient), accompanying person's number of turns, the average number of unique words used by the patient, the average turn length for the patient, and the average number of repeated questions.

There are other approaches such as component analysis e.g., the Principle Component Analysis (PCA) which can be used to reduce the dimensionality of the features. However PCA is better suited to reducing very large datasets (e.g., with hundreds of features) and also, by using feature selection methods directly affected by the classifier in question, we ensure we identify the most important features for the task at hand. This also enables us to arrive at a subset of features that needs extracting as opposed to PCA-based reduction which would still require us to extract all features prior to dimensionality reduction. Applying the PCA resulted in a classification accuracy rate of 92.5% for the Perceptron, the SVM and the LR, however, the accuracy rate for the SGD, the AdaBoost and the

Table 3.7: *The most significant (top 10) features with the highest contributions for the classification between the ND and the FMD patients using the RFE on the train set.*

Rank	Feature Name
1	NeuAVUniqueWords
2	APsNoOfTurns
3	PatAVUniqueWords
4	PatAVTurnLength
5	AVNoOfRepeatedQuestions
6	PatFailureExampleEmptyWords
7	PatAVFillers
8	PatAVAllWords
9	PatMeForWhoConcerns
10	PatAVPauses

RF dropped to 87.5%, 72.5% and 67.5% respectively. The McNemar test and 5 repeats of 2-fold cross validation test showed these declines were statistically significant.

Using only the most significant (10) features instead of all 22 features resulted in a better performance for most of the five nominated classifiers. The mean accuracy of correct diagnosis prediction improved to 92.5% across all classifiers with standard deviation of 2.74%. While the correct classification rate of the RF dropped from 93% to 90%, the accuracy rate for the linear SVM, the LR rose from 90% and 93% to 95% and 95% respectively. The accuracy rates for the remaining three classifiers stayed the same (see Figure 3.2). The McNemar test and 5 repeats of 2-fold cross validation test, however, did not show any significant differences between the accuracy rates of the classifies using the 10 features.

In order to show that a feature is statistically important in discriminating between two or more classes, it is necessary to first determine whether the values of the feature are distributed normally (i.e. follow the Gaussian distribution). For the features which are normally distributed, we can then use the **Student's t-test** to show the values of the feature are significantly different in the classes. For the features which are not normally distributed we can use a **non-parametric test** to show the significance difference.

To determine the statistical normality of the 22 features extracted from the manual transcripts of the conversations, we used two normality tests: Shapiro-Wilk test(Shapiro and Wilk [1965]) and D'Agostino's K-squared test. For the features which could pass

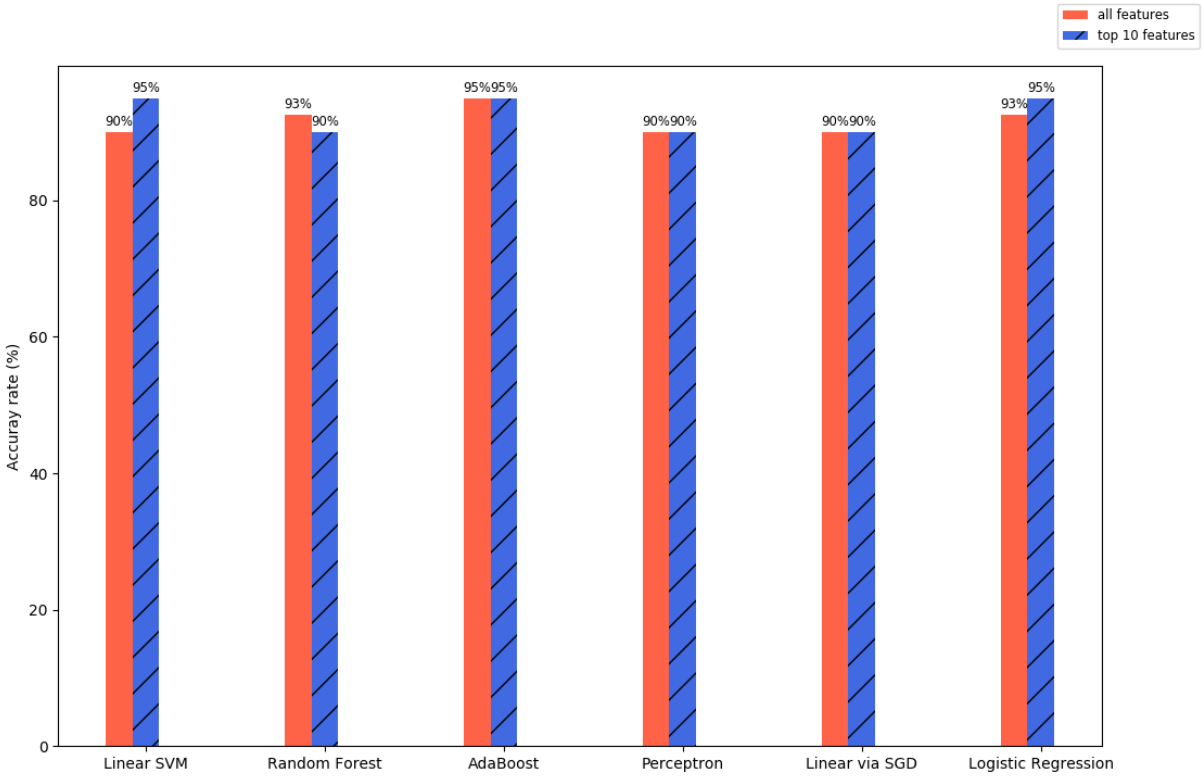


Figure 3.2: *Comparison of accuracy rates using individual classifiers based on all features and the most significant (10) features.*

the normality tests we then applied the two tailed Student's t-test and calculated the p-values, and for the non-normal features we applied the Mann-Whitney U test (Mann and Whitney [1947]), which is an effective way to show the significant difference for the values which are not normally distributed.

The results showed that 10 features out of 22 had values which were significantly different between the two classes (FMD and ND). Table 3.8 shows the 10 features with significance difference between the two classes using the statistical tests. Note that three features AVNoOfRepeatedQuestions, PatDontKnowForExpectation and PatMeForWho-Concerns could not pass the Shapiro-Wilk normality test, but they passed the D'Agostino normality test. We could consider these three features as “soft” normal features.

Using the 10 most statistically significant features we calculated the accuracy rates of the classifiers. The mean accuracy of correct diagnosis prediction improved to 93.75% across all classifiers with standard deviation of 2.09%. Figure 3.3 compares the accu-

Table 3.8: *The 10 most statistically significant features using the normality tests (Shapiro-Wilk and D'Agostino) and then parametric (Student's t-test) for normal (norm.) features and non-parametric (Mann-Whitney U test) for non-normal (non-norm.) features. Feature marked with star were in the top 10 features in Table 3.7.*

No.	Feature Name	Shapiro-Wilk	p-value	D'Agostino	p-value
1	NeuNoOfTurns	non-norm.	0.0339	non-norm.	0.0339
2	*APsNoOfTurns	non-norm.	0.0006	non-norm.	0.0006
3	*AVNoOfRepeatedQuestions	non-norm.	0.0053	norm.	0.0086
4	*PatAVTurnLength	norm.	0.0004	norm.	0.0004
5	*PatAVUniqueWords	norm.	6.5e-6	norm.	6.5e-6
6	PatAVEmptyWords	norm.	0.0199	norm.	0.0199
7	PatDontKnowForExpectation	non-norm.	0.0269	norm.	0.0446
8	*PatMeForWhoConcerns	non-norm.	0.0015	norm.	0.0014
9	*PatAVAllWords	non-norm.	3.4e-5	non-norm.	3.4e-5
10	*PatAVFillers	non-norm.	0.0004	non-norm.	0.0004

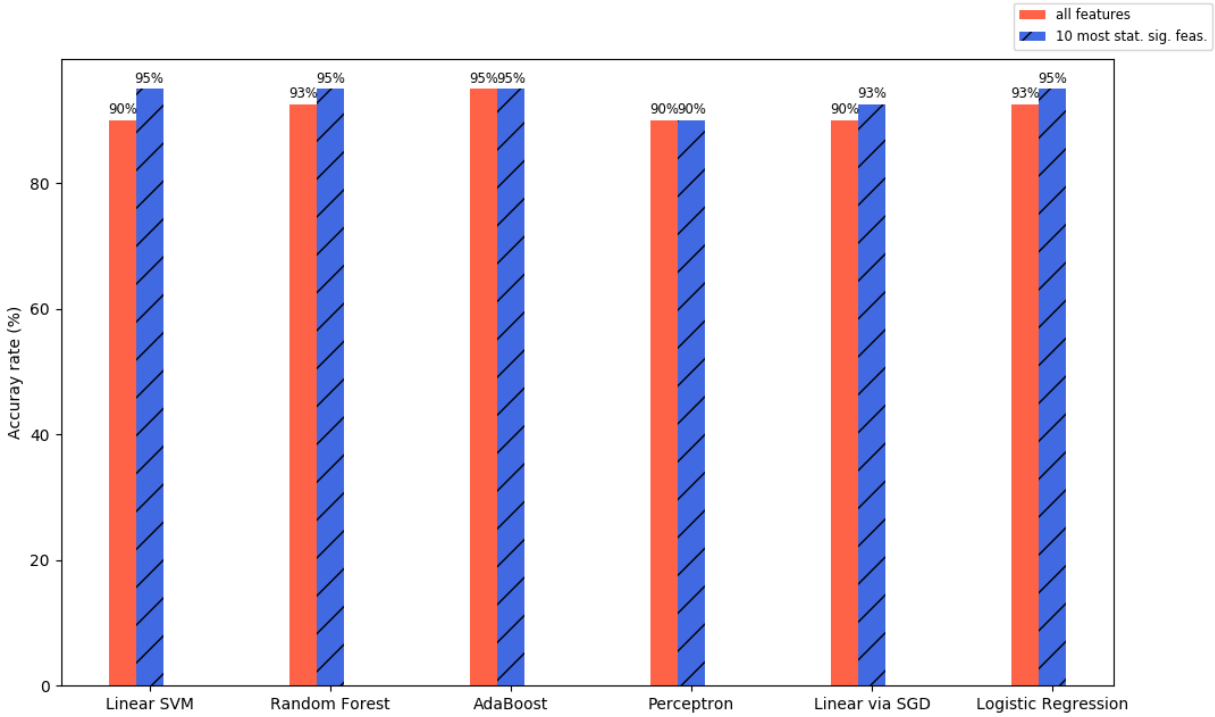


Figure 3.3: *Comparison of accuracy rates using individual classifiers based on all features and the 10 most statistically significant features.*

racy rates gained by these 10 features to all 22 features. While the accuracy rate of the AdaBoost and the Perceptron remained the same, accuracy rate of the remaining four classifiers rose to 95%. The McNemar test and 5 repeats of 2-fold cross validation test,

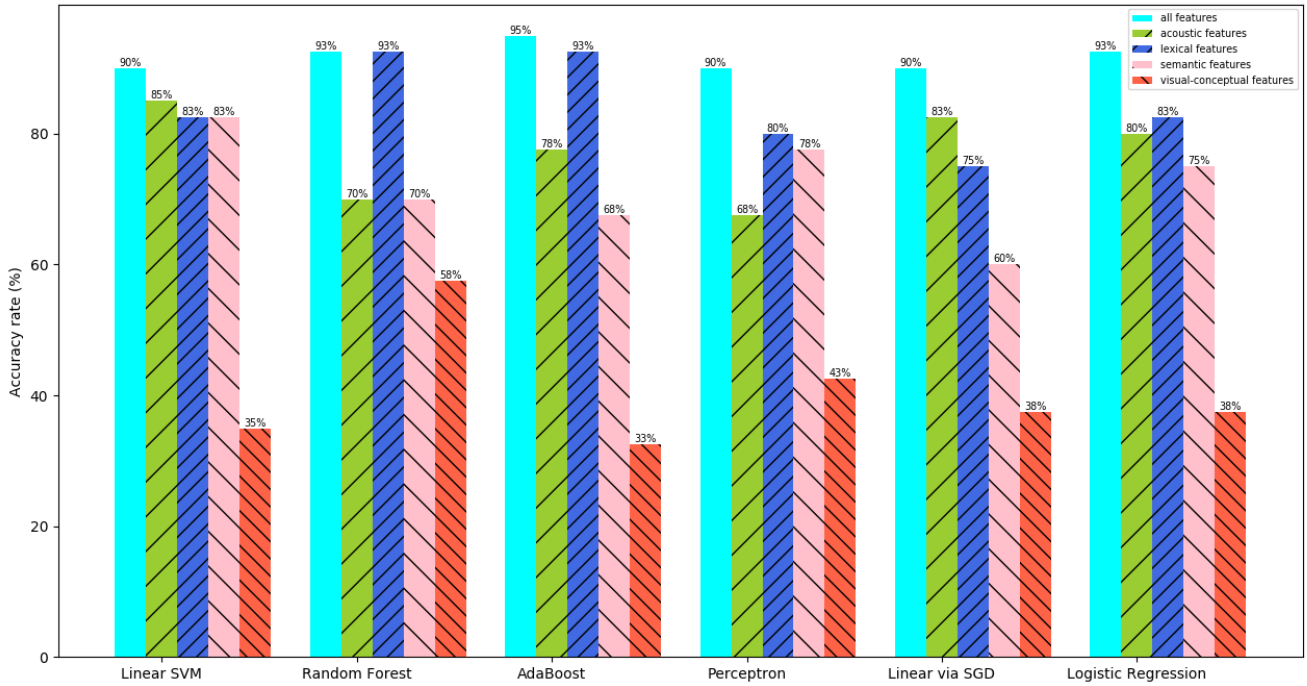


Figure 3.4: *Classification accuracy rates for different types of features acoustic, lexical, semantic and visual-conceptual, as well as all features.*

did not reveal any significant differences between the accuracy rates of the classifies using these 10 features. Comparing the 10 most statistically significant features with the top 10 features from the RFE approach, 7 features are common between the two selections. The accuracy rates from the former are slightly higher but not significantly different.

3.3.2 Feature type importance

In order to identify the relative diagnostic contribution of individual feature types, the classification task was repeated using only acoustic, only lexical, only semantic and only visual-conceptual features. The results are presented in Figure 3.4; however, the importance of feature type depends on the classifier itself to some degree. So, in brief, visual-conceptual features were the least discriminant with an accuracy rate ranging from 58% for the RF to 33% for the AdaBoost (an average accuracy rate of around 41% across all the classifiers), whereas the lexical features were the most important feature types with an average of around 85% contributions for the six classifiers. Acoustic and semantic features

were the second and third most important feature types with an average of around 77% and 73% across the six classifiers respectively.

Looking again to Tables 3.7 and 3.8 we can see that none of the top features in both tables includes the visual-conceptual features. That indicates that they contributed less discriminative information than the two classes. However, the top 10 features from the RFE feature selection and the 10 most statistically significant features includes the remaining three features types.

3.4 Discussion

The early diagnosis of ND and their distinction from normal ageing or FMD is a challenging clinical task and recently, the number of referrals from primary care to specialist memory clinics has increased considerably [Royal College of Psychiatrists, 2016]. However, many patients referral are associated with subjective memory concerns, but not dementia [Bell et al., 2015c; Larner, 2014; Menon and Larner, 2011]. Currently the decision to refer is based on a GP's interpretation of the history given by patient and informant (such as partner, friend or family member) and the result of short screening tests. Although these tests have a high sensitivity, they have a low specificity for dementia [Boustani et al., 2005; Hessler et al., 2014].

This study explored whether the insights gained by an expert qualitative study of detailed transcripts can be used to develop an automated screening process for ND. We tested a simplified version of the automatic analysis of conversation, focussing on machine learning and classification. Six standard machine learning classifiers were used. The accuracy rate of all the classifiers were high (an average accuracy of 92% across the classifiers) and we could not find statically significant differences (using 5 repeats of 2-fold cross validation test) between the classifiers except for a significant difference between theLR and the SVM . In the subsequent LR classifier was chosen as the nominated classifier

Using the most significant (10) features from the RFE selection approach as well as using (10) most statistically significant features resulted in slightly improved overall performance than using all initially defined features (an average of 92.5% and 93.75%

respectively). Visual-conceptual features were the least important for the classification, while the lexical features were the most important features followed by the acoustic and semantic features respectively.

In addition, the study has identified some new interactional features with diagnostic potential now requiring further in-depth analysis using methods such as CA: quantitative features extracted from the contributions which APs and neurologists made to the conversation were amongst the most significant features. The contributions of these individuals were not studied in the previous qualitative studies of memory clinic encounters by Jones et al. [2015] and Elsey et al. [2015]. Whilst the interactional role of APs in clinic conversations requires more research, the conversational role of caregivers to people with dementia (i.e. individuals with more significant cognitive problems than those exhibited by the patient group described here) has been studied by Perkins et al. [1998], focussing on turn taking, repair and topic management. They found that caregivers had a key role in successful conversations. For instance, caregivers used touch, gaze and the patient's name before talking, to achieve better responses from patients. Greater familiarity between patient and caregiver reduced dysfluencies, mishearing and misunderstanding, while unfamiliarity between the interviewer and the patient resulted in fewer topic initiations.

It is possible that the differences in neurologists' communication behaviour in encounters with ND patients on the one hand and FMD patients on the other, which we picked up by automated CA in this study, are due to the fact that they became aware of the diagnosis relatively early in the consultation. Future studies will need to examine whether less expert clinicians (for instance those working in primary care) would change their communication in similar ways and whether they could be made more aware of that fact that they are adjusting their conversational style (which could help with the diagnostic process).

This study has several limitations. Although the recruitment of patients first referred to a memory clinic with cognitive concerns increases the clinical validity of our findings, our recruitment method means that the findings cannot be readily generalised to patients complaining of memory problems in other settings, for instance in primary care. Furthermore, we were only able to analyse a relatively small number of conversations (30). The

patients whose interactional behaviour we studied, however, represented two neurologically well-characterised groups. Importantly, our study did not compare patients with ND with healthy controls but with patients with FMD, enhancing the practical relevance of our findings.

We assumed perfect accuracy of transcription by ASR (close to the manual transcripts), which is not an uncommon first step in this research area. Looking ahead, this part of an automated analysis of conversation will be one of the most difficult aspects and will need to be the focus of further studies. It is possible that features not described here would perform better diagnostically if less perfect transcripts than used in this study were employed in a fully automated diagnostic procedure. Furthermore, in this initial proof-of-concept study we focused on a relatively small number of features described by Elsey et al. [2015]. There are, however, potentially many other distinctive semantic, acoustic and lexical features that could be extracted from audio or video recordings which may further improve the classification accuracy.

Following chapters will include more in-depth analysis of the classifiers, Additional data allows us to analyse e.g. confusion between multiple classes.

It is worth mentioning that, initially, when we started the study, the group of neurologists that we were working with, were interested in discriminating between the FMD and the ND groups, but gradually they recruited more patients with MCI and finally included HC subjects. They wanted to gradually add the new groups to avoid the complexity of four-way classification. That is why we started with a binary classifier (FMD vs. ND) in Chapter 3 , and Chapter 6, and then in Chapter 7 we included the MCI patients and finally, in Chapter 8 we put together all the four patient groups.

3.5 Summary

Automatic analysis of conversation for diagnosing dementia is a relatively new field of study. Most studies worked on extracting acoustic features from the audio recording of people and pass them to a classifier to distinguish between healthy group vs. AD and/or MCI groups. The accuracy rate of binary classifiers (between two groups) were considerably high, however, as the number of groups increased, the accuracy rate of the classifiers dropped considerably. Some studies extracted other types of features such as lexical/syntactical and semantic features from the manual transcripts of the speech or they have used ASR to automatically convert the speech to text. A number of studies achieved a low accuracy rate using the ASR comparing to the manual transcripts, but some reported results as good as using the manual transcripts. Different studies focused on different datasets which has made it difficult to compare between their results.

In this chapter, we introduced the pipeline of our automatic system to identify dementia, which compromises a speaker diarisation unit to identify the speakers and the segments of their speech in a conversation. The output of the diarisation is given then to an ASR to convert to text. Then a feature extraction unit uses the output of the ASR as well as the audio file of the conversation to extract different types of features. Finally the features are passed to a classifier to categorise the entire conversation.

A pilot study is conducted to find an answer to the first research question, concenterating the feasibility of developing an automatic tool to identify dementia through analysis of conversation.

The results of our initial study on our dataset, the Hallamshire, suggests that automated analysis of conversation has the potential to improve the screening and triage procedures for patients with possible ND. The improvement of case selection for referral to specialist clinics would mean that those at high risk of developing dementia could be seen more quickly, whilst those with FMD could be reassured at an earlier stage in the clinical management pathway. Although further work is required to develop our method into a screening tool that could be deployed in primary care, the approach described here has the advantage of being non-invasive and usable in a wide range of health care settings.

Despite its limitations, this study demonstrates the feasibility of translating interactional findings derived from the qualitative study of transcripts into features which can be automatically extracted and analysed. Our findings show that such an automated process has the potential to improve the early identification of patients at high risk of developing dementia. At the same time our study provides further support for the validity of analysing conversation.

Chapter 4

Automatic speech recognition

Contents

4.1 Spontaneous speech recognition 64

4.1.1 Front-end processing 65

4.1.2 Acoustic modelling 66

4.1.3 Language modelling 68

4.1.4 Search (decoding) 69

4.1.5 LVCSR challenges 70

4.2 Deep neural networks 72

4.2.1 RNN/LSTM and TDNN 75

4.2.2 End-to-end ASR 78

4.3 Semi-supervised learning 80

4.4 State-of-the-art 82

4.5 Automated transcription 86

4.5.1 Baseline ASR 87

4.5.2 Adding additional data 88

4.5.3 Improving the Acoustic Model (AM) 90

4.5.4 Improving the Language Model (LM) 91

4.5.5 WER per speaker group 93

4.6 Discussion 95

4.7 Summary 98

In the previous chapter we introduced the pipeline of our automatic system to detect dementia analysing conversation. In this chapter we introduce the ASR component of the system. So this chapter is an investigation to find an answer to the second research question, what kinds of speech technologies are needed for developing dementia detection system. The chapter is structured as follows:

Section 4.1 is an introduction to spontaneous speech recognition, different components of the long vocabulary continuous speech recognition system and the main challenges of the system.

Section 4.2 discusses about the deep neural networks and in particular Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM).

Section 4.3 is about 'semi-supervised learning'. Using this approach we can improve both acoustic and LMs of ASR.

Section 4.4 lists some of the state-of-the-art results reported by other authors on the common datasets dedicated to conversational speech recognition.

Section 4.5 gives details of the ASRs we trained for dementia detection and the results gained by the systems.

Section 4.6 and 4.7 are the discussion and the summary of this chapter, respectively.

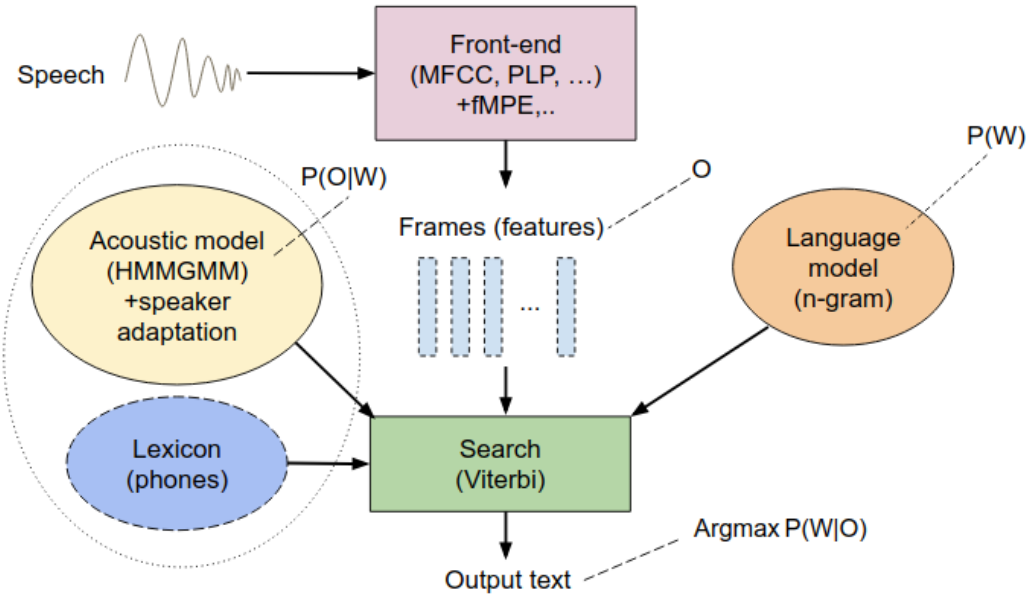


Figure 4.1: *General architecture of a conventional LVCSR system.*

4.1 Spontaneous speech recognition

In recent years, Large Vocabulary Continuous Speech Recognition (LVCSR) has extended to spontaneous speech for various applications such as telephone conversations, lectures, seminars, meetings, etc. At the moment, using state-of-the-art ASR, WER of non-spontaneous speech (e.g. for reading mode) can easily reach better than 5%, however, for natural spoken speech, the error rate is still high due to the acoustic and linguistic differences between the written (formal) and the spoken (informal) language and the unpredictable nature of spontaneous speech. The structure of written language is well-formed and known, while natural spoken language is normally loosely structured with a high degree of flexibility and variance. The mismatch between training and testing conditions is also a fact and reflects the low level of robustness of the current LVCSR systems [Saon and Chien, 2012]. Thus training a decent spontaneous speech recognition system is a challenging task. The state-of-the-art of the speech recognition system at the moment has not solved totally the classical problems such as background noise, channel distortions, foreign accents, casual and disfluent speech, and unexpected topic change which can be easily handled by human listeners.

The main purpose of an LVCSR system is to convert the acoustic signals of continuous speech to a sequence of corresponding words. Although the acoustic signal is highly variant, in a very short interval (around 10-15 ms) it is assumed that it remains invariant and it is possible to extract a number of features representing the speech for the short interval (these feature vectors are called observations).

Words can be represented as a sequence of phones. There are around 45 phones in the English language and normally dictionaries (lexicons) are required to determine the way phones construct a given word. An ASR attempts to find a mapping between the observations and the phones guided by a language model (search). However, this requires the training of good representatives for each phone in advance using the available dataset (input: observations and output: related phones). The recognised phones, then, can be used for producing the outputs of the ASR. Mono-phones, however, are not usually used, as it is very hard to find the borders between phones, both inside words and between two consequent words, since neighbouring phones affect each other.

In addition to the current phone, AM normally takes into account a number of preceding phones as well as succeeding phones (e.g, tri-phones: the previous phone + current phone + the next phone).

Language model, which is normally trained on a general corpus, provides more generic information about how likely a word is to appear in a phrase or sentence. These information can boost the ASR performance by narrowing down the search space.

Figure 4.1 shows a general architecture of a modern LVCSR system which consists of four major components front-end processing, AM, LM, and search (decoding). More details are given in the following paragraphs.

4.1.1 Front-end processing

The main purpose of this part is to extract acoustic feature vectors from the input wave files. This is normally carried out by framing the input speech signals using the short term Fast Fourier Transform (FFT) of a window of 25-30 ms length at each 10 ms. The energies of the neighbouring frequencies within each frame are then binned together using the Mel scale filter bank followed by applying logarithm and Discrete Cosine Transform

(DCT). The output representation of each frame is 13-dimensional Mel Frequency Cepstral Coefficient (MFCC). In order to capture the dynamics of the features, normally deltas and delta-deltas are calculated as well, i.e. 39 dimensions in total.

In early years of ASR history, a number of studies reported slightly better results using the Perceptual Linear Prediction (PLP) coefficients for feature extraction for noisy environments compared to MFCC [Cui et al., 2003; Rajnoha and Pollák, 2011].

In order to improve the performance of system, the acoustic features can be normalised using methods such as Cepstral Mean Subtraction (CMS) and Cepstral Variance Normalisation (CVN). Transformation techniques such as LDA are also used at this stage to make the feature vectors distributed with diagonal co-variance Gaussians (feature dim reduction), which in turn helps to have much easier computations for the AM. In recent years other feature level techniques have been proposed such as noise robustness methods (e.g. Stereo Piece-wise Linear Compensation for Environment (SPLICE) [Droppo et al., 2001] and Quantile Equalisation (QE) [Hilger and Ney, 2006]), speaker adaptive approaches (e.g. Vocal Tract Length Normalisation (VTLN) [Eide and Gish, 1996], feature-space Maximum Likelihood Linear Regression (fMLLR) [Gales et al., 1998]), and discriminative techniques (e.g. feature-space Minimum Phone Error (fMPE) [Povey et al., 2005]).

In recent years more robust features have been extracted from the audio files specially for training Deep Neural Network (DNN)s, for instance the log-scaled mel spectrogram (Petridis and Pantic [2016]; Salamon and Bello [2017]; Salamon et al. [2017]) and deep bottleneck features (Mun et al. [2016]; Nguyen et al. [2013]; Yu and Seltzer [2011]).

4.1.2 Acoustic modelling

The AM aims at training the acoustic representatives (statistics) for each output unit (e.g. phone, tri-phone, word). One of the most important approaches for AM is Hidden Markov Model (HMM)s [Rabiner, 1989] which can represent sequential data (like speech) by a set of states, transitions and the probabilities of each state and the transition between the states.

Figure 4.2 displays an example of an HMM with 6 states, the transition proba-

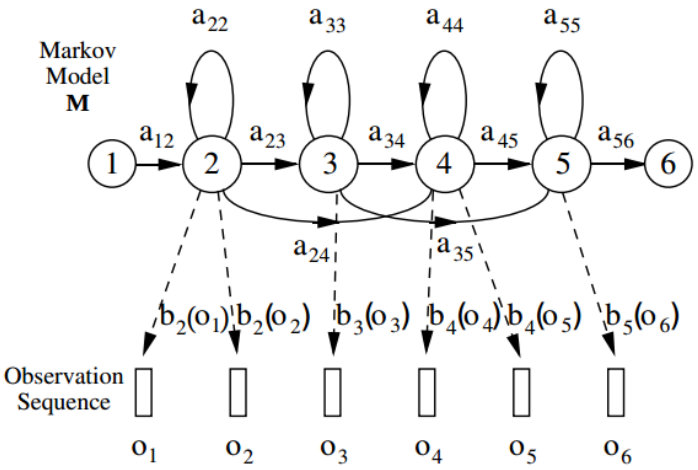


Figure 4.2: *Example of an HMM with 6 states for observation sequence o_1 to o_6, output probabilities for each state $b_j(o_t)$, and transition probabilities a_{ij}'s [Young et al., 2006].*

bilities between the states (a_{ij}'s) and the probability distribution of observing a feature vector in a certain state, $b_j(o_t)$ (output probabilities). The sequence of the states $X = x(1), x(2), ..., x(T)$ are unknown or hidden, however, given the Markov model M, the likelihood of observing $O = o_1, o_2, ..., o_T$ can be calculated by summing up over all possible state sequences as Young et al. [2006]:

$$P(O|M) = \sum_{X} a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \tag{4.1}$$

The output probabilities normally are represented by a Gaussian Mixture Model (GMM), means and co-variance of K Gaussian components, each with a weight (all the weights together sums up to 1). The probabilities for the states, transitions and the weights can be calculated using the generative Maximum Likelihood (ML) criteria (or minimising the empirical risk in respect to the joint likelihood loss [Deng and Li, 2013]) using a number of the Expectation-Maximisation (EM) (or Baum-Welch method [Levinson et al., 1983]) steps or iterations.

Despite the great success of the HMMs, there are a few limitations of using these stochastic models. The main issue is the assumption that the sequence of observations is independent. So their probabilities can be written as a product of the individual ob-

servation, and the assumption that the distribution of the individual observation can be parametrised by GMMs. Also, the Markov assumption (that the probability of being at a state at time t, only depends on the state at time $t-1$) may not be totally true for speech [Rabiner, 1989].

In recent years, discriminative algorithms based on DNN have shown a better performance and more robustness comparing to the conventional HMM-GMM approach [Abdel-Hamid et al., 2014; Bengio, 2009; Graves and Jaitly, 2014a; Hinton et al., 2012, 2006; Seide et al., 2011]. Discriminative learning can be used in two ways: Using a discriminative model directly or employing a discriminative training objective function to a generative model [Deng and Li, 2013].These approaches use different criteria to train the AMs such as Minimising Classification Error (MCE), Maximum Mutual Information (MMI), Minimum Phone Error (MPE), and Maximum Word Error (MWE) [Deng and Li, 2013; Saon and Chien, 2012].

4.1.3 Language modelling

The LM provides information about the sequences of words in a general context (or a special domain), e.g. what is the probability of the word 'are' following the word 'there'?

According to the chain rule of probability, the probability of seeing a sequence of words w_1 to w_n (or w_1^n) can be written as the product of the conditional probabilities of seeing previous words as:

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)...P(w_n|w_1^{n-1}) = \prod_{k=1}^{n} P(w_k|w_1^{k-1}) \tag{4.2}$$

The most common LM approach, the n-gram, only takes into account $N-1$ preceding words to calculate the probability of seeing a sequence of words:

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-N+1}^{k-1}) \tag{4.3}$$

Generally the Maximum Likelihood Criteria Estimation (MLE) is used to calculate the conditional probabilities by counting the occurrence of $N-1$ previous words before

word n from a corpus and normalising it (dividing by the total number of occurrence of $N-1$ previous words before any words). For a bi-gram (when $N = 2$; C stands for count) the probability can be written as:

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)} = \frac{C(w_{n-1}w_n)}{C(w_{n-1})} \tag{4.4}$$

The main issue of n-gram LM is zero probabilities for unseen words in training dataset. There are techniques to deal with this issue such as smoothing by adding one or k to the counts, back-off (using n-1 gram to estimate the probabilities) and interpolation (mixing with probabilities from other models) e.g. Kneser-Ney (KN) and Good Turing (GT) smoothing [Chen and Goodman, 1996].

4.1.4 Search (decoding)

Decoders (e.g. Viterbi algorithm [Forney, 1973]) attempt to find the optimal sequence of words or hypotheses for an input feature vector sequence using the acoustic and LMs. This is performed usually by the equation 4.6 taken from the well-known Bayes' rule (4.5) that combines the acoustic and LM together (note that the denominator is ignored since the observation sequence is the same for all sequence of the words):

$$P\left(W|O\right) = \frac{P\left(O|W\right).P(W)}{P(O)} \tag{4.5}$$

$$\hat{W} = \underset{w}{argmax}\ P\left(W|O\right) = \underset{w}{argmax}\ P\left(O|W\right).P(W) \tag{4.6}$$

where W is a sequence of words, $W = w_1, w_2, ..., w_n$, O is a sequence of observed feature vector (pre-processed audio files into feature vector sequence) and $O = o_1, o_2, ..., o_t$. The posterior probability $P\left(W|O\right)$ is computed from the AM and the prior probability of words $P(W)$ obtained by the LM.

Viterbi beam search (Viterbi [1967]) is one of the common solutions for performing efficient The Weighted Finite-State Transducer (WFST) [Mohri and Pereira, 2002] is a technique which can efficiently combine different sources of knowledge together (such as

HMMs, pronunciation dictionaries, grammars, LM and context trees). When combining large amounts of different knowledge, however, the computational cost of running them grows considerably up.

4.1.5 LVCSR challenges

In addition to the general problems of an LVCSR and informal and unpredictable structure of spontaneous speech, there are a number of other fundamental challenges. The following is a list of the major issues which had been gathered in an early survey by Shriberg [2005]. These issues still have not been solved, and have been in the focus of a number of recent studies such as Moore [2015]; Nakamura et al. [2008]; Sani et al. [2015]; Verkhodanova and Shapranov [2015].

Hidden punctuation and sentence boundaries: Punctuation is an essential part of written language which is very helpful for 'automatic downstream processes' (e.g. parsing, information extraction, summarization) as well as human readability. However, spoken language lacks explicit punctuation and relies mostly on prosody (pitch, duration, stress and intonation). The outputs of current ASRs, however, are streams of words without any punctuation. There have been a number of studies dealing with this issue by combining the ASR models with pauses [Wald, 2013], or use other approaches including knowledge-based information, machine translation techniques [Bell et al., 2015b], supervised machine learning approaches [Blanchard et al., 2016], and DNN based approaches (e.g. Tilk and Alumäe [2015] and Chan et al. [2016]).

Dealing with disfluency: Disfluencies (e.g. filled pauses, reputation, repair, false start) are non-separable parts of natural speech which occur very frequently (appears in up to one-third of our natural utterances) and downgrades the performance of ASR severely. There have been various efforts for resolving these natural phenomena such as detecting and removing disfluencies or clean-up before speech recognition [Kaushik et al., 2010], developing statistical modelling of disfluencies [Lease et al., 2006], transition-based techniques [Wu et al., 2015], and detecting or predicting disfluencies using different approaches such as using prosodic information and pauses, syntax, Conditional Random Fields (CRF)s and DNNs [Cho et al., 2015; Christodoulides et al., 2015].

Turn-taking and overlapping: In spontaneous speech, speakers take turns not always in a sequential manner, but often in a competitive approach and they may start a turn before the current speaker's turn has finished. The overlapping speech affects ASR and introduces a considerable amount of errors to the system. Some techniques to reduce the effect of overlapping include source separation [Heittola et al., 2013] and auditory scene analysis [Okuno et al., 2007], recording each speaker using a separate channel [Dat et al., 2016], and multi-speaker LM. Overlap detection can be a part of a speaker diarisation system.

Emotions and para-linguistic: What is heard from a natural spoken conversation, is normally more than only the sequences of words, and often the speech is mixed up with emotions (e.g. laughter, anger, stress) which are difficult to capture correctly by an ASR. A number of different approaches have been proposed for emotion detection such as prosodic feature extraction [Yu et al., 2009], using machine learning classifiers [Kumar and RangaBabu, 2015] and DNN based techniques [Laffitte et al., 2016; Rawat and Mishra, 2015].

For the purposes of this study, dealing with all of the above-mentioned challenges is not essential. Punctuation and sentence boundaries are very useful for reading a text and are essential parts of CA. As mentioned above, it would be possible to combine the outputs of the ASR with a corpus of English sentences to add punctuation to the output of ASR, however, for natural language processing or feature extraction purposes, it is also possible to use algorithms and techniques which do not rely directly on the sentence boundaries or particular punctuation. For instance, the BoW ignores the punctuation, capital letters, and also very common words (e.g. 'the' in English). Also, displaying emotions (e.g. laughing, crying) in the outputs of ASR similarly can be ignored (although ASRs can be trained to distinguish different spoken noise such as laughter). Dealing with overlapping speech can be a part of a speaker diarisation system and some disfluencies can be captured by ASR applying an in-domain LM. It is worth mentioning that due to having all of these issues in our dataset, we expect a high WER for the ASR and we will investigate how the error will effect the results of our dementia detection system.

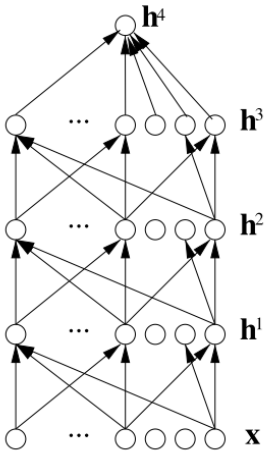


Figure 4.3: *Multi-layer neural network with feed-forward way from the input x, through the hidden layers h^j, to the network output h^4 [Bengio, 2009].*

4.2 Deep neural networks

DNNs are the current common approaches for machine learning in general and speech recognition in particular, which can outperform the other conventional methods if enough training data and memory space are provided and high speed Central Processing Unit (CPU)s (or Graphics Processing Unit (GPU)s) are employed.

A DNN, simply, is a feed-forward Artificial Neural Network (ANN) or MLP with more than one hidden layer between its inputs and outputs (see Figure 4.3) [Bengio, 2009; Hinton et al., 2012; Jiang, 2010]. Each hidden unit, j, typically uses a 'logistic function' (e.g. hyperbolic tangent) to map its total input from the previous layer x_j to a scalar value, y_j which is then sent to the next layer [Hinton et al., 2012]:

$$y_j = logistic(x_j) = \frac{1}{1+e^{-x_j}}, x_j = b_j + \sum_i y_i w_{ij} \tag{4.7}$$

where b_j is the bias of unit j, i is an index over units in the previous layer and w_{ij} is the weight on a connection to unit j from unit i. For a multiclass classification task, the output unit j converts its total input x_j into a class probability, p_j using the 'softmax

nonlinearity' [Hinton et al., 2012]:

$$p_j = \frac{e^{x_j}}{\sum_k e^{x_k}} \tag{4.8}$$

where k is an index over all classes.

DNNs can be discriminatively trained by 'backpropagating' derivatives of a cost function (or objective function or errors) measuring the discrepancy between the target outputs and the actual outputs produced for each training case. In a forward propagation the costs are calculated and in a backward propagation the partial derivatives of the costs with respect to the weights are calculated and can then be used to update the weights. For softmax output function, the cost function C is the cross-entropy between the target probabilities, d, and the outputs of softmax, p [Hinton et al., 2012]:

$$C = -\sum_j d_j log(p_j) \tag{4.9}$$

where the target probabilities taking values of one or zero, are the supervised information provided for training the DNNs. For a phone recogniser, for instance, the softmax outputs can be used as the phone's posteriors, i.e. an output node for each phone and its value is the posterior of observing that phone.

For large training sets, it is typically more efficient to compute the derivatives on a small random "mini-batch" of data rather than the whole set, before updating the weights in proportion to the gradient. The "stochastic gradient descent" method for updating the weights, then, can be improved by using a momentum coefficient $0 < \alpha < 1$ that smooths the gradient computed for mini-batch, t, damping oscillation across ravines and speeding progress down ravines [Hinton et al., 2012]:

$$\Delta w_{ij}(t) = \alpha \Delta w_{ij}(t-1) - \epsilon \frac{\partial C}{\partial w_{ij}(t)} \tag{4.10}$$

One of the major issues of the DNNs is "overfitting" in which the layers learn well the parameters from the training data, however, facing unseen test data, they perform poorly and cannot discriminate the classes as expected. To reduce the effects of this problem,

large weights can be penalised in proportion to their squared magnitude, or the learning process can stop when the performance on a held-out validation set starts getting worse. However, the most effective approach for reducing overfitting is pre-training which provides much better starting points and makes faster discriminative fine tuning. Techniques such as Restricted Boltzmann Machine (RBM) and autoencoders are used for pre-training DNNs. RBM uses one forward step (prediction) and one backward step (using the prediction guess for the initial input) in order to predict its input' probability distribution (normally by the KullbackLeibler Divergence (KLD) criteria [Bengio, 2009; Hinton et al., 2012, 2006]).

"Dropout" is another technique to deal with overfitting [Cheng et al., 2017]. In this approach the activations are multiplied by random zero-one masks only during the training stage. For instance, the dropout probability p=0.5 means half of the masks are ones. Dropout technique allows for the training of robust networks which are not too fitted to the training set input data. However, setting a proper dropout probability and locating it in deep neural network layers is not straightforward.

DNNs can be trained as a multiple classifier on a frame-level cross entropy criteria, however, since the ASR is naturally a sequence classification, a sequence-discriminative criteria (MMI, MPE, or state-level minimum Bayes risk (state-level Minimum Bayes Risk (sMBR)) can be used to estimate the HMM states. This is known as a hybrid DNN-HMM approach [Bourlard and Morgan, 2012; Veselỳ et al., 2013].

For the utterance u at time t, the HMM output for the state s can be obtained by the softmax activation function as the following equation [Veselỳ et al., 2013]:

$$y_{ut}(s) = P(s|O_{ut}) = \frac{e^{a_{ut}(s)}}{\sum_{s'} e^{a_{ut}(s')}} \tag{4.11}$$

where O_{ut} is the observation for the utterance u at time t, and $a_{ut}(s)$ is the activation function in the output layer for the state s.

A pseudo log likelihood can be obtained as:

$$log(P(O_{ut}|s)) = log(y_{ut}) - log(P(s)) \tag{4.12}$$

where, $P(s)$ is the prior probability and can be gained from the training data.

For the sequence of all observations $O_u = O_{u1}, ..., O_{uT}$, the MMI objective function can be calculated as:

$$F_{MMI} = \sum_u log(\frac{P(O_u|S_u)^K P(W_u)}{\sum_W P(O_u|S_u)^K P(W)}) \tag{4.13}$$

Where W_u is the reference word sequence for utterance u, and $S_u = S_{u1}, ..., S_{uT}$ is the sequence of states for W_u.

MPE/sMBR criteria can be written as:

$$F_{MBR} = \sum_u \frac{\sum_W P(O_u|S_u)^K P(W) A(W, W_u)}{\sum_{W'} P(O_u|S_u)^K P(W')} \tag{4.14}$$

Where $A(W, W_u)$ is the number of correct phone labels for MPE or correct state labels for sMBR.

4.2.1 RNN/LSTM and TDNN

Data in a feed-forward neural network flows in one direction, from the input layer towards the output layer, and the network only looks at the current input data. However, in a RNN, A recurrent neuron has a loop (directed connections in time series) which allows it to look at the previously seen input data (known as short term memory) as well as the current input data. The internal state of RNN stores the previous input data. Therefore, RNN can capture the contextual information which allows it to model the sequential data better than the feed-forward neural networks. Figure 4.4 shows an RNN with one input, one output and one recurrent hidden layer with a dotted cycle [Lipton et al., 2015].

In a RNN weights are updated using the backpropagation through time approach [Williams and Peng, 1990]: each RNN can be considered as a sequence of neural networks. The errors backpropagate from the last time-stamp towards the first time-stamp. RNNs, however, suffer from two issues: exploding and vanishing gradient. In updating the weights, the gradient measures how much the output changes in respect to a the change in input. In the exploding gradient (high slope), the weights

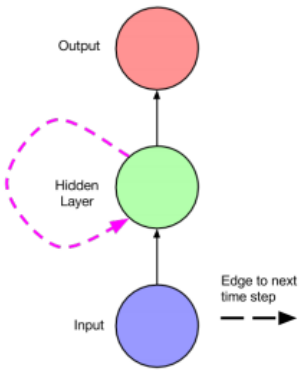


Figure 4.4: *A simple RNN with one input, one output and one recurrent hidden unit [Lipton et al., 2015].*

increases drastically, while in the vanishing gradient the slope gets very small which causes it practically to loose any learning. In addition, as the length of the sequence increases, the computational cost of calculations rises drastically. The Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] is introduced to resolve these issues. LSTM can learn long time memory as well as short term. This happens using gated cells. Initially, there were two types of gates: input and output gates, then Gers et al. [1999] introduced other gate known as forget gates. These three gates act similar to what we can do with the memories in computers, i.e. read (input gate), write (output gate), and delete (forget gate). In training the networks, LSTM learns which data should be kept and which can be ignored. The gradient can be controlled accordingly thereby avoiding the exploding or vanishing issues.

Recently, LSTMs have been successfully used for the AM of LVCSR [Sak et al., 2014] as well as LM [Sundermeyer et al., 2012].

Figure 4.5 shows an LSTM architecture with a recurrent project layer and an optional non-recurrent projection layer suggested for the AM by Sak et al. [2014]. Normally the LSTM network consists of an input layer, a recurrent LSTM layer and an output layer, however, they introduced two more layers, one recurrent and one optional non-recurrent layer as in Figure 4.5.

For an input sequence $x = x_1, ..., x_T$ and an output sequence $y = y_1, ..., y_T$ (time

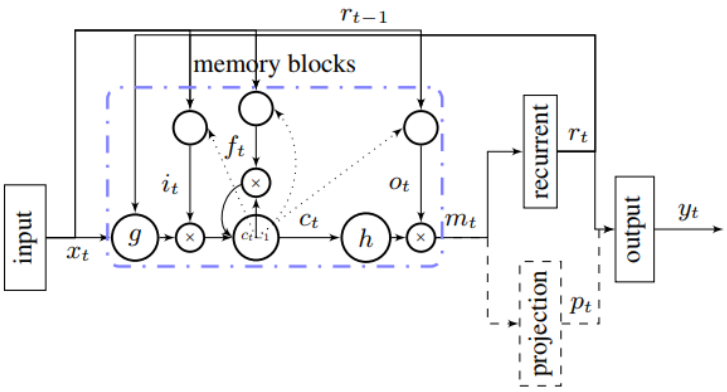


Figure 4.5: *LSTM with a recurrent projection and an optional non-recurrent projection layer [Sak et al., 2014].*

$t = 1, ..., T$), the activation units equations are [Sak et al., 2014]:

$$i_t = \sigma(W_{ix}x_t + W_{ir}r_{t-1} + W_{ic}c_{t-1} + b_i) \tag{4.15}$$

$$f_t = \sigma(W_{fx}x_t + W_{rf}r_{t-1} + W_{cf}c_{t-1} + b_f) \tag{4.16}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g(W_{cx}x_t + W_{cr}r_{t-1} + b_c) \tag{4.17}$$

$$o_t = \sigma(W_{ox}x_t + W_{or}r_{t-1} + W_{oc}c_t + b_o) \tag{4.18}$$

$$m_t = o_t \odot h(c_t) \tag{4.19}$$

$$r_t = W_{rm}m_t \tag{4.20}$$

$$p_t = W_{pm}m_t \tag{4.21}$$

$$y_t = W_{yr} r_t + W_{yp} p_t + b_y \tag{4.22}$$

Where W denotes weights, b denotes biases, σ is the logistic sigmoid function, i, o, and f indicate input, output and forget gates, c shows cell activation vector, m cell output vector, $\odot$ element wise product, g and h, cell input and cell output activation functions, r and p recurrent and non-recurrent unit activations.

Similarly to the AM, RNNs and LSTMs can be used for LM. RNNs take into account the whole history of the words in training and testing which can overcome the issues of the conventional n-gram based LM, i.e. only considering n previous/next words and backing-off. Combining RNN/LSTM with an n-gram approach results in an efficient and robust LM [Mikolov et al., 2010; Sundermeyer et al., 2012].

The Time Delay Neural Network (TDNN) Waibel et al. [1990] is a feed-forward neural network in which the neural network units are arranged in a hierarchy and get input from their lower layers activations as well as time delayed inputs (previous data, in time t_{-1}, t_{-2}, etc.). This allows shift-invariance (no need to prior alignments) and models long time context. However, for modern AM other techniques should be combined with LSTMs to make it efficient. For instance, Peddinti et al. [2015] applied sub-sampling to improve the speed and performance of LSTMs. Figure 4.6 shows the sub-sampling which needs a low number of connections and efficient computations using LSTMs (e.g. 10 times faster). Instead of considering all the previous time steps, only in a selection of them activations are computed.

4.2.2 End-to-end ASR

Recently, end-to-end ASR systems have been introduced to simplify the complex pipeline of ASR. They attempt to directly find a map between speech and words. In a hybrid DNN-HMM approach, we need the alignment information of the trained HMMs, prior to building the DNNs which may take time and resources. Connectionist Temporal Classification (CTC) [Graves et al., 2006; Graves and Jaitly, 2014b] is an objective function introduced to train RNNs directly without knowing the alignments between input and

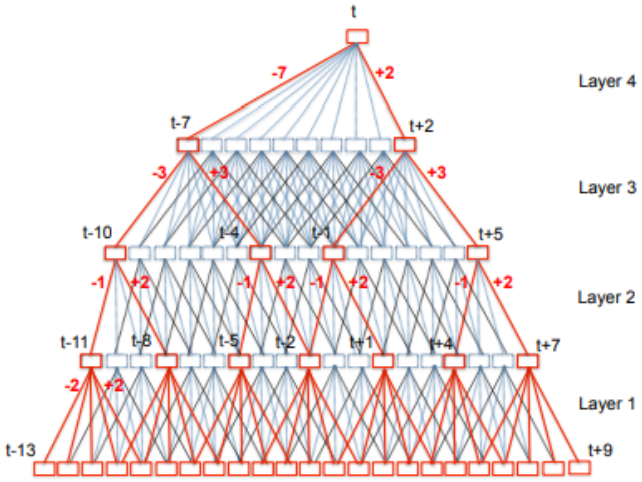


Figure 4.6: *Sub-sampling technique to improve computation cost of LSTMs. Sub-sampled connections (red lines), original LSTMs connections (blue and red lines) [Peddinti et al., 2015].*

target sequences. The output layer compromises a single unit for each individual label as well as a blank label representing the 'null emission'. For an input sequence $x = x_1, ..., x_T$ and the output vector y_t at time t, the probability of emitting label can be denoted as:

$$P(k, t|x) = \frac{e^{y_t^k}}{\sum_{k'} e^{y_t^{k'}}} \tag{4.23}$$

where y_t^k is element k of y_t. A CTC alignment 'a' is gained using the chain rule and by multiplying probability emission labels over time-stamps as:

$$P(a|x) = \prod_{t=1}^{T} P(a_t, t|x) \tag{4.24}$$

Removing repeated and blank labels from the alignments using the β operator, the probability of an output transcript y can be written as:

$$P(y|x) = \sum_{a \in \beta^{-1}(y)} P(a|x) \tag{4.25}$$

For a given target transcript y^*, the objective function can be written as:

$$F_{CTC} = -log(P(y^*|x)) \tag{4.26}$$

Decoding for the CTC networks can be done using a beam search algorithm which also enables the integration with an LM [Graves and Jaitly, 2014b]. However, for an efficient decoding approach, WFSTs are used for decoding. Miao et al. [2015] introduced a combination of a grammar, a lexicon and a token WFST to use for decoding CTC networks.

An alternative to CTC is encoder-decoder networks. Bahdanau et al. [2016] used a bidirectional RNN to encode the input sequence as feature representatives and then an Attention-based Recurrent Sequence Generator to decode the representatives to a sequence of labels. The attention mechanism tempts to select the temporal locations over input sequence that needed to be updated and used to predict the output data.

4.3 Semi-supervised learning

Generally, making annotations ('faithful transcripts') ready for spontaneous speech recognition is a challenging task. For broadcast news, as Li et al. [2015] mentioned, even given the closed caption text, a lot of work is still needed to make the annotations ready for the ASR, especially due to the differences between the actually uttered phrases and what is transcribed (e.g. filler words and repairs).

The 'semi-supervised' approach is a solution for training ASRs with a low amount of 'labelled data' and a considerable amount of 'unlabelled data'. A conventional semi-supervised technique is known as 'self-learning' in which, first a small amount of the labelled data is used to make a poor LM for the ASR and then, automatically use it to decode the larger set of unlabelled data (automatic transcripts). The transcripts can be filtered (selecting a subset of segments) using measures such as confidence scores. These outputs, finally, can be used to make a new and perhaps stronger language and acoustic models for the ASR [Kemp et al., 2004; Novotney and Schwartz, 2009; Zavaliagkos and Colthurst, 1998].

However, in recent years, there have been attempts to develop more effective approaches to semi-supervised learning mostly based on DNNs. For instance, for the lightly supervised scenario (e.g. closed captions are available for the broadcast news), Li et al. [2015] found out that using the conventional matching technique to filter the automatic transcripts can result in missing some informative data (e.g. when caption the text is wrong but the ASR output is right) and also generates extra errors (e.g. when both the caption text and the ASR are wrong but they match each other). Using binary classifiers and performing verification they could recover the informative data and reduce the errors. They used a mixture of GMM and DNN approach for training AMs for the ASR. Later they extended their work to a semi-supervised approach without the limitation of the lightly supervised assumption, i.e. they used a significant amount of unlabelled data to boost the AMs [Li et al., 2016].

Graph-based semi-supervised learning is another approach, in which the labelled and unlabelled data are jointly used to construct weighted graphs in which the nodes represent the data samples and the edges contain the similarity between the data samples. The closer data samples receive the same or similar labels, while the dissimilar data is given different labels [Kanda et al., 2016]. Although the performance of this approach is relatively high, it suffers from a computationally high complexity (constructing and keeping the weighted graphs in memory).

Dhaka and Salvi [2017] proposed a frame-based phone classification on the Texas Instruments-Massachusetts Institute of Technology (TIMIT) database using semi-supervised sparse auto-encoder (an encoder DNN followed by a decoder DNN, representing data by sparse matrix) by combining the supervised cost function of a deep classifier with the unsupervised cost function of the auto-encoder. Using the graph-based algorithm, however, they reported a slightly better performance comparing to their proposed approach.

Committee-based semi-supervised training has been proposed by Vu et al. [2011], in which first, a core and some complementary AMs are trained using the labelled data. Then using each one of those, a large unlabelled dataset is decoded by the ASR to produce multiple transcripts. The transcripts are lined up and the filtering is performed on the

segments with a certain level of agreements in the committee. Finally, the selected data and the original labelled data are combined to make a new AM. They trained a primary model (baseline) using the normal DNNs with filter bank features, two other DNNs with MFCC and PLP features, one filter bank-based Sigmoid-unit-type Recurrent Neural Network (SigRNN) and one filter bank-based LSTM. Agreement of the SigRNN and LSTM resulted in the best performance for their experiments.

4.4 State-of-the-art

The early ASRs, which appeared around 70 years ago, could barely recognise digits and/or a limited number of single words. Nowadays, the ASRs can recognise continuous speech with a relatively high accuracy rate and many applications have been introduced using the ASRs which can help us in everyday life (e.g. Amazon Alexa, Google OK). The ASR improvements have been achieved owing to the advances in memory, hardware, speech technology, machine learning, and more importantly the common datasets that the educational and the industrial companies have provided over time for researchers in this field of study.

There have been a number of well-known databases particularly dedicated to recognise continuous and spontaneous speech including

- **Texas Instruments-Massachusetts Institute of Technology (TIMIT)** [Garofolo, 1993]: is a reading style continuous speech corpus of American English with 630 speakers who uttered 10 sentences. Data was collected by a microphone with a 1-channel and 16000 sample rate (pcm). The dataset designed for acoustic-phonetic studies providing both phone and word level time alignments which allows phone recognition as well as normal word evaluation for ASRs.

- **Switchboard (SWB)** [Godfrey et al., 1992]: is a large scale multi-speaker corpus of American English conversational speech over telephone lines (2-channel ulaw with 8000 sample rate) with around 2500 conversations lasting between 3 and 10 mins. Hub5 2000 is an evaluation data set of 20 telephone conversations of SWB.

- **Call Home (CH)** [Alexandra et al., 1997]: consists of 120 telephone calls (2-channel mu-law, sample rate: 8000) with family and friends spoken by native American speakers each lasting about 30 mins.

- **Fisher (Fsh)** [Cieri et al., 2004]: is a collection of 5850 conversational telephone speech (2-channel ulaw with 8000 sample rate) with time-aligned transcript each around 10 mins. In contrast to SWB and CH the speakers were assigned to specific topics to talk about and some of the topics were similar to the topics in SWB conversations.

- **Wall Street Journal (WSJ)** [Consortium et al., 1994; Garofolo et al., 1993]: is a large scale read style speech corpus from the Wall Street Journal news. WSJ recorded by over 240 speakers with American English accent and 78000 utterances (73 hours). Data was recorded by close-talking head-mounted microphone as well as a secondary microphone (sampling rate: 16000, type: 1-channel pcm compressed). 4000 utterances were spontaneous speech uttered in dictation mode. Wall Street Journal Cambridge (WSJCam) is a similar dataset to WSJ recorded by 140 British speakers (92 for training set and 48 for testing set) [Fransen et al., 1994; Robinson et al., 1995].

- **International Computer Science Institute (ICSI)** [Janin et al., 2003]: comprises of 75 weekly meeting conversations at International Computer Science Institute in Berkeley, each speech between 17 and 103 mins. Speakers were wearing close-talking microphones (sample rate: 48000). Each meeting was between 3 and 10 (average of 6) speakers.

- **Augmented Multi-party Interaction (AMI) meeting corpus** [Renals et al., 2010]: with over 100 hours worth of interactions of multi-party meetings (in three different rooms) recorded by close-talking and far-field microphones (microphone array) as well as cameras. The speakers were mostly non-native English speakers. The audio wave files down-sampled from 48000 to 16000 sample rate.

- **Computational Hearing in Multi-source Environments (CHiME)** [Barker et al., 2013]: is a series of challenges dedicated to distance speech recognition operating in a robust and human-like environmental condition. In the recent challenge

(CHiME-5) [Barker et al., 2018], they provided over 40 hours of training data collected by 6 Kinect microphone arrays and 4 wearable binaural microphone pairs in 16 homes (2 extra homes for development and 2 more for evaluation purposes). They used a dinner party scenario with 4 speakers (2 acting as hosts and 2 as guests) for the challenge. Speakers, who were friends, moved around and talked to each other about different topics in kitchen, dining room and living room. The recording are available in single-array as well as multiple-array tracks (32 microphones per session, 6 microphone arrays with 4 microphones, plus 4 participants wearing 2 microphones, wave files sample rate: 16000).

The **word error rate (WER)** is the standard measure for evaluating the performance of an ASR, which can be written as [Holmes and Holmes, 2001]:

$$WER = 100 \times \frac{C(substitutions) + C(deletions) + C(insertions)}{N(reference)}\% \tag{4.27}$$

where N is the total number of words in the reference and $C(substitutions)$, $C(deletions)$, $C(insertions)$ are the number of substitutions (for wrong words), deletions (for missing words) and insertions (for extra words) errors in the hypothesis respectively. ASR accuracy rate can be defined as the number of words correctly recognised as a proportion of the total number of words. Note that the ASR accuracy and $100-WER$ are not always equal (WER could be higher than 100% when summing up the three errors together makes a bigger number than the total number of words in the reference).

In the following paragraphs, we will be giving an overview of the state-of-art performance of ASRs based on DNN techniques.

Hinton et al. [2006] proposed the Deep Belief Net (DBN) which is a single multilayer generative model obtained by combining a stack of RBMs. The DBN-DNN used for the Bing Mobile Voice Search (BMVS) application that has a high degree of acoustic variability (noise, music, accent, sloppy pronunciation, hesitation, repetition, etc). The initial results based on only using only 24 of training data has shown a 69.6% accuracy (5.6% improvement compared to the best HMM-GMM trained with the MPE criteria). They then extended the training data to 48 hours which led to a significant increase

in accuracy (71.7%). Applying the HMM-DNN recipe to other tasks such as the SWB (with over 300 hours training data), English Broadcast News (50 hours data), Google Voice input (5870 hours data) and Youtube (1400 hours data) all has resulted in a better performance comparing to the HMM-GMM recipes [Hinton et al., 2012]. Seide et al. [2011] also reported a significant improvement in recognition by combining the context-dependent DNN with HMMs. They could reduce the WER error for the SWB task from 40.9% to 27.5% in a single pass Speaker Independent (SI) model, and additionally, further improvement achieved by a multi-pass adaptive approach (dropping WER to 25.2%).

Recently, using RNNs for AM and combining n-grams with neural network LM Saon et al. [2015] reported 8.0 % WER for the SWB dataset evaluated by the Hub5 2000 evaluation set, while a group from Microsoft [Xiong et al., 2016] obtained 6.3 % WER for the same test set combining RNN LM with bidirectional LSTM (Bidirectional Long Short-Term Memory (BLSTM)) for AM.

Xiong et al. [2016] achieved 11.9 % WER for CH dataset, and IBM [Saon et al., 2015] 12.5 % WER.

In a recent work using RNNs, a WER of around 40 % has been reported for ICSI corpus [Enarvi and Kurimo, 2016].

Chen et al. [2014] expanded the Automatic Speech Attribute Transcription (ASAT) framework to spontaneous speech recognition, using the lattice rescoring approach for the SWB corpus. The ASAT consists of two key elements: (a) bank of attribute detectors with confidence scores, and (b) an evidence merger, which combines the low-level attribute scores into higher level evidence like the phoneme posterior. The output of the attribute detector is stacked together and made a supervector. They gained a slight improvement comparing to the baseline ASR for the SWB task particularly using the deep merger (around 3%).

Abdel-Hamid et al. [2014] used the Convolutional Neural Network (CNN) (originally developed for image recognition tasks) with a limited weight sharing scheme on the TIMIT dataset which has resulted in 6 to 10% better performance than the conventional DNNs. Graves et al. [2013] used LSTM for the TIMIT dataset and could get 17.7% Phone Error Rate (PER). In their recent work, combining the LSTM with CTC and integrat-

ing the output with a LM in the decoding phase, they could achieve an state-of-the-art accuracy of 6.7% for the WSJ dataset [Graves and Jaitly, 2014a].

Hori et al. [2016] obtained 22.6 % WER using the AMI dataset, while Swietojanski and Renais [2016] reported 26.2 % WER for this dataset.

Barker et al. [2013] reported that the baseline DNN-based ASR trained by Kaldi for CHiME-5 challenge, achieved 47.9% WER using the binatural microphone pairs, while using Kinect microphone array (with beamforming) they gained an 81.3% WER.

Due to the complexities of conversations between the neurologists and the patients in our study, the current commonly available ASRs (e.g. Google Go, Amazon Alexa) can not be used directly for automatic transcription (mostly they are good for recognition in reading style). Despite the recent advances in the speech recognition area in general, unfortunately at present, in medical applications there is not any appropriate database available to train our own ASR. As we will show in the next sections, the well-known databases on their own are not enough for training a good ASR and we need a considerable amount of data which is recorded in a similar conditions as the interviews between the neurologist, the patients and the APs.

4.5 Automated transcription

This section aims at describing the ASR component of the system. The ASR should deal with the challenges of spontaneous speech recognition which are listed at the beginning of this chapter. In addition, the quality of the recordings is not as good as the quality of common datasets available for speech community. They were recorded by a single normal microphone which was not located close to the speakers. Also, there were not any alignments for the transcripts. So we had to make the alignment manually ready for the ASR[1]

In order to automatically transcribe, we start with a baseline ASR trained using HMM-GMM approach for the AM and n-gram with KN/GT smoothing for LM of the

[1]Since giving the whole audio file of a conversation (e.g. 20 minutes) at once to the ASR to process, requires a considerable amount of memory, we have to first use the alignment to segment the audio file into smaller parts and then pass the smaller audio segments to the ASR to process.

ASR. Then we improve the acoustic and LMs to boost the performance of the ASR.

4.5.1 Baseline ASR

The first step towards automatic transcription is to use the manually produced segments for the conversations. In this chapter we only focus on the ASR part of our dementia detection system, assuming that the segmentation information is available for the ASR. This will allow us to see the effect of adding ASR to the system. In the next chapter, we will add the speaker diarisation unit of the dementia detection system to provide automatic segments and see the effect of diarisation and ASR together on the system.

Using the manually produced transcripts, speaker turn segmentation were prepared and since dealing with short length segments was a challenging task for ASRs, all the segments lasting less than 0.5 second and overlapping segments were removed from the data. The final data consisted of approximately 8 hours of spontaneous speech from 81 speakers (some doctors appeared in multiple interviews), with 6266 utterances with mean length of 4.6 seconds. In total, the data set comprised of 30 conversations with an average recordings time of 16 minutes.

The Kaldi toolkit [Povey et al., 2011] was used for speech recognition. For the LM, a set of 3- and 4-gram models were trained using KN or GT, and the model with the lowest perplexity was chosen. The dataset for training the LM came from the transcripts of the training set itself, i.e. in-domain LM (more details about the LM in Section 4.5.4 Table 4.3).

Table 4.1 shows the average WER and the standard deviation in brackets, for the baseline ASRs. We trained three baseline ASRs and as we go through the table from top to bottom, WER decreases considerably.

First, we trained a model (WSJCam), using the Kaldi's standard recipe on WSJCam dataset. Decoding on the model for Hallamshire (Hal) dataset resulted in 91.4% WER. Maximum A Posterior (MAP) adaptation of the WSJCam on the Hal dataset (WSJCam+MAP_Hal) could not improve the performance of the ASR significantly (86.3% WER). Therefore, another model trained using only the Hal dataset itself (Speaker Adaptation Training (SAT)_Hal). The 'leave-one-out' cross validation approach

was used for training the ASR. For this model, the HMM-GMM based AM, the training process included: (1) mono training using only 13-dimensional MFCC as acoustic input features, (2) delta training (adding deltas and delta-deltas to the input features, i.e. MFCC with 39 dimensions), (3) LDA_MLLT training (applying LDA and MLLT feature transformation), and (4) SAT. Following this approach, we obtained 69.5% average WER (13.4% standard deviation). As you can notice, the average WER is still high, which reflects the very challenging nature of the dataset. Since the amount of data was very limited (around 8 hours), following the DNN recipes on the dataset could not improve the results.

Table 4.1: *Baseline speech recognition results: the average WER with the standard deviation in brackets.*

Model	avg. WER (sd) [%]
WSJCam	91.4(6.1)
WSJCam+MAP_Hal	86.3(8.12)
SAT_Hal	**69.5(13.4)**

4.5.2 Adding additional data

A commonly used approach to increase the performance of ASRs is to add out of domain data, when there are not enough data to train a robust ASR. However, the out of domain data should be selected carefully. If there are considerable differences between the two datasets (e.g. recording conditions, the style of speech and the language used by the speakers), adding dataset may not result in a better performance.

We had access to around hundred hours of a new dataset, ("Seizure (Sez)" dataset), which was recorded in a similar way as the Hal dataset (i.e. interviews between neurologists and patients with or without seizure (epilepsy)). The Sez dataset consisted of 241 recordings between doctors and patients, each interview lasting between 20 to 40 mins. The interactions were transcribed manually for different CA studies. The transcripts, however, were not suitable to train ASRs, i.e. there was no speaker segmentation information available. Manual alignment of Sez dataset would take a considerable amount of

time and efforts. Therefore, we needed a way to select a subset of the data which can be easily segmented in an automatic approach.

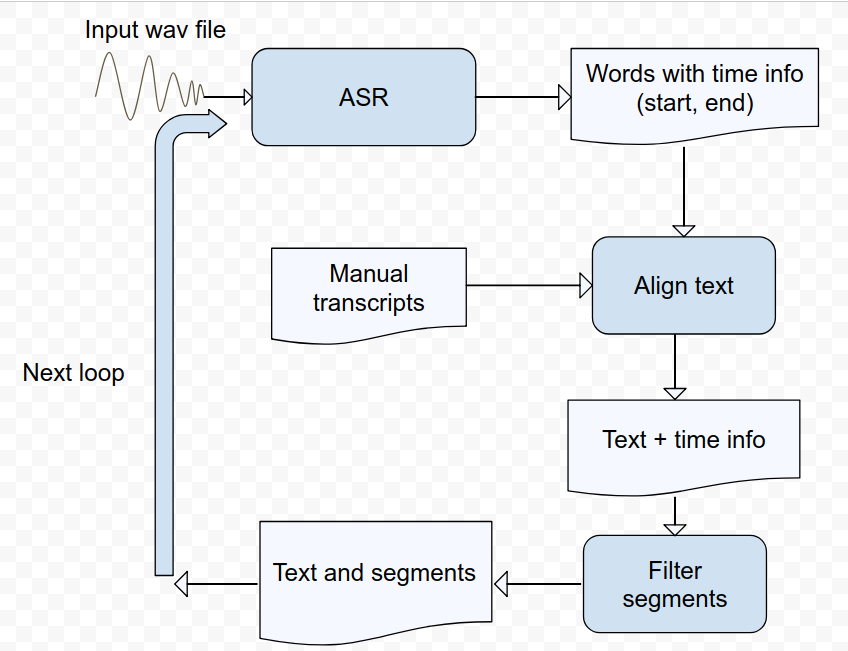


Figure 4.7: *Block diagram of the automatic segmentation system.*

We followed the conventional approach of semi-supervised training to produce automatic segmentation for the Sez dataset and mix the dataset with the Hal dataset to let us train a better ASR. The process of automatic segmentation, however, needed to be repeated a few times to find the best segmentation (the segmentation which achieves the lowest WER for the whole dataset).

Figure 4.7 shows the block diagram of the automatic segmentation system. First, the input wave files are given to the initial ASR (trained on the previous available dataset, i.e. Hal dataset) to generate the automatic transcripts with timing information (start and end time for each word). Then these transcripts are aligned with the manual transcripts to make segments. Since the performance of such ASR is not perfect, not all of the generated segments are reliable. Using a filtering algorithm we can choose the segments which are recognised with a high accuracy rate (we are confident about them). These segments, then, are joined to the previous dataset to train a new ASR with more training data. The whole process can be repeated. Each loop can generate more data which should carefully be selected/filtered to train a better ASR. This continuous until we reach an optimum

point[1] (no further improvement can be seen). Using this technique we managed to recruit 50 hours and 16 minutes data (28k utterance, 597 speakers and average utterance length of 6.3 seconds) out of the whole 100 hours of Sez data. Note that here, we only wanted to select a subset of the Sez dataset with the lowest possible errors in segmentation which then can be added to the Hal dataset to improve the ASR acoustic and language models.

4.5.3 Improving the AM

Mixing the Sez dataset with the Hal dataset allowed us to have enough data to follow a recent Kaldi recipe using LSTM-TDNNs [Cheng et al., 2017]. In this approach, first, we trained a SAT HMM model and then followed a deep neural network recipe for LSTMs and LSTMs on top of it. The input layer of the neural network consisted of 100 neurons for the input i-vectors[2] as well as 40 neurons for 40-dimensional high resolution MFCC input features. There were two layers of LSTMs with 800 neurons each after the input layer. The next layer was an LSTM layer with 800 neurons and 200 recurrent-project as well as 200 non-recurrent-project layers. The pattern of 2 layers of LSTMs + LSTM layer was repeated for three times, i.e. the total network consisted of 3 layers of LSTMs and 6 layers of LSTMs. The output layer is constructed from 4555 neurons for each possible context dependent phonemes.

Table 4.2: *Improved speech recognition results: the average WER with the standard deviation in brackets.*

Model	avg. WER (sd) [%]
SAT_Sez_Hal	54.8(13.7)
LSTM-TDNN_Sez_Hal	**40.9(12.9)**

Table 4.2 shows the effect of adding around 50 hours recordings from the Sez dataset to the Hal dataset (8 hours). Similar to the baseline ASR, we used the leave-one-out cross validation for training the ASRs and then calculated the average WER for the ASRs. Following the SAT training (SAT_Sez_Hal), we obtained WER 54.8%, which is a remarkable

[1]We tried using the segments with the highest confidence scores, but it was not as good performance as choosing the segments with the lowest WER.

[2]The intermediate vectors representing speaker characteristics, were originally introduced for speaker verification by Dehak et al. [2009]. These acoustic features have also been used for speech recognition.

drop of almost 15% comparing to SAT_Hal (Table 4.1). The WER for the LSTM-TDNN recipe (LSTM-TDNN_Sez_Hal), however, reached at a 40.9% WER (another additional 14% decrease).

4.5.4 Improving the LM

We further investigated the effect of improving the LM on the performance of the ASR. In order to demonstrate the effect of an ideal LM, we trained a LM using both the training and the testing transcripts together. We call this the 'oracle' LM, as it can be considered as an upper baseline (an upper bound golden standard).

To improve the LM, we trained n-gram LMs on Fsh [Cieri et al., 2004] and SWB [Godfrey et al., 1992] transcripts, and then interpolated them with the LM trained on the Hal data (using different weights, e.g. 0.2 for LM1 and 0.8 for LM2).

There are a number of metrics for evaluating an LM including: perplexity, vocabulary size, Out Of Vocabulary (OOV) and coverage.

Perplexity is the standard measure to evaluate a LM [Jurafsky and Martin, 2008]. Perplexity shows how well a model can predict probability distributions such as n-grams. For a test set of words, perplexity is the inverse probability of the test set words normalised by the number of words. Intuitively, perplexity can be seen as a branching factor, i.e. how many words an average the model could choose. The lower perplexity, the better model prediction.

Vocabulary size is the total number of words in the train set and **OOV** is the number of unknown words, i.e. not seen in the train set. Generally for a LM bigger vocabulary size and lower OOV are desirable.

Coverage of an n-gram shows the percentage of known n-grams in a test set. To calculate the coverage of n-grams for the LMs we used the equation from Wu and Matsumoto [2014], i.e. the number of unique n-grams in the test set which are seen in the train set as well, divided by the total number of unique n-grams in the test set. Obviously more coverage (i.e. close to 100%) indicate a better LM.

Table 4.3 lists the evaluating metrics for the six trained LMs. Since we followed the leave-one-out cross validation approach we trained 30 LMs, therefore all of the metrics

in the table are the average of the metrics.

The first LM, 'Hal_LM', is the original LM used for the SAT_Hal ASR (in-domain 3/4 gram LM using KN or GT smoothing) with a perplexity around 150. The vocabulary size was 4024 and OOV =74. The number of 3-grams was over 16k with a high coverage of 92.9% (i.e. only around 7% unknown 3-grams). However, the coverage for 54k 4-grams dropped to 70.7%. This means that 3-grams had a better coverage for this LM.

Interpolation with Fsh and SWB transcripts, 'Hal_Fsh/SWB_LM' caused a decline in the perplexity from 150 to 106 (a drop of 44). The interrelated LM consisted of many words (vocabulary size of 54k) which resulted in a lower OOV (11 vs. 74) and increased number of 3-grams (48k) 4-grams (606k). The coverage of 3-grams reached almost 100% and the coverage of 4-grams jumped to around 99%.

However, the lowest perplexity was gained by the Oracle (Orc) model ('Hal_Orc_LM') with 17.4 which used all of the train and test set data. So the OOV reached 0 and the coverage for both 3-grams and 4-gram increased to 100%.

Mixing the Sez dataset with the Hal dataset improved the perplexities for both in-domain and interpolated LMs, a drop of around in 45 perplexity comparing the 'Hal/Sez_LM' to 'Hal_LM' (around 105 vs. 150), and a 12 perplexity decrease comparing 'Hal/Sez_Fsh/SWB_LM' to 'Hal_Fsh/SWB_LM' (around 94 vs. 106), although, the perplexity for the oracle LM 'Hal/Sez_Orc_LM' slightly increased, in comparision to the 'Hal_Orc_LM', most likely due to enlarging the text of the training set. Obviously adding the Sez dataset increased the vocabulary size, number of 3/4 grams and coverage and decreased the OOV. The coverage of 3-grams and 4-grams increased comparing to the LMs without the Sez dataset. Therefore this mixing improved the performances of the LMs in terms of the metrics.

Using the LMs listed in Table 4.3, the average WER was recalculated and listed in Table 4.4. The results shows that improving the LM can reduce the average WER for the ASRs drastically. For the baseline ASR (SAT_Hal model), first three rows in the table, the LM interpolation ('Hal_Fsh/SWB_LM') decreased the WER by 1.2%, while the oracle LM, could improve the performance of the ASR considerably to 54%(14.5% improvement).

For the SAT_Sez_Hal ASR (rows 4 to 6), the LM interpolation resulted in 0.8% de-

Table 4.3: *Evaluating metrics for the six LMs. For training the LMs we used the leave-one-out cross validation approach (i.e. for each model 30 LMs). Orc: oracle, LM: LM, Fsh: Fisher dataset, SWB: Switchboard dataset. PPL: average perplexities, #vocab: vocabulary size, #OOV: average number of out of vocabulary words, #3-g: average number of 3-grams, 3-g cov.: average coverage of 3-grams, #4-g: average number of 4-grams 4-g cov.: average coverage of 4-grams.*

LM	PPL	#vocab	OOV	#3-g	3-g cov.	#4-g	4-g cov.
Hal_LM	149.8	4024	74	16365	92.9%	54426	70.7%
Hal_Fsh/SWB_LM	106.1	53907	11	48046	99.9%	606547	98.6%
Hal_Orc_LM	**17.4**	4098	0	16563	100%	55617	100%
Hal/Sez_LM	105.3	11667	26	29754	99.0%	188952	91.4%
Hal/Sez_Fsh/SWB_LM	94.1	55643	8	48325	99.9%	616725	98.8%
Hal/Sez_Orc_LM	18.1	11693	0	29780	100%	189292	100%

crease in WER, and the improvement from using the oracle LM was approximately 20% (comparing 34.9% WER vs. 54.8%).

Similarly, for the DNN based ASR ('LSTM-TDNN_Sez_Hal') (the last three rows in the table), the interpolation improved the WER from 40.9% to 40.0%, and using the oracle LM 'Hal/Sez_Orc_LM' we achieved a WER 25.8%. Therefore, improving the LM can reduce the WER considerably, especially if we improve the coverage of n-grams and reduce OOV.

Finally, we focused on training RNN based LMs, which are the current state-of-the-art for training a robust LM for ASRs. We managed to train 15 RNN-based LMs (out of the total of 30 for the leave-one-out cross validation). For those models, we obtained around 0.6% further WER improvement (i.e. an estimation of 39.4% WER comparing to 40.0% WER for the Hal/Sez_Fsh/SWB_LM). However, due to the computational and memory limitations that we have had in our university, we preferred to choose the Hal/Sez_Fsh/SWB_LM for the next experiments in this project.

4.5.5 WER per speaker group

The Hal dataset consists of conversations between doctors, patients and accompanying person(s). The doctors mostly talked with a clear and more formal language in the inter-

Table 4.4: *Speech recognition results for different LMs.*

Model	Language Model	WER(std) [%]
SAT_Hal	Hal_LM	69.5(13.4)
SAT_Hal	Hal_Fsh/SWB_LM	68.3(13.3)
SAT_Hal	Hal_Orc_LM	54.0(16.6)
SAT_Sez_Hal	Hal/Sez_LM	54.8(13.7)
SAT_Sez_Hal	Hal/Sez_Fsh/SWB_LM	54.0(13.8)
SAT_Sez_Hal	Hal/Sez_Orc_LM	34.9(16.2)
LSTM-TDNN_Sez_Hal	Hal/Sez_LM	40.9(12.9)
LSTM-TDNN_Sez_Hal	Hal/Sez_Fsh/SWB_LM	40.0(13.0)
LSTM-TDNN_Sez_Hal	Hal/Sez_Orc_LM	**25.8(13.1)**

actions, comparing to the patients and the accompanying persons. Since we are interested in the whole conversation (not only the patient's segments), as a further analysis, we investigate the WER per different speaker type in a different patient group (ND/FMD).

Figure 4.8 shows the WER for the LSTM-TDNN_Sez/Hal ASR with Hal/Sez_Fsh/SWB_LM LM, which was split per different speaker (patient, neurologist and accompanying person(s)) in different patient group (ND vs. FMD). On the whole the words uttered by the speakers in the FMD patient group were recognised better than those in the ND group (38% WER vs. 44%), although the neurologists in the ND group had a slightly better WER (34% vs. 33%).

The WER for the patients in the ND group was the highest amongst the different speakers in the different patient group with a 51% WER (considering the error bar, we had patients in the ND group with over 65% WER), while the patients in the FMD group had 39% WER (12% less). The second worse WER was for the accompanying person(s) in the ND group with 48% WER (the corresponding speakers in the FMD group had 41% WER).

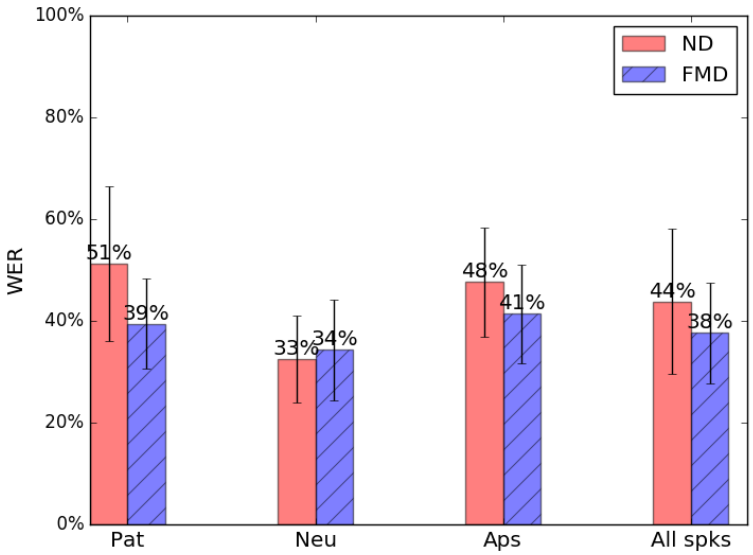


Figure 4.8: *WER per speaker for different patient groups: ND, FMD. Pat=patient, Neu=neurologist, Aps=accompanying person(s), All spks=all speakers.*

4.6 Discussion

The aim of this chapter was to find an answer to the second research question, what types of speech technologies are needed for developing dementia detection system. The conversations between the doctors and the patients in our main dataset, the Hal dataset, were spontaneous and recorded in a real hospital with normal microphones. It is not applicable to use the current commercial ASR systems (e.g. 'OK Google' and 'Siri') to recognise these conversations(we need to train our own ASR). They are trained mostly with read style and command mode speech for individual speakers, and therefore not a good match for the speaking style of our data. Besides in conversations there are other complexities such as turn taking and overlapping speech which can not be handled well by these ASRs. A diarisation unit is needed to segment the audio file into smaller parts and pass the audio of a single speaker to the ASR. There are a number of commonly used datasets dedicated to spontaneous speech recognition. However, they are either recorded using telephony channels (e.g. SWB and CH) or with good quality microphones and/or arrays of distance microphones (e.g. AMI, CHiME-5).

In terms of style of the conversation, the Hal dataset is not like a broadcast TV show

and neither a meeting. TV shows normally are recorded by high quality microphones and people mostly talk clearly and speak formally. In a meeting, which is normally recorded by arrays of high quality microphones, there are many speakers, and depending on the topic, people also try to talk more formal (e.g. academic speaking). The main speaker may ask some questions from the audience and he/she controls the flow of conversation.

To train an ASR for our dementia detection system, we started with training a few ASRs on the available datasets. Although not mentioned inside the chapter, the SWB dataset was the first dataset we choose to train, however, due to the its differences with the Hal datasets (i.e. in recording channels (telephone vs. normal microphone) and the English accents (American vs. Yorkshire British), topics, etc), we obtained a high WER around 95% for the ASR. Similar results gained from the AMI and WSJCam datasets with around 91% WER for both datasets (see WER for WSJCam in Table 4.1). MAP adaptation could not improve the performance of the ASR, therefore, we started to use parts of the Hal dataset itself for training ASRs, i.e. n-1 recordings to train the ASR and 1 to test (leave-one-out cross validation). Since the amount of data for training was not enough to train a DNN ASR, we added out of domain data from the Sez dataset to boost the acoustic and LMs.

Despite the improvements that we gained by following different approaches, the WER still seems high, which shows the challenging nature of the Hal dataset. Training a decent LM was essential in improving the performance of the ASRs. We trained LMs with lower perplexities and OOV and higher n-gram coverage. Interpolating the LM with Fsh and SWB datasets, improved the perplexity and n-gram coverage of the LM and in turn reduced the WER of the ASR. However, mixing with the Sez dataset improved the LM and ASR considerably. We also showed that having oracle LM can boost the ASR and reduce the WER from 40% to 25.8%, see Table 4.4). The oracle LM (the perfect LM) had 0 OOV and 100% coverage (nothing unknown) which resulted in the best result.

In terms of speakers group we found out that generally, the conversations in the FMD group had lower WER than those in the ND group and the neurologists had the lowest WER amongst the speakers (around 33%) in both the FMD and the ND groups. However, patients and the APs in the ND group had the worst WER. The neurologists used far

more clear and formal language and patients in the FMD group could talk more clearly since they had fewer conversational issues which let them to provide longer and more natural responses. While the patients in the ND group were struggling to provide clear responses. There were lots of gaps in their talk and they relayed on the APs to complete their answers to the asked questions.

In this chapter we trained ASR unit of our dementia detection system focussing on reducing the WER. Despite having a relatively high WER (40%) for the ASR, in the following chapters (Chapter 6 and 8) we will investigate how these errors can affect the classifier results (i.e. compared to the features extracted from the manual transcripts). Also we will look at different types of features (acoustic, lexical, etc.) which can be extracted automatically from the audio files which can be affected by the ASR errors.

It is worth mentioning that due to having a high number of experiments needed for training the ASRs (30 ASRs, 4 models, i.e. 30 * 4 = 120 LMs) and limitation that we have had in using university's computational resources e.g. CPU/GPUs and memories, we can not claim that we trained the best possible ASRs with the best possible tuning techniques (i.e. tuning the parameters for one ASR is easy, but finding the best parameters for 30 ASRs is not).

4.7 Summary

Spontaneous speech recognition is a challenging task due to its unstructured and unpredictable nature and a significant mismatch between training and testing conditions. In addition, issues such as disfluency (filled pauses, repair, false start, etc.) and overlap in conversation should be dealt by ASR. The conventional ASR architecture consists of a front-end processing to extract acoustic features from the input stream, an AM to built decent representatives for different phonemes (or tri-phones), a LM to model the word order in general utterances, and a decoding unit to use the combination of the acoustic and LMs for the unseen (test) acoustic stream and find the best match text output. High speed CPUs and GPUs, recently, has enabled us to apply deep neural network approaches for speech recognition task. Commonly DNNs can be used in hybrid mode for AM, e.g. HMM + LSTM-TDNN.

For the ASR unit of our dementia detection system, we firstly trained an HMM model using SAT training of Kaldi toolkit. Since we had a limited amount of data (Hal dataset with total recording time of 8 hours), we had difficulties to apply the DNN recipes to train ASR. DNN recipes need hundreds or thousands hours of data to train a robust ASR. We had given extra data (Sez) recorded in similar conditions as our dataset, but without segmentation. Using a conventional approach similar to the semi-supervised training we could select a subset of Sez dataset, around 50 hours to mix with the Hal dataset.

Combining the extra dataset allowed us to followed the recent Kaldi recipe of LSTM-TDNNs. Comparing to the SAT model, LSTM-TDNN based ASR performed with a considerably lower WER (42.7% vs. 55.8%).

As a further improvement we investigated boosting the LM using interpolating with another dataset (Fsh/SWB) as well as applying RNN based LM. These improvements slightly (2-3%) reduced the ASR performance, however, the oracle LM could remarkably decrease the WER (up to 25.8%). This indicates the challenging unstructured behaviour of the spontaneous speech. Despite having a high WER , in the next chapters we will investigate whether replacing the outputs of the ASR with manual transcripts of the conversations can still be useful for dementia detection.

Chapter 5

Speaker diarisation

Contents

5.1 Diarisation 101

5.1.1 Diarisation architecture 102

5.1.2 Diarisation toolkits 104

5.1.3 State-of-the-art 105

5.2 Baseline diarisation for dementia detection 110

5.2.1 Effect of overlapping speech and within-turn gaps 111

5.2.2 Word diarisation error rate 113

5.3 I-vector based diarisation 114

5.4 Discussion 115

5.5 Summary 118

This chapter aims at investigating an answer to the second research question, that is what type of speech technologies needed for dementia detection system through analysis of conversation, focusing on the speaker diarisation unit of the system. A Speaker diarisation unit determines who talks when in a conversation. It splits the audio into segments (showing start and end times of each segment) and specifies which speaker each segment belongs to by assigning the segments to a number of speaker ids in the conversation. Due to the challenging nature of the recordings of our major dataset, the Hal dataset, it is essential that we can train a (or use an existing) diarisation toolkit with the best diarisation performance. We also need to nominate a classifier as our final classifier in the dementia detection system.

Section 5.1 introduces the speaker diarisation systems and the general components of the system.

In **Section 5.2** we discuss about a few baseline diarisation units for our dementia detection system and find the best baseline unit.

In **Section 5.3** we train i-vector based diarisation units and compare their performances with the best baseline speaker diarisation unit.

Section 5.4 includes the discussion part of this chapter.

Section 5.5 summarises the chapter.

5.1 Diarisation

Audio diarisation is the task of labelling and categorising audio sources within a spoken document [Tranter and Reynolds, 2006]. Speaker diarisation refers to the task of identifying "who spoke when?" in an audio (and/or video) recording. The number of speakers and the amount of the spoken audio are normally unknown. Thus diarisation can be considered as unsupervised identification of speakers and their intervals (segments) of speech within an audio stream [Miro et al., 2012; Moattar and Homayounpour, 2012; Tranter and Reynolds, 2006].

Speaker diarisation can be used for applications of audio/video processing and information retrieval ranging from telephone conversations, broadcast news, Total Variability (TV) shows, and movies to conference, lectures and meetings. A diarisation challenge was initially introduced by the National Institute of Standard and Technology (NIST). The Rich Transcription (RT) and subsequent Rich Meeting Transcription (RMT) projects were dedicated to speaker diarisation (RT02, ..., RT07 and RT09[1]).

There are three primary domains for using speaker diarisation: broadcast news, recorded meetings and telephone conversations [Tranter and Reynolds, 2006]. The nature of data for these three domains are different. For instance, in telephone conversations, the recording channel and environment is typically different for each recording and there are normally 2 or 3 speakers each one using a different microphone [Moattar and Homayounpour, 2012]. Broadcast news normally are recorded by good quality microphones and/or cameras while in contrast, meetings are recorded by a single or an array of far-field microphones. Thus the quality of meeting data is somewhat poorer. The broadcast news is often recorded in read mode and normally with more speakers (e.g. hosts, reporters, audience) whilst the meeting by nature is more spontaneous with fewer speakers and normally fewer utterance turns per speaker [Miro et al., 2012].

[1] http://www.itl.nist.gov/iad/mig/tests/rt

5.1.1 Diarisation architecture

Most of the speaker diarisation systems are developed based on two major approaches: the bottom-up and the top-down. As illustrated in Figure 5.1 (a), the top-down method starts with one cluster (or a few clusters) for the speakers and eventually adds more clusters as the process progress, while the bottom-up approach starts with more clusters than expected speakers (usually single cluster for each segment) and gradually merge the clusters together to reach an optimum number of clusters. If the final number of clusters is smaller than the optimum number, the process is called under-clustering; when the number is more than the optimum, it is known as over-clustering. Both approaches are generally modelled by HMMs with states represented by GMMs corresponding to each speaker, and transitions corresponding to the speaker's turns.

The bottom-up approach (known as Agglomerative Hierarchical Clustering (AHC) is the most common approach for diarisation, which trains a number of speaker clusters and then successively merges the clusters until only one cluster remains per speaker. Initialisation can be carried out in different ways such as k-means and uniform initialisation. Initially the audio stream is over segmented, exceeding the maximum number of speakers. Then iteratively, those are closely matching are identified and merged together, i.e. each iteration results in one reduction. Each cluster is modelled by a GMM and as a new cluster is merged, the two previous GMMs of their clusters are used to train a new GMM. Standard distance metrics are used to identify the close clusters. After each merging, normally a reassignment of the frames to the new clusters (e.g. via Viterbi realignment) is needed, and the whole process is repeated until a stopping criterion is reached (see Figure 5.1 (b)).

A typical diarisation system may include the following components [Le et al., 2007; Miro et al., 2012; Moattar and Homayounpour, 2012]: speech activity detection, speaker change detection/speaker segmentation, speaker clustering and re-segmentation.

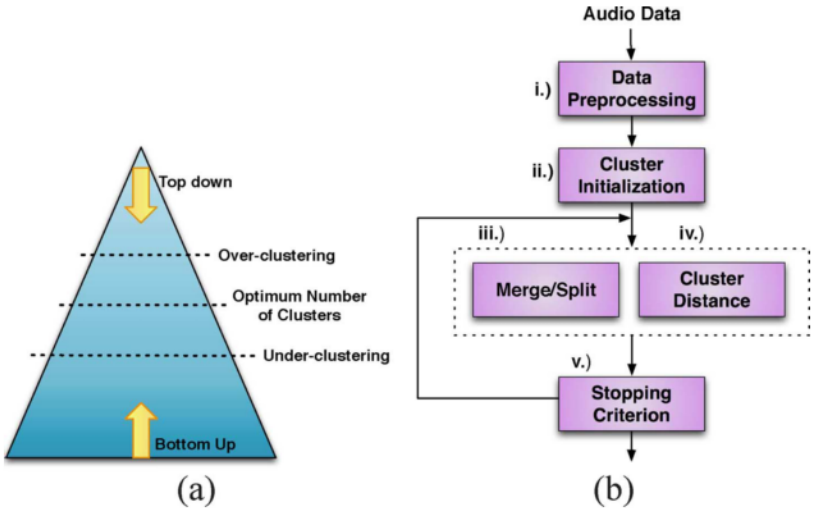


Figure 5.1: *General diarisation system. (a) Alternative clustering schemes. (b) General speaker diarisation architecture [Miro et al., 2012].*

5.1.1.1 Speech activity detection

An important initial step for speaker diarisation is Speech Activity Detection (SAD)[1]. The SAD labels the segments as speech or non-speech and directly affects the diarisation error. Poor SAD can significantly decrease the diarisation performance. The SAD process can be performed separately and before diarisation. Some SAD approaches use feature extraction, energy based thresholds and pitch estimation. However, model-based approaches often perform better [Le et al., 2007]. In these approaches, models are usually trained for speech and non-speech separately. Drawbacks of this approach become apparent when there are significant differences between the training and the testing conditions. Hybrid approaches have been proposed in which, first, energy based decisions are made and then, the models improve the decisions [Le et al., 2007].

5.1.1.2 Speaker change detection or speaker segmentation

Speaker segmentation is the core step for diarisation, which attempts to split the audio stream into 'speaker homogeneous' segments, or it can be reviewed as detecting changes in speakers or turns. Speaker change detection is normally performed by comparing the

[1]sometimes it is called VAD

acoustic segments in two consecutive sliding windows [Le et al., 2007; Miro et al., 2012; Moattar and Homayounpour, 2012]. In order to compare the similarity/dissimilarity between two segments a number of distance measures have been proposed including relative cross entropy or the KLD [Siegler et al., 1997], Bayesian Information Criterion (BIC) [Chen and Gopalakrishnan, 1998], Cross Likelihood Ratio (CLR), Normalised Cross Likelihood Ratio (NCLR), Log Likelihood Ratio (LLR) [Reynolds, 1995], Information Change Rate (ICR), and Probabilistic Linear Discriminant Analysis (PLDA) [Prince and Elder, 2007].

5.1.1.3 Speaker clustering

While the segmentation step works on the adjacent windows to determine whether they belong to the same speaker, clustering attempts to identify and merge together the same speaker segments anywhere in the audio stream. Measuring similarity is performed by the same measures used for the segmentation.

5.1.1.4 Re-segmentation

The numbers of change points are often higher than the real number of speaker changes due to the high level of false alarms of speaker clustering errors. Therefore, it is often necessary to realign the adjacent segments of the same speakers together or perform re-segmentation. So as a further refinement for diarisation, another round of segmentation is needed between two neighbouring segments (to make sure they belong to two different speakers or they belong to a single speaker). Generally for realignment, the Viterbi decoding is applied in which the audio stream is re-segmented based on the current clustering prior retaining on the new segmentation [Le et al., 2007; Miro et al., 2012].

5.1.2 Diarisation toolkits

There are a number of diarisation toolkits which are designed particularly for research purposes. The following is a number of the current common open-source diarisation toolkits:

- **CMUseg** Siegler et al. [1997]: is an open source C and C++ toolkit developed at Carnegie Mellon University (CMU). The KLD distance is used as a distance measure for segmentation and clustering is based on a simple agglomerative technique.

- **Speech Recognition Research at the University of Twente (SHoUT)**[1] Huijbregts [2008]: is a C++ toolkit which has been developed by Huijbregts as a part of his PhD. It uses agglomerative model based diarisation applying the BIC distance measure for segmentation and clustering criterion.

- **Laboratoire d'Informatique de l'Universit du Mans (LIUM)**[2] [Meignier and Merlin, 2010]: is a Java based toolkit developed by the researchers from University of Maine (France) which uses DNN by distance measures such as BIC, CLR and NCLR.

- **AudioSeg**[3] Gravier et al. [2010]: is a C based toolkit which provides different types of segmentation and speaker clustering such as silence/audio activity detection, segmentation using BIC, LLR or KLD distances, GMM/HMM based clustering and Viterbi segmentation.

- **Diarisation Toolkit (DiarTK)**[4] Vijayasenan and Valente [2012]: is a C++ based toolkit developed in Idiap research institute which uses a non-parametric clustering and realignment based on a technique known as agglomerative information bottleneck and avoids explicit GMM speaker modelling.

- **Kaldi diarisation toolkit**[5]: extracts i-vectors from the input recordings and uses PLDA as a scoring metric. The AHC is used for clustering and a threshold learned from the data is used to stop clustering.

5.1.3 State-of-the-art

The most commonly used corpora for diarisation are those introduced by the NIST Speaker Recognition Evaluation (SRE) (e.g. [Greenberg et al., 2014, 2013; Martin and

[1] http://shout-toolkit.sourceforge.net/
[2] http://www-lium.univ-lemans.fr/diarisation/doku.php/welcome
[3] https://gforge.inria.fr/projects/audioseg
[4] http://www.idiap.ch/scientific-research/resources/speaker-diarisation-toolkit
[5] https://github.com/kaldi-asr/kaldi/tree/master/egs/callhome_diarization/v1/diarization

Greenberg, 2009, 2010a,b; Sadjadi et al., 2017]). In addition there have been a few challenges dedicated to diarisation such as Multi-Genre Broadcast (MGB) Bell et al. [2015a], and Arabic acMGB-2 and 3 [Ali et al., 2016, 2017]

The standard diarisation error metric is Diarisation Error Rate (DER) defined by RT evaluations (NIST Fall Rich Transcription on meetings 2006 Evaluation Plan 2006) as:

$$\text{DER} = \frac{\sum_{s=1}^{S} dur(s)(max(N_{ref}(s), N_{hyp}(s)) - N_{correct}(s))}{\sum_{s=1}^{S} dur(s) N_{ref}(s)} \tag{5.1}$$

where S is the total number of speaker segments in which the reference and hypothesis speaker pairs are the same, $N_{ref}(s)$ and $N_{hyp}(s)$ indicate the number of speakers in the reference and the hypothesis respectively, $N_{correct}(s)$ is the number of speakers that correctly matched between the reference and the hypothesis, and $dur(s)$ is the segment length [Huijbregts, 2008; Miro, 2006].

The DER consists of three different sources of errors: missed speech, E_{MISS} (from the segments which exist in the reference but are missed in the hypothesis), false alarm speech, E_{FA} (from the segments which exist in the hypothesis but not in the reference), and speaker error E_{SPKR}.

$$\text{DER} = E_{SPKR} + E_{FA} + E_{MISS} \tag{5.2}$$

The speaker error E_{SPKR} itself consists of two errors: the number of incorrectly assigned speakers and the speaker overlap error. The overlap error is the most significant source of errors [Huijbregts et al., 2012; Miro et al., 2012; Yella and Valente, 2012] for diarisation tasks. In particular, for recorded meetings, where detecting and treating the overlaps reliably remains unsolved problem [Miro et al., 2012]. However, overlapping speech is inevitable in a natural conversation. A significant amount of our spontaneous conversations include overlapping speech. It generally occurs when speakers try/compete to take a turn or show their agreement or disagreement with the current speaker by back channelling. Therefore it is desirable to find efficient ways to deal with this issue.

A number of studies focused on dealing with overlapping speech. Otterson and Ostendorf [2007] showed that assigning the overlapping speech according to the labels of two

nearest neighbouring segments can considerably reduce the diarisation error and removing the overlapping segments directly from the input stream does not necessarily improve diarisation, however,"robust speaker clustering" may perform better than removing overlaps.

Boakye et al. [2008] proposed an HMM-based overlap segmenter using different kinds of features (MFCC features, Root Mean Squared (RMS) energy and Linear Predictive Coding (LPC) coefficient) and applying Diarisation Posterior Entropy (DPE) measure based on frame-level speaker likelihoods. After detecting the overlaps, the system modifies the segments using a post-processing component.

Motivated by CA studies, Yella and Bourlard [2014] proposed an acoustic based overlap detector using long-term conversational features such as silence/speech statistics captured in windows around 3 to 4 sec. These features are used for estimating the probabilities of the overlap and the single-speech speaker categories. The main hypothesis of the work is that the most silent segments are less overlapped. Testing their approach on three meeting datasets: AMI [Renals et al., 2010], NIST RT09[1] and ICSI [Janin et al., 2003], they could improve DER considerably (AMI: from 30.4% to 24.2%, NIST RT09: from 33.9% to 31.5%, and ICSI:from 33.3% to 30.9%).

Some works on diarisation were based on the concept of 'i-vector' which was originally introduced for speaker verification by Dehak et al. [2011]. Using the Joint Factor Analysis (JFA) [Kenny et al., 2007] as a feature extractor, they trained a model containing both the channel and the speaker information. The factor analysis defined a new low-dimensional space (opposing the GMM based JFA with high-dimensional super-vectors). In this space, each utterance was represented by an intermediate vector or i-vector.

Sell and Garcia-Romero [2014] first applied an unsupervised segmentation (1-2 second length with 0.5 second overlapping with the preceding and following segments) on the input audio stream. Then extracted i-vectors for the segments. The dimension of the i-vectors was further reduced by a conversation-dependent PCA. PLDA was used as a score metric to determine whether a pair of segments belong to the same speaker. The AHC was used for clustering and a threshold learned from the unlabelled data used as the stopping

[1]http://www.itl.nist.gov/iad/mig/tests/rt/

criterion.

Garcia-Romero et al. [2017] used a DNN and replaced the two-step generative processes of extracting i-vectors and learning the PLDA scoring function with an efficient single embedding while learning a scoring metric to discriminate between pairs of embeddings. Using PCA, they projected the embeddings into a conversation-dependent space adapting the PLDA to the conversation. The conventional AHC was used for clustering the segments and a threshold based approach was applied to stop the clustering. Furthermore, Variational Bayes (VB) re-segmentation [Sell and Garcia-Romero, 2015] refined the borders. They trained a DNN using 10K utterances taken from Fisher English [Cieri et al., 2004], NIST SRE04 [Martin and Przybocki, 2004], NIST SRE05 [Martin et al., 2005], NIST SRE06 [Przybocki et al., 2006], NIST SRE08 [Martin and Greenberg, 2009] datasets and used the utterances to learn PLDA scoring function. The CH corpus [Alexandra et al., 1997], with 500 recordings, was used for the evaluation. They gained 12.8% DER knowing the number of speakers in advance and then using the VB re-segmentation reduced the error to 9.9% (comparing to the normal i-vector approached with 13.6% and 11.2% respectively). While using the oracle threshold, they reached at DER of 12.6% without VB re-segmentation and 10.3% with VB re-segmentation.

Recently, for the inaugural DIHARD Challenge[1], Sell et al. [2018] investigated replacing i-vectors with x-vectors (introduced by Snyder et al. [2016]). The models using x-vector performed considerably better for both two tasks of Track1 (25.94% DER vs. 28.06% DER) and Track2 (39.43% DER vs. 40.42% DER). Further improvements by VB re-segmentation and fusion resulted in their best performance (for Track1 23.99% DER , for Track2 37.19% DER).

In recent studies, linking speakers across different recordings has been shown to improve performance for diarisation tasks. The main idea is to use the previous recordings of the same speakers to improve the clustering accuracy for the current recording, which ultimately results in better accuracy for the overall diarisation process. Techniques such as JFA, an extension to the TV parametric adaptation technique using the GMMs, are normally applied for the speaker linking task. Ferras and Bourlard [2016] used two differ-

[1]https://coml.lscp.ens.fr/dihard/

ent diarisation techniques (an Information Bottleneck (IB) diarisation, a discriminative approach, and an HMM-GMM based approach, generative method) in parallel followed by a fusion step (combining two diarisation results according to the maximum vote between the two) prior to the speaker linking. They reported a relative improvement of 7% using the fusion approach over their test dataset and also at least 25% relative gain in the JFA technique in comparison to the TV.

Milner and Hain [2016]; Milner et al. [2015] trained a DNN based SAD model and a DNN based speaker segmentation stage followed by the speaker linking using the BIC criteria for the Sheffield MGB[1]) challenge. They obtained the final DER of 57.2% for the linked speaker diarisation, and 50.1% without linking. The University of Cambridge team [Wang et al., 2016], however, have achieved DER rates of 47.5% and 40.2% for the linked and unlinked diarisation respectively. They have developed a DNN-HMM hybrid approach for segmentation which is used for all the four evaluation tasks of the challenge. Using a 40-dimensional filter bank with PLP encoding first they used a Viterbi decoding for the DNN-based SAD task followed by a divergence based Speaker Change-point Detection (SCD) [Wang et al., 2016].

Sinclair and King [2013] focused on addressing the roots of various challenges for diarisation over the three recent RT evaluation datasets (RT05, RT07 and RT09) and using a number of oracles (ideal assumptions, models). They have found out that having the oracle number of speakers (i.e. knowing the real number of speakers in advance) only gives a little improvement in DER in their diarisation system and it often slightly reduces the accuracy. The oracle SAD model can eliminate both the missing speech and false alarm speech errors but at the price of introducing more speaker errors, and arguably the authors have concluded that the performance of a diarisation system is not highly dependent on the SAD output. They also tried 'ideal cluster initialisation' rather than the random initialisation, and found it can be effective for meeting conditions, yet not useful for the end-to-end conversations. The ideal segmentation/clustering models (e.g. in terms of approach itself, the number of Gaussians), however, reduces the E_{SPKR}, significantly. More importantly, they have reported that ignoring the overlap did cost 19.11%

[1]The challenge included four different tasks: speech-to-text transcription, alignment, longitudinal speech-to-text transcription, and longitudinal speaker diarisation and linking [Bell et al., 2015a]

E_{MISS} reduction in performance, however, assigning the overlaps to one of the speakers can halve this cost, and further assigning to the second speaker resulted in another error decrease by half.

Zajıc et al. [2017] trained a CNN as a regressor (i.e. producing values between 0 and 1; 0 = no change and 1 = a speaker change) for SCD. A conversation is split into short segments which are represented by i-vectors. Then the output of the trained CNN is used to refine the statistics gained for the segments, resulting in improved performance. To train the i-vectors, they used the NIST SRE 2004, 2005 and 2006 corpora combined with the SWB-1 release2 and the SWB-2 phase 3. They used the English part of the CH corpus covering only two speaker conversations (109 telephone conversation: 35 conversations for training CNNs and the rest for testing). They gained 7.8% DER for their suggested system while the baseline with constant length windowing segmentation obtained DER 9.2%.

5.2 Baseline diarisation for dementia detection

In order to perform the speaker diarisation, initially two common diarisation toolkits were used: SHoUT [Huijbregts, 2008] and LIUM [Meignier and Merlin, 2010]. SHoUT is a toolkit which is designed especially for meeting speaker diarisation. LIUM, on the other hand, is designed for performing speaker diarisation on broadcast news and telephone conversation applications.

The audio files of the interviews were passed to the two diarisation toolkits. Table 5.1 shows the components of the diarisation errors when using the two different systems. As can be seen, the DER using the SHoUT system is marginally lower (around 5%) than the LIUM, 45.7% compared to 49.5%. The speaker error for the LIUM with 20.2%, was considerably more that of the SHoUT system with around 13.5%. However, the missing speech error in the SHoUT diarisation system was approximately 3% greater than the LIUM system and comprised the most significant portion of the DER. The speaker false alarm error, however, did not make a major contribution to the DER of either diarisation system. Since the SHoUT diarisation system outperformed the LIUM, the SHoUT toolkit was used in the subsequent experiments. Note that the conditions of our recordings were

much closer to the meeting conversations than the broadcast/telephone ones for which LIUM was originally designed.

Table 5.1: *Diarisation error (consisting of the missing speaker error: E_{MISS}, false alarm error: E_{FA}, and speaker error: E_{SPKR}) for the Hal30 data using the SHoUT and LIUM toolkits.*

Diarisation system	E_{MISS}	E_{FA}	E_{SPKR}	DER
SHoUT	32.2%	**0.0%**	**13.5%**	**45.7%**
LIUM	**29.3%**	**0.0%**	20.2%	49.5%

Because our dataset is different to the standard datasets in terms of speaking style,it is difficult to evaluate the performance of our diarisation module. For instance, as mentioned before DER for AMI was around 24% [Yella and Bourlard, 2014] and for DIHARD challenge [Sell et al., 2018] Track1 was 24% for Track2 and 37%.

5.2.1 Effect of overlapping speech and within-turn gaps

The speaker diarisation of conversations with a low level of overlapping speech normally produces fewer speaker errors than that of conversations with more overlapping segments. There are different strategies for dealing with overlapping speech. Overlapping speech either can be left in for the system to handle, or it can be detected and then totally or partially removed from the audio files, after which the diarisation is applied only to the non-overlapping speech.

The outputs of diarisation systems consist of a number of gaps within the same speaker's segments (we refer to this as the 'within-turn gaps'). These types of gaps detected by the diarisation systems are either real gaps (or jingle, noise, or a very short utterance of the other speakers), or they have been detected by mistake. Inspection revealed that there were frequent within-turn gaps in our data, and that it should be possible to detect these gaps and remove them from the output of the diarisation system. Figure 5.2 displays the automatic process of removing the within-turn gaps and merging the short segments of speech by the speaker 1.

Removing the within-turn gaps, however, would result in producing very long segments for the speakers, which in turn enforces more memory allocation and increased

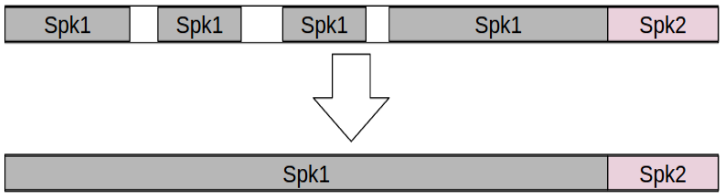


Figure 5.2: *Example of removing the within-turn gaps. After eliminating the gaps (here, three gaps), four shorter segments of speaker 1 are merged together to form a longer segment.*

computational load for the ASR. Thus in order to solve this issue, after removing the gaps and joining the segments together, the new long segments (of > than 20 second length) were divided into a number of smaller segments.

Table 5.2: *DER for Hal30 data using the SHoUT toolkit after removal of overlapping segments (no-overlaps: NoV), within-turn gaps (no-gaps: NoG, and both (no-overlaps/no-gaps: NoV_NoG).*

Diarisation system	E_{MISS}	E_{FA}	E_{SPKR}	DER
SHoUT_NoV	38.6%	**0.0%**	**11.3%**	49.9%
SHoUT_NoG	**2.4%**	**0.0%**	22.3%	**24.7%**
SHoUT_NoV_NoG	3.4%	**0.0%**	23.4%	26.8%

Table 5.2 shows the DER for the Hal30 after removal of overlapping speech (SHoUT_NoV), the within-turn gaps (SHoUT_NoG) and both (SHoUT_NoV_NoG). Having eliminated the overlapping segments (comparing the SHoUT_NoV to the SHOUT in Table 5.1), the DER increased slightly (around 4%). Speaker error reduced by 3%, however, the missing speaker error were increased by approximately 7% while the false alarm remained the same.

Removing the within-turn gaps by applying the above-mentioned post-processing (SHoUT_NoG), however, resulted in the best diarisation performance amongst the three different diarisation approaches (with 24.7% DER). The missing speaker error measure was lowered significantly from over 32.2% to 2.4%, although speaker error was almost double that of SHoUT_NoV. Eliminating both the overlapping segments and the within-turn gaps (SHoUT_NoV_NoG) did not improve the diarisation performance further.

5.2.2 Word diarisation error rate

Although the DER is the most common error measure for diarisation procedures, it is only based on the duration of the segments and does not indicate to what extent the diarisation outputs would be useful for ASRs alignments when word boundaries are detected, i.e, how many words would be assigned to the correct speaker if a perfect ASR was used. Therefore, for evaluation purposes, using forced alignments and the manual reference transcripts, we attempted to calculate some measures to indicate how the outputs of the diarisation systems would be useful for ASRs.

The measure which we used for this purpose is based on the ratio of the number of words, not assigned to the right speakers, to the total number of words (we refer to this measure as word diarisation error, Word Diarisation Error Rate (WDER)). This can be calculated by adding two errors (equation 5.3): the ratio of the number of missing words to the total words (missing words ratio, E_{MWR}) and the ratio of the number of words assigned to the wrong speakers (words assigned to the wrong speakers ratio, E_{WAWSR}):

$$\text{WDER} = E_{MWR} + E_{WAWSR} \tag{5.3}$$

Table 5.3 shows the WDER for the four diarisation systems (SHoUT, SHoUT_NoV, SHoUT_NoG, and SHoUT_NoV_NoG). SHoUT without the within-turn gaps (SHoUT_NoV) performed best and would lose only 5.4% of the total words if a perfect ASR was used after the diarisation system. This includes 3.6% of words assigned to wrong speakers and 1.8% of words missing.

Table 5.3: *Word diarisation error for the baseline diarisation systems.*

Diarisation system	E_{WAWSR}	E_{MWR}	$WDER$
SHoUT	23.4%	6.2%	29.6%
SHoUT_NoV	24.4%	6.4%	30.8%
SHoUT_NoG	**3.6%**	**1.8%**	**5.4%**
SHoUT_NoV_NoG	3.7%	3.6%	7.3%

The results displayed in Tables 5.2 and 5.3 confirm that the best diarisation system was SHoUT_NoG, which only lost 5.4% of the words expected to be recognised by the

ASR.

5.3 I-vector based diarisation

First following the Kaldi diarisation recipe for the CH dataset (with 500 conversations between friends over phone), a model was trained for the SWB and SRE2004, SRE2006 and SRE2008 datasets (named Kaldi(SWB) diarisation model). The model consist of 2048 GMM components and i-vector dimensions of 128. 32K utterances from the training set were used for learning the Universal Background Model (UBM) prior to the i-vector feature extraction. PLDA scoring was used for clustering and a threshold based criterion for stopping the AHC clustering approach. The threshold based criterion resulted in 12.7% DER for the CH diarisation task. Also we repeated the recipe knowing the number of speakers per conversation as a criterion to stop clustering. This way ended up to a further improvement and reduced the diarisation error to 12.1% (named Kaldi(SWB_Number of Speakers (NumSp))).

We used these two diarisation systems for evaluating the Hal dataset (Hal30). Table 5.4 shows the details of the diarisation error. Knowing the number of speakers in advance (Kaldi(SWB_NumSp)), gained 31.3% DER, slightly better than the model with the thresholds (Kaldi(SWB)) with 31.6% DER. Speaker false alarm and missing errors for both models were not important (0% and 0.1% respectively).

Similarly, we trained two diarisation models without and with knowing the number of speakers in conversation trained by Seizure dataset mixed with half of the Hal dataset (the other half were used for testing, i.e. two models trained without having overlap between training and testing sets). These two models were called Kaldi(SezHal) and Kaldi(SezHal)_NumSp respectively. As can be seen from the table, these two diarisation systems, had better performances since they were trained using the conversations recorded with similar conditions to the Hal data. The best i-vector based diarisation system Kaldi(SezHal)_NumSp achieved 20.6% DER which is around 4% better than SHoUT_NoG (with 24.7% DER in Table 5.1).

It is worth mentioning that the mentioned-above post-processing step (section 5.2.1)

Table 5.4: *DER for the Hal30 data using the Kaldi diarisation trained by the switchboard (SWB) or seizure data mixed with half of the Hal data (held-out approach)(SezHal), with or without knowing the number of speakers (NumSp).*

Diarisation system	E_{MISS}	E_{FA}	E_{SPKR}	DER
Kaldi(SWB)	**0.1%**	**0.0%**	31.5%	31.6%
Kaldi(SWB)_NumSp	**0.1%**	**0.0%**	31.2%	31.3%
Kaldi(SezHal)	**0.1%**	**0.0%**	22.5%	22.6%
Kaldi(SezHal)_NumSp	**0.1%**	**0.0%**	**20.5%**	**20.6%**

on the i-vector based diarisation models did not change the results, therefore, we ignored the step.

The word diarisation error for these models are summarised in Table 5.5. Comparing to Table 5.3, i-vector based diarisation systems, generally have less word diarisation error, however, surprisingly, SHoUT_NoG have a better performance with 5.4% WDER comparing to the best i-vector based diarisation system (Kaldi(SezHal)_NumSp with 7.2% WDER).

Table 5.5: *Word diarisation error for the i-vector based diarisation systems.*

Diarisation system	E_{WAWSR}	E_{MWR}	**WDER**
Kaldi(SWB)	15.4%	4.4%	19.8%
Kaldi(SWB)_NumSp	13.9%	5.9%	19.8%
Kaldi(SezHal)	4.5%	3.3%	7.8%
Kaldi(SezHal)_NumSp	**4.2%**	**3.0%**	**7.2%**

5.4 Discussion

The main dataset for our study, the Hal, contains interviews between doctors, patients and accompanying person(s) (if present). The style of the conversations is similar to that of a general meeting, although in a meeting people tend to speak more clearly and in a more formal manner using better structured language and a higher level of articulation.

In the experiments presented in this chapter, we aimed to find the best diarisation tool for the Hal dataset conditions. Initially two pre-trained diarisation tools were cho-

sen: LIUM (tuned for broadcast news and telephone conversation) and SHoUT (trained for meeting condition). The performance of two systems were evaluated using DER. The SHoUT diarisation system outperformed the LIUM with 5% less overall diarisation error and in particular with around 7% less speaker errors. Using a post-processing technique (merging gaps in segments and re-segmenting long portions) the overall DER was further reduced to around 25%. The i-vector based diarisation system trained by the Kaldi CH recipe (tuned for telephone conversation condition) had a relatively poor performance for the Hal dataset. Therefore, a new diarisation system was trained with a similar Kaldi recipe, but using Seizure data and half of the Hal dataset (avoiding any overlap with the test recordings). This diarisation system outperformed the best SHoUT diarisation system.

We expect that, with more data available we would be able to further improve the diarisation unit for our system.

The DER, although widely used, does not show how well the diarisation component will work when its outputs is passed to ASR. Therefore, we introduced the word-based WDER measure which is compromised of the missing word rate and the words wrongly assigned to speaker errors.

Although most of the i-vector based diarisation systems were better in terms of DER, the SHoUT_NoG were slightly better in terms of WDER (5% vs. 7% for Kaldi(SezHal)_NumSp). However, the i-vector based Kaldi diarisation system will be used as the final diarisation unit of our system (for convenience we will refer to this as the Kaldi diarisation in the following chapters).

In our study, we are interested in the conversations between doctors and patients, and will use a diarisation tool to determine who talks when. The speaker diarisation task for our study, thus, is not necessarily as difficult as the general diarisation problem (unknown number of speakers and conditions, audio mixing up with music, jingles, etc), since the number of speakers for each conversation can be assumed to be known and limited between two to and maximum of four speakers, and the speaker information such as gender and age is also available.

The conditions of the interviews of the doctors with the patients and/or accompanying

person(s) (APs) (who may help the patients to answer the questions) is close in style to the general meeting condition for diarisation, apart from the fact that the doctors mostly ask the questions, which would make their speech less spontaneous compared to the patients or the APs. Therefore we may expect to see that both the diarisation and speech recognition tasks for the doctors' part result in a lower error rate than observed for the patients or the APs.

5.5 Summary

Speaker diarisation system aim at segmenting input audio streams into smaller segments and assigning the segments to the speakers in a conversation. The general architecture of the system comprises of acoustic beamforming for a meeting condition (in cases with multiple distant microphones), a speech activity detection unit to split audio into speech and non-speech (e.g. noise, music, laugh) segments, a speaker segmentation unit to segment the audio into smaller portions, with each neighbouring portion assumed to belong to a different speaker (speaker change detection), and finally speaker clustering to merge the smaller segments that belong to the same speakers. A second pass of the diarisation system may needed to further re-segment and cluster segments.

A few commonly used diarisation toolkits were listed in the chapter and in particular we used the SHoUT, LIUM and Kaldi diarisation toolkits.

A new measurement for evaluating the performance of different diarisation system was introduced (WDER). WDER is word-based and therefore more suitable for assessing how well a particular diarisation performance affects the subsequent ASR. Regarding both the DER and the WDER, between the two best diarisation systems, the Kaldi diarisation system was chosen, as it produced better classification results in the dementia detection system.

Chapter 6

Feature extraction

Contents

6.1 Introduction 121

6.1.1 Extended acoustic features 121

6.1.2 Extended lexical features 123

6.1.3 Word vector features 125

6.2 Classification results 129

6.2.1 Effect of different feature types 129

6.2.2 Confusion matrix 130

6.2.3 Feature selection 131

6.2.4 The Receiver Operating Characteristic (ROC) curve 134

6.3 Discussion 134

6.4 Summary 138

In **Chapter 3** initially we focused on the feature extraction and classification units of our dementia detection system, assuming the other components were available. A number of classifiers and features were introduced to confirm the proof of concept for the automatic dementia detection.

After introducing the ASR and the diarisation units of the system in **Chapter 4** and **Chapter 5**, this chapter focuses again on the feature extraction component of the automatic system, exploring more features with the aim of narrowing down the final features of the system. The chapter is organised as follows:

Section 6.1 describes the expansion of our initial feature set to a numbers of different types of features, including the extended acoustic, the extended lexical, and the word vector based features.

Section 6.2 provides the classification results of the automatic dementia detection system, as well as discussion about the importance of different feature types, feature selection and robustness of the classifier.

Section 6.3 and **Section 6.4** provides the discussion and the summary of this chapter respectively.

6.1 Introduction

Due to the lack of standard datasets and ethical difficulties around the sharing of data collected from people with medical conditions, still it is hard to find robust and widely used set of features based on audio and speech which can help in identifying dementia. Exploring and identifying such features continues to be an important part of current studies. Moreover, different studies have been based on different datasets each with a limited number of examples recorded in different conditions (e.g. patient describing a picture or completing a task) and as a result different types of features have been introduced by different research groups.

In **Chapter 3** we introduced a set of CA-inspired features (refer to **Section 3.2.2** and Table 3.3). Due to time limitations, we chose not to focus on extracting visual features automatically. Therefore, out of the 22 features, the two visual-conceptual features were removed. The remaining 20 CA-inspired features consisted of three feature types: acoustic, lexical and semantic features. In this chapter we will extend the acoustic and lexical features and also we will add the **word vector features** as an additional feature type. Note that in order to avoid the confusion between the new acoustic and lexical features with the previous acoustic and lexical features (see Table 3.4) we call them the **extended acoustic** and the **extended lexical features**.

6.1.1 Extended acoustic features

In our baseline CA-inspired features we already introduced a number of acoustic features based on the length and the number of turns for each speaker role and the pauses for patients in conversation. However, many studies introduced other acoustic features (see **Section 3.1**). Adding more acoustic features will let us to explore and identify more important features which can help in identifying dementia from the audio recordings of the conversations.

Using the well known 'Praat vocal toolkit' [Boersma et al., 2002], the total number of 36 acoustic features (12 for each speaker role: neurologist, patient and accompanying person(s)) were extracted from the audio files. Since we were interested in features usually

Table 6.1: *20 CA-inspired features: acoustic, lexical, semantic. Note: two visual-conceptual features from Table 3.3 were removed.*

Type	**Features**
Acoustic	APsNoOfTurns PatNoOfTurns NeuNoOfTurns APsAVTurnLength PatAVTurnLength PatFailureExampleAVPauses NeuAVTurnLength PatAVPauses
Lexical	PatAVUniqueWords NeuAVUniqueWords APsAVUniqueWords PatAVAllWords
Semantic	PatMeForWhoConcerns PatFailureExampleEmptyWords PatFailureExampleAllTime PatDontKnowForExpectation PatAVFillers PatAVEmptyWords AVNoOfRepeatedQuestions AVNoOfTopicsChanged

seen in formal CA transcripts, we extracted a number of features including the prosodic features (average overall duration, pitch, intonation, and silence), the features capturing creakiness and breathiness (difference between the first harmonic and the harmonic close to the first, second and third formants: H1-A1, H1-A2, H1-A3 [Gordon and Ladefoged, 2001; Khan et al., 2015]; difference between the two first harmonics: H1-H2), and features related to the vocal stability (jitter, shimmer, harmonics-to-noise and noise-to-harmonics ratios). In order to extract these acoustic features, we gave the timing information from the ASR outputs (i.e. start and end time of the words and silences) to the Praat. We calculated the pitch, intonation, duration, etc. on the speech (spoken words) of the audio files, and only the average silence length feature was extracted on the silence segments. Table 6.2 shows these features for the three speaker roles.

It is worth mentioning that there are other toolkits for extracting the acoustic features. For instance, the Open Smile toolkit [Eyben et al., 2010] is widely used for extracting different types of acoustic features. However, for our system, we preferred to use the Praat to extract the acoustic features.

Table 6.2: *List of the extended acoustic features.*

Speaker role	Feature
'Neu' (neurologist)	average overall intonation, pitch, duration and silence(**NeuAvgIntonation**, **NeuAvgPitch**, **NeuAvgDuration NeuAvgSil**); difference between the first harmonic and the harmonic close to the first, second and third formants(**NeuAvgH1-A1**, **NeuAvgH1-A2**, **NeuAvgH1-A3**); difference between the two first harmonics (**NeuAvgH1-H2**); local jitter and shimmer(**NeuAvgGitterLocal**, **NeuAvgShimmerLocal**); harmonics-to-noise and noise-to-harmonics ratios(**NeuAvgMeanHNR**, **NeuAvgMeanNHR**)
'Pat' (patient)	average overall intonation, pitch, duration and silence(**PatAvgIntonation**, **PatAvgPitch**, **PatAvgDuration PatAvgSil**); difference between the first harmonic and the harmonic close to the first, second and third formants(**PatAvgH1-A1**, **PatAvgH1-A2**, **PatAvgH1-A3**); difference between the two first harmonics (**PatAvgH1-H2**); local jitter and shimmer(**PatAvgGitterLocal**, **PatAvgShimmerLocal**); harmonics-to-noise and noise-to-harmonics ratios(**PatAvgMeanHNR**, **PatAvgMeanNHR**)
'APs' (accompanying person(s))	average overall intonation, pitch, duration and silence(**ApsAvgIntonation**, **ApstAvgPitch**, **ApsAvgDuration ApsAvgSil**); difference between the first harmonic and the harmonic close to the first, second and third formants(**ApsAvgH1-A1**, **ApsAvgH1-A2**, **ApsAvgH1-A3**); difference between the two first harmonics (**ApsAvgH1-H2**); local jitter and shimmer(**ApsAvgGitterLocal**, **ApsAvgShimmerLocal**); harmonics-to-noise and noise-to-harmonics ratios(**ApsAvgMeanHNR**, **ApsAvgMeanNHR**)

6.1.2 Extended lexical features

As mentioned in **Chapter 2**, some studies focused on extracting the lexical features. Blanken et al. [1987]; Bucks et al. [2000] reported that the number of nouns produced by people with dementia were significantly less than the healthy controls. Jarrold et al. [2014] found out that 11 out of 14 of their Part Of Speech (POS) features were statistically significant in differentiating between different types of dementia. Therefore,

Table 6.3: *List of the extended lexical features.*

Speaker role	**Feature**
'Neu' (neurologist)	average number of verbs, nouns, adjectives, adverbs, pronouns, wh_words(e.g, who), determiners, conjunctions, cardinals, existential(e.g., there is), prepositions (**NeuAvgVerb**, **NeuAvgNoun**, **NeuAvgAdjective**, **NeuAvgAdverb**, **NeuAvgPronoun**, **NeuAvgWh_word**, **NeuAvgDeterminer**, **NeuAvgConjunction**, **NeuAvgCardinal**, **NeuAvgExistential**, **NeuAvgPreposition**, **NeuAvgOtherPOS**)
'Pat' (patient)	average number of verbs, nouns, adjectives, adverbs, pronouns, wh_words(e.g, who), determiners, conjunctions, cardinals, existential(e.g., there is), prepositions (**PatAvgVerb**, **PatAvgNoun**, **PatAvgAdjective**, **PatAvgAdverb**, **PatAvgPronoun**, **PatAvgWh_word**, **PatAvgDeterminer**, **PatAvgConjunction**, **PatAvgCardinal**, **PatAvgExistential**, **PatAvgPreposition**, **PatAvgOtherPOS**)
'Aps' (accompanying person(s))	average number of verbs, nouns, adjectives, adverbs, pronouns, wh_words(e.g, who), determiners, conjunctions, cardinals, existential(e.g., there is), prepositions (**APsAvgVerb**, **APsAvgNoun**, **APsAvgAdjective**, **APsAvgAdverb**, **APsAvgPronoun**, **APsAvgWh_word**, **APsAvgDeterminer**, **APsAvgConjunction**, **APsuAvgCardinal**, **APsAvgExistential**, **APsAvgPreposition**, **APsAvgOtherPOS**)

we were interested in extracting the lexical features, i.e. different POS for the words in conversations.

Penn Treebank part of speech tags [Taylor et al., 2003] were assigned to the words uttered by each type of speaker in the conversations. The number of the Penn Treebank' tags were originally 36, however, similar tags (e.g. different types of verbs) were joined together to make more general, higher-level tags. The tags were gathered under 12 different groups for each speaker role (a total of 36 features for the three speaker roles). Table 6.3 shows these tags for the three speaker roles, including average number of verbs, nouns, adjectives, adverbs, pronouns, 'wh' words, determiners, conjunctions, cardinals, existential and prepositions.

6.1.3 Word vector features

Machine learning algorithms work on vectors of numbers, and word embedding is a technique which is widely used to convert text to numbers; instead of a word, a series of numbers are used. Traditionally, techniques such as BoW [Harris, 1954] and Term Frequency-Inverse Document Frequency (TF-IDF) [Sparck Jones, 1972] were used for word embedding with some success. Recently, more successful approaches have used deep learning techniques to produce vector representing words. Two recently introduced techniques are 'Word to Vector (W2vec)' [Mikolov et al., 2013a,b] and 'GloVe' [Pennington et al., 2014] which are both based on the co-occurrences of words, taking into account the context (neighbouring words) in a text.

The 'W2vec' is trained using a simple three layer deep neural network (input, hidden and output layers). It can learn the word vector using two techniques: skip-gram and Continuous Bag of Words (CBoW). The skip-gram aims at predicting the context from a given word, while the CBoW attempts to predict a word given a context. Generally, the skip-gram can capture more information than that captured by the semantics of a single word, and using the negative sub-sampling technique, the skip-gram technique can outperform the CBoW. Mikolov et al. [2013a,b] demonstrated that the resulting 'W2vec' vectors exhibits some interesting properties, for instance, that vector("King") - vector("Man") + vector("Woman") is very close to the vector("Queen"). Despite the amazing advantages of the 'W2vec', it has some limitations including not taking into account the global co-occurrence of the words in the whole corpus. The 'GloVe' adds the benefits of the matrix factorisation approaches to the skip-gram to capture the global statistical information. Instead of focusing only on the probabilities of words in the context, the ratio of co-occurrence probabilities are taken into account. In fact, the 'GloVe' attempts to associate the logarithm of ratios of co-occurrence probabilities with the vector differences. The authors of the 'W2vec'[1] and 'GloVe'[2] have both shared their pre-trained models for public use.

One of the main applications of the word vector encoding techniques is sentiment

[1]http://mccormickml.com/2016/04/12/googles-pretrained-word2vec-model-in-python/
[2]https://nlp.stanford.edu/projects/glove/

analysis - the problem of identifying opinions or moods in a piece of text. A popular benchmarks for sentiment analysis is the ICL Internet Movie Database (IMDB) containing 50000 movie reviews associated with positive or negative sentiments (half for training and half for testing). Inspired by the language modelling and probabilistic latent topic models, Maas et al. [2011] introduced a model for the vector representation and achieved an accuracy of 88.89% for the binary classification task. Le and Mikolov [2014] extended the 'W2vec' model to make vectors representing paragraphs or documents ('Document to Vector (Doc2vec)'). The main idea was to add an extra token (ID) for each document to the content while training the BoW or skip-gram model. They reported 92.58% accuracy for the sentiment analysis task of IMDB, however, other researchers have struggled to reproduce the same outcomes [Lau and Baldwin, 2016]. Combining CNNs with BLSTMs Shen et al. [2017] resulted in a classification rate of 89.7% for the IMDB sentiment analysis task. Random embedding substitution obtained 88.98% accuracy by using a normal LSTMs and 89.71% using BLSTM. Yenter and Verma [2017] also reported 89% accuracy using CNN and LSTM. In addition to the sentiment analysis, the word vector has been used in various NLP tasks such as semantic queries [Bordawekar and Shmueli, 2017], semantic textual similarity [Lau and Baldwin, 2016], document analysis [Park et al., 2018], and text understanding [Gao et al., 2017].

Recently, word vector has been used in a number of different tasks involving spoken language. Tao et al. [2016] applied the 'Doc2vec' (an expansion of 'W2vec' to documents) to the ASR outputs of non-native English speaker taking the TOEFL internet-Based Test (iBT) to score (measure) the responses, and they observed a considerable amount of improvements comparing to using TF-IDF features. Audhkhasi et al. [2017] used the 'GloVe' embedding to initialise the final dense layer of their deep neural network to directly convert acoustic features to words and reported reasonably low WER on the SWB and CH dataset.

The use of word vector for detection of pathologies and para-linguistic information in speech is very novel. Lopez-Otero et al. [2017] used the 'GloVe' word vector to detect depression from the transcripts produced by the ASR on the de-identified speech (modifying voice characteristics for privacy reasons). They split each turn of a speaker into a

series of words. Each turn was then represented by summing up the normalised 'GloVe' word vector of the turn. They also considered applying a weight coefficient to the vectors allowing them to assign more importance to the rarer words. Then, they reduced the dimension of the turn vectors by using the PCA algorithm and used an SVM classifier to classify between depression and non-depression speech. They gained 80% classification accuracy using de-identified speech recognised by ASR (with 37.3% WER).

For classification tasks we need to use the word vector representations of the individual words in a transcript in a way that enables us to distinguish the different classes. This section describes the four different approaches we have investigated. The first two are based on composing a vector from the individual word vector and using these vectors to train a classifier as per usual; the third method uses a cosine similarity as a measure of how different a word vector is to typical word vector found in the labelled/known classes; the fourth approach models the vectors from the first two approaches in a sequential model.

Average/Variance of word vector

Assume a corpus C consists of n documents, $D_i; 1 \leq i \leq n$. Each document consists of a number of words, $D_i = w_{i1}, ..., w_{ij}$ and each word can be converted to a vector V with d dimensions as $V(w_{ij})$ using one of the pre-trained word vector algorithms like 'W2vec' or 'GloVe'. Ignoring the non important words in a text (stop list) as well as replicated words, we can make a new vector by calculating the average of the word vector appended to the variance of the word vector as:

$$AV(D_i) = [\mu(D_i), \sigma(D_i)] \tag{6.1}$$

where μ and σ are the average and variance and $AV(D_i)$ has dimensions $2 * d$. The first proposed approach uses the $AV(D_i)$ vectors for training a classifier.

Difference between Average/Variance of word vector

The second approach is similar but based on a feature vector derived as the difference between $AV(D_i)$ and a vector combined over all training documents in each class. That is, for a supervised classification task with m known classes, $c_{1...m}$, we can make m combined AV vectors: $AV(c_l); 1 <= l <= m$. The feature vector in this second approach is found by summing the differences between $AV(D_i)$ and each $AV(c_l)$. We refer to this vector as

$DiffAV$:

$$DiffAV(D_i) = \sum_{l=1}^{m}(AV(c_l) - AV(D_i) \tag{6.2}$$

Cosine similarity between word vector

As a third approach for representing documents, we calculate the cosine similarity between the word vector of a document and the word vector of each class. The value of the cosine similarity will be normalised (sum up to one). We refer to this as $CosWV$ (cosine word vector) and define it as:

$$CosV(c_l, D_i) = \sum_{j=1}^{k}\sum_{t=1}^{r} cos(V(w_{ij}), V(w_{lt})) \tag{6.3}$$

$$CosWV(D_i) = \left[\frac{1}{M}CosV(c_1, D_i), .., \frac{1}{M}CosV(c_m, D_i)\right] \tag{6.4}$$

where $M = Max(CosV)$ and k and r number of words in document D_i and class c_l respectively.

Sequence word vector

For the fourth and final approach, we extract fixed length frames of the whole document using a sliding window over the text (we have used the length of 80 words with a 25% overlap) and computing the AV and $DiffAV$ vector of each frame gives us $SeqAV(D_i) = [AV(D_{i1}), ..., AV(D_{if})]$ and $SeqDiffAV(D_i) = [DiffAV(D_{i1}), ..., DiffAV(D_{if})]$.

From these approaches, which were introduced in Mirheidari et al. [2018a], we choose the first measure (AV) to represent the meaning of the conversations. This feature, then, is added to the other features to complete our system's feature set. Using the Principal Component Analysis (PCA) approach we could reduce the dimension of the vectors from 600 to 7. Almost the same results can be achieved using the reduced dimensions.

Note that in the paper we showed that the CosWV will give us the best classifier accuracy for the Hallamshire dataset, however, since we use leave-one-approach to split data into train and test sets, each time the CosWV values changes for different data samples, and so it is not possible to add the word vector feature to our other features.

6.2 Classification results

As mentioned before in Chapter 3, the LR classifier was nominated as the final classifier for the dementia detection system. In addition to the baseline CA-inspired features, in this chapter three additional feature sets are introduced. Now we want to investigate the contribution of the different feature types in the classification task.

6.2.1 Effect of different feature types

Table 6.4 shows the accuracy of the LR classifier for different features types and the three levels of automation to classify between ND and FMD patients of the Hallamshire dataset (15 ND and 15 FMD). Each row shows the classifier accuracy with increasing transcription automation and segmentation ranging from the manual transcript to the automatic segmentation and transcription (Diarisation (Diar)+ASR). The CA-inspired features for the ASR and the Diar+ASR were produced using the output text and the timing information provided by the ASR outputs. Note that due to the diarisation error we could not manage to extract the word vector features for three conversations (affected results are marked with stars in the table).

Table 6.4: *Accuracy of the LR classifier to classify between ND and FMD patients using different feature types, the three levels of transcription automation and segmentation. Inside the brackets is the number of features. CA: CA-inspired features, E-LX: extended lexical features, E-AC: extended acoustic features, WV: word vector features, *: due to the errors caused by the diarisation systems, 3 out of 30 word vector based results were missing. Combined features: final column shows results for all feature combined together.*

Level of automation	CA(20)	E-LX(36)	E-AC(36)	WV(7)	Combined features(99)
Manual Transcript	**90.0%**	76.7%	66.7%	70.0%	76.7%
ASR	**96.7%**	80.0%	73.3%	70.0%	83.3%
Diar+ASR	**90.0%**	83.3%	66.7%	74.1%*	**90.0%**

For the manual transcript (row 1), accuracy of the classifier trained on the baseline CA-inspired features was 90.0% (the best result amongst the four feature types). Using the extended lexical features, the accuracy of the classifier dropped to 77%, while for the extended acoustic and the word vector features the accuracy were 67% and 70% respec-

tively. Putting together all the features (the combined features), however, resulted in 77% accuracy (13% less than the CA-inspired features). Combining all features together not necessarily result in a better accuracy rate for the classifier (some features may correlated negatively together, i.e. putting them together cause more confusion for the classifier). That is why sometimes feature selection can achieve a better result than using all the features.

For the ASR, the classifier trained on the CA-inspired features achieved 97% accuracy, while the accuracy for the extended lexical features was 80.0%. The extended acoustic features and the word vector features gained 73% and 70% accuracy respectively. Accuracy of the classifier with the combined features was also 83%.

Similarly for the Diar+ASR, the accuracy of the classifier trained on the CA-inspired features achieved the best result among the four feature types with 90.0%. The second highest accuracy was achieved by the extended lexical features with 83%. The word vector features and the extended acoustic features came third and fourth with 74% and 67% respectively. Using the combined features, the accuracy achieved 90%.

On the whole, the CA-inspired features always achieved the best classifier accuracy. The second most important features were the extended lexical features. The extended acoustic features and the word vector features had lower accuracy.

Combining all features together (99 features) resulted in the same or slightly better classifier accuracy than the extended lexical features, but was not as good as the accuracy achieved by the CA-inspired features. Some features in the combined features are not complementary and they may correlated together negatively. Therefore, to have a better result it is needed to find a subset of the features with the highest contributions.

6.2.2 Confusion matrix

Confusion matrix is a table which can be used to show the performance of a classifier. The rows of this matrix shows the predicted class by the classifier and the columns shows the real or true classes. Table 6.1 shows the confusion matrix for the classifier (Diar+ASR) trained using the 99 features. 14 out of 15 FMD were predicted correctly (93% correct) and only one (7%) was confused with ND, while 13 out of 15 ND were predicated truly

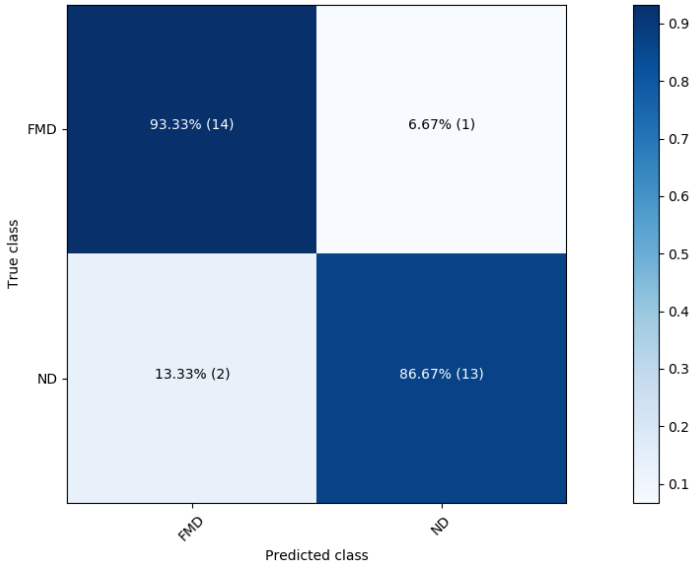


Figure 6.1: *Confusion matrix for the classifier (Diar+ASR) using all 99 combined features.*

(87%) and 2 were confused with FMD (13%). This shows that predicting the FMD group was slightly easier (with less confusion) than the ND.

6.2.3 Feature selection

In this section, as in **Section 3.3.1**, using the RFE technique (on the train set) the important features (the top n features) were selected from the combined features. The results of the classifier accuracy for the three levels of transcription automation and segmentation (manual transcript, ASR, Diar+ASR) using the top n features ($n = 1, ..., 10$) are shown in Figure 6.2.

For the manual transcript (red line) using the top 1 feature, the accuracy of the classifier was 77%. Adding one more top feature resulted in a 83% accuracy. For the top 3 to the top 10 except for the top 8, the classifier achieved an accuracy of 97%. Using the top 8 features, the classifier achieved 100% accuracy.

For the ASR (grey line), the top 7 to the top 10 features all resulted in 100% accuracy, while for the Diar+ASR (green line), the top 3 to the top 5 features had 100% accuracy.

As can be seen from the figure, deciding on the optimum number of top features is unknown and there is not a clear answer to it, i.e. it may differ from one classifier to another classifier. The minimum n with the highest accuracy rate for the three transcription

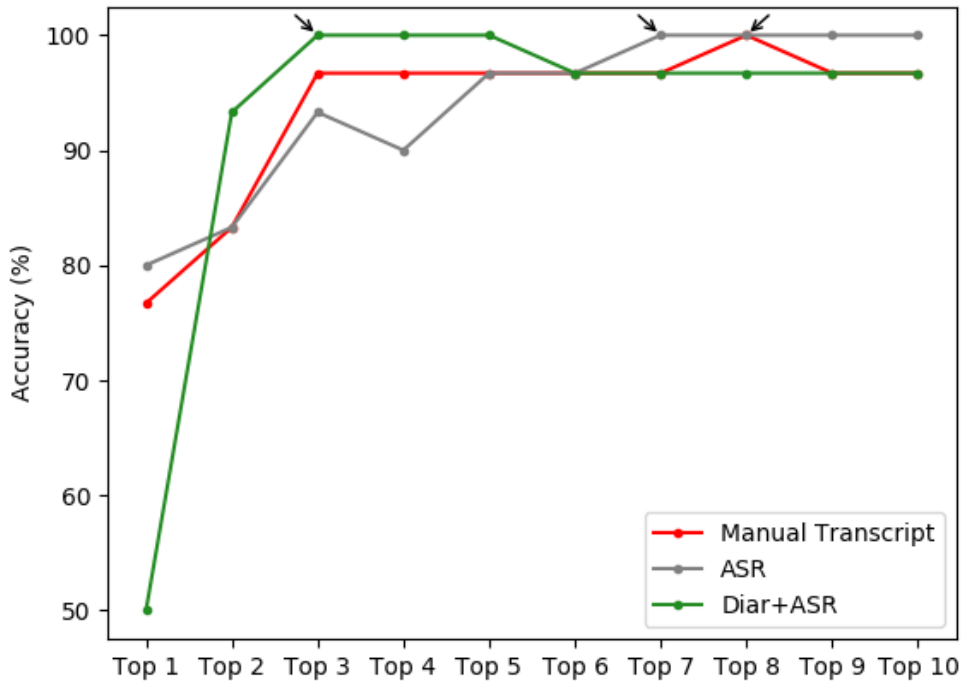


Figure 6.2: *The classifier accuracy for the three levels of transcription automation (manual transcript (hatched red), ASR (grey), Diar+ASR (hatched green)) using the top n features from the combined features. $n = 1, ..., 10$. The lowest number of features to achieve a 100% accuracy, for the three levels of transcription automation are marked by arrows.*

automation are marked with arrows.

The list of the top 10 features is shown in Table 6.5. As expected from the previous section, the CA-inspired features are the dominant top features with the highest ranks for the three levels of transcription automation. In addition it is seen that features from all the three speaker roles are in the top 10 features.

For the manual transcript (the first column), 8 out of 10 important features are acoustic features and the remaining 2 are the lexical features. Note that, some of the features which were chosen by the RFE in Table 3.7 could be found in Table 6.5, but not all of them. Since, here the classifier had a combination of 99 features to select using the RFE approach (remove one feature at time with the lowest contribution in classification) which let the extended acoustic and word vector features come in to the top 10 list as well.

For the ASR (the second column), 5 important features are acoustic, 3 are lexical and 2 are word vector features. Comparing to the first column, 7 features are the same, which shows despite the errors caused by of the ASR most of the important features remained

Table 6.5: *Top 10 features with the highest contributions in classification between ND and FMD patients for the three levels of transcription automation (Manual transcript, ASR, and Diar+ASR). WV_col1: The word vector features column 1. *: were seen in Table 3.7. AC: acoustic features, LX: lexical features, SM: semantic features, and WV: word vector features.*

Rank	Manual Transcript	ASR	Diar+ASR
1	PatAVAllWords*(LX)	PatAVAllWords*(LX)	NeuAVTurnLength(AC)
2	NeuAvgH1-A1(AC)	NeuAvgH1-A1(AC)	APsNoOfTurns*(AC)
3	APsNoOfTurns*(AC)	APsNoOfTurns*(AC)	NeuAvgH1-A2(AC)
4	APsAVTurnLength(AC)	APsAVTurnLength(AC)	WV_col3(WV)
5	PatAvgMeanHNR(AC)	APsAvgH1-H2(AC)	PatFailureExampleEmptyWords*(SM)
6	ApsAVUniqueWords(LX)	PatAVTurnLength*(LX)	WV_col1(WV)
7	PatNoOfTurns(AC)	WV_col1(WV)	WV_col5(WV)
8	NeuNoOfTurns(AC)	APsAVUniqueWords(LX)	NeuAvgH1-A1(AC)
9	PatAVTurnLength*(AC)	WV_col4(WV)	WV_col2(WV)
10	NeuAvgH1-H2(AC)	NeuAvgH1-A2(AC)	NeuAVUniqueWords*(LX)

the same (70%).

For the Diar+ASR (the last column), fully automatic system, however, the most important features are very different (only 3 are the same as the first column). The errors of the diarisation and the ASR together has made much more effect on selecting the top 10 features. 4 features are acoustic, 4 are word vectors and 1 semantic feature and 1 lexical features. The turn length for the neurologist and the number of turns for the accompanying person (NeuAVTurnLength and APsNoOfTurns) are the top two important features, followed by the average number of unique words uttered by the neurologists (NeuAVUniqueWords) and the number of empty words uttered by the patients in response to recalling the last memory failure example (PatFailureExampleEmptyWords).

It worth mentioning that we followed the approach introduced in Chapter 3 to find the most statistically significant features using the normality tests (Shapiro-Wilk and D'Agostino) and then parametric Student's t-test for normal features and non-parametric Mann-Whitney U test for non-normal features. We found 28 out of 99 features were statistically important features for the three level of automation (manual transcript, ASR and Diar+ASR), however, the accuracy of the classifier using those 28 features were 83.3%, 93.3% and 90.0% respectively, which were less than the accuracy gained by using the RFE feature selection.

6.2.4 The ROC curve

As it was mentioned in **Chapter 2**, both sensitivity and specificity are important for the cognitive diagnostic process. For instance, if a cognitive test can recognise the ND patients with 100% accuracy, but it cannot correctly recognise most of the non-ND patients, its sensitivity is high but its specificity is low.

Similarly, for a classifier, it is important to have high sensitivity and high specificity[1]. The ROC curve shows the true positives against the false positives for different settings (thresholds) of a classifier, i.e. compares the sensitivity against the specificity.

Figure 6.3 shows the ROC curve for the classifier for the fully automated transcription using the 99 features. The dashed red line shows the chance level (i.e. 50% for a binary classifier). The curve below the chance level (down right) could be considered as a weak classifier (here, there is not any curve below the chance level). As can be seen, by increasing the false positive rates, the true positive rates are not going to drop considerably. Note that we could not use the standard leave-one-out cross validation approach here (the test set should have samples from more than one class) and instead we used the k-fold ($k = 15$) cross validation. The accuracy of the classifier trained by the k-fold ($k = 15$) cross validation was 90%. The average ROC Area Under Curve (AUC) was 92%, which indicates a robust classifier (with a high sensitivity and a high specificity).

For the top 3 features the ROC AUC was 98% with classifier accuracy of 100% (Figure 6.4).

6.3 Discussion

In addition to the initial baseline CA-inspired features (which already consisted of some acoustic, lexical and semantic features), new features were introduced for the three levels of automation for transcription and segmentation: the extended acoustic, the extended lexical and the word vector features. Therefore, we had four feature types: acoustic, lexical, semantic and word vector features.

Considering the feature types separately, for the three levels of automation for

[1]Often these are referred to as recall and precision respectively

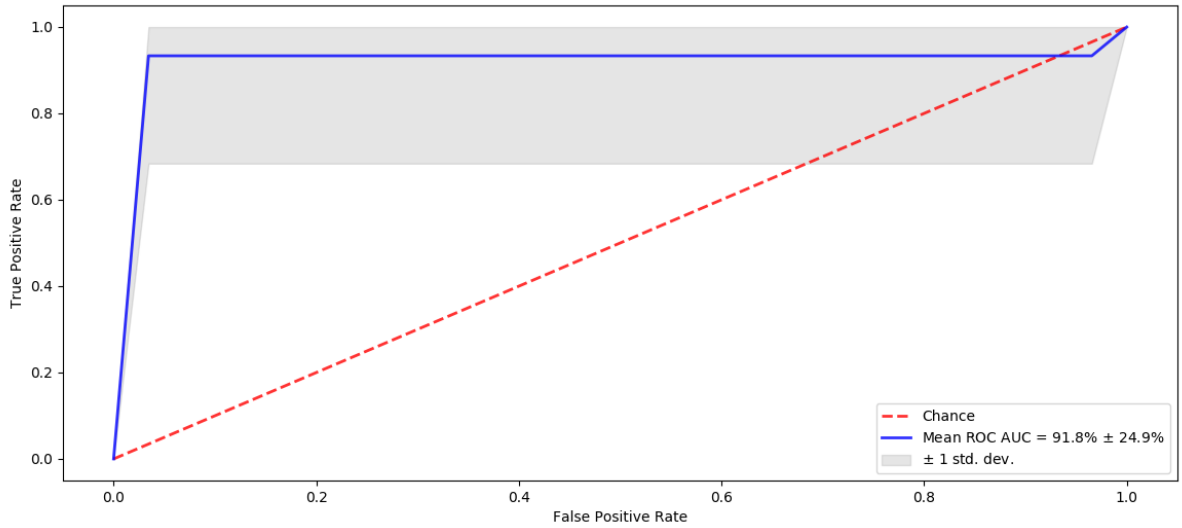


Figure 6.3: *ROC curve for the classifier of Diar+ASR using the 99 features (k-fold ($k = 15$) cross validation, accuracy of the classifier: 90%).*

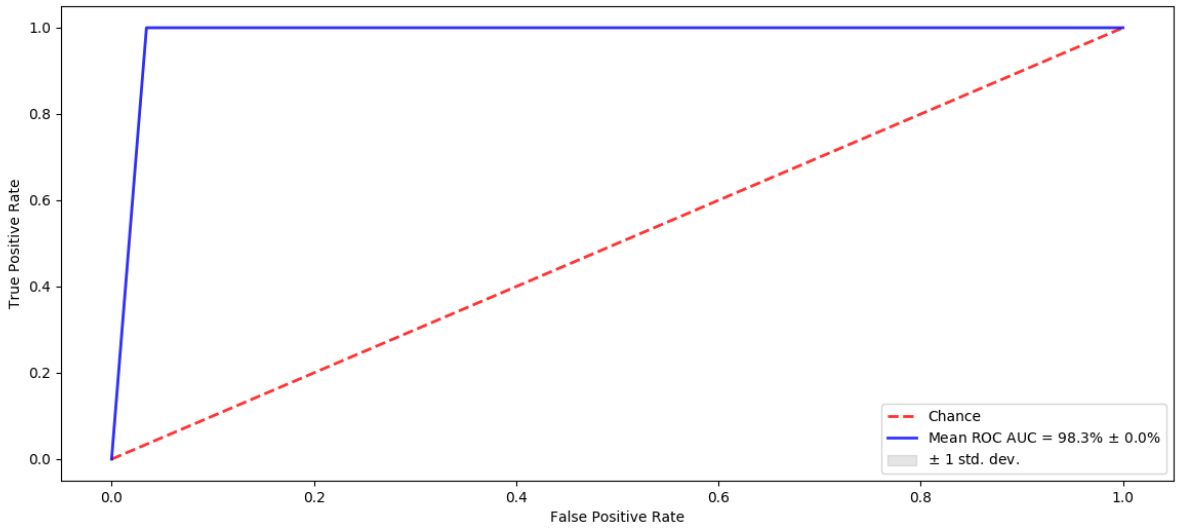


Figure 6.4: *ROC curve for the classifier of Diar+ASR using the top 3 features (k-fold ($k = 15$) cross validation, accuracy of the classifier: 100%).*

transcription and segmentation, the most important features mostly were acoustic features. However, the second important features for the manual transcript were the lexical features but for the ASR the second important features were the lexical and the word vector features, while the second important feature type was the word vector feature for the Diar+ASR. The least important features for the Diar+ASR were the

lexical and the semantic features.

Having different top 10 features for the three levels of automation for transcription and segmentation, shows that the errors caused by the ASR and the diarisation unit changes the extracted features, however, despite having these errors the binary classifier can classify with a high accuracy rate of 90%. This might be because of the fact that the binary classification between the two patient groups was not a complicated task. As we add more patient groups to the classification, the effects of the errors of the ASR and the diarisation unit will be more.

Comparing the top 10 out of 99 features for the manual transcript to the top 10 features in Table 3.7 (out of 22), three of the features (PatAVAllWords, APsNoOfTurns, PatAVTurnLength) were the same but the rest were different. One important reason for the difference is that in Chapter 3 we used the ranking from the five classifiers together to find the top 10 features, but here we only used the ranking of a single classifier (LR). The other reason is that we provide the classifiers wit different sets of features. Although 20 of the features were the same, the RFE approach recursively removes a feature with the lowest contribution in classification task. So it is possible in the middle of the selection process, to have some features with rankings close together which let the approach choose different features to remove. Also some of the new extended features were important in discriminating between the two patient groups.

Al-Hameed et al. [2018] extracted 812 acoustic features from the patient-only segments of the audio files of the Hallamshire dataset (same as our manual transcript, except they manually separated the patient parts of the conversations and manually aligned the words in the segments). They trained five classifiers and their best classifier, the linear SVM, gained 87% accuracy using all features, and 97% accuracy using the top features (both for the top 7 features using the SVM wrapper approach and the top 24 features using the Pearson's Filter). Comparing to their results, for the manual transcript, the CA-inspired features achieved 90% accuracy (3% better than their accuracy for 812 features), and the combined features had 77% accuracy (10% less than their result). However, for all the three levels of automation, using the feature selection approach (top n features), the accuracy of our classifier could achieve 100% accuracy (3% better than their result for the

top features with 97%).

It should be mentioned that our extended acoustic features (even for the manual transcript) were extracted completely automatically from the conversations using the ASR alignments and the Praat toolkit, i.e. first we determined the start and end time of each word from the alignments produced by the ASR and then we passed the timing information to the Praat. As we know, the ASR alignments at both the phoneme and the word levels are error-prone. These errors affect the extracted acoustic features, while Al-Hameed et al. [2018] used the manual alignments, which hardly can be produced by the automatic approaches.

In addition, since the baseline CA-inspired features already had high contributions to classifications (see Table 3.5), adding new features types did not improve the overall classification accuracy much. The CA-inspired features included the acoustic, lexical features and semantic. Looking at the top 10 features for the three levels of transcription automation, we can find acoustic features and lexical features as the most important features especially the turn related features (e.g. APsNoOfTurns, NeuAVTurnLength) and the lexical features (e.g. PatAVAllWords, APsAVUniqueWords).

The ROC curve analysis showed that the binary LR classifier that we used for our experiments, was robust, i.e. it was not showing low sensitivity nor low specificity, however, we expect that as we add more patient groups to the classifier (in the next chapters), the curve would tend more to the chance line, since the task of the classification gets harder and the classifier makes more mistakes.

6.4 Summary

In this chapter, two feature types were extended (acoustic and lexical) and one new feature type was introduced, i.e. the word vector features.

The four feature types were extracted for the three levels of transcription automation (the manual transcript, the half automated (ASR), and the fully automated (Diar+ASR)) and used to train the classifier to classify between the two patient groups: ND and FMD.

The feature types with the highest classification accuracy were the CA-inspired with a 90% accuracy and the extended lexical features with a 83% accuracy for the fully automated transcription. Putting together the four feature types (combined features) resulted in the same accuracy.

Using the RFE feature selection approach, the accuracy of the classifier achieved 97% using the most significant (2 or 7) features for the fully automated transcription (same accuracy achieved for the manual transcript using the most significant 3 features). The majority of the features were acoustic features. The next important feature types were the word vector and lexical features. The least important features, however, were the semantic features.

Finally, using the ROC curve analysis we showed that the classifier was robust with an average 98% ROC AUC using the most significant (3) features for the fully automated transcription (92% ROC AUC using all 99 combined features).

Chapter 7

Intelligent Virtual Agent

Contents

7.1 Introduction 141

7.2 Verbal fluency tests 141

7.3 Using an IVA to elicit conversation 142

7.3.1 IVA datasets 144

7.4 Results 147

7.4.1 Confusion matrix 148

7.4.2 Feature selection 149

7.4.3 The Receiver Operating Characteristic curve 152

7.4.4 Comparing neurologist-led to IVA-led conversations 152

7.5 Discussion 154

7.6 Summary 157

In **Chapter 3**, we introduced our automatic dementia detection system and in **Chapters 4**, **5** and **6** we focused on different components of the system including the ASR, the speaker diarisation unit, the classifier and the feature extraction. As a final step towards full automation, in this chapter, we will introduce an Intelligent Virtual Agent (IVA) (an animated talking head displayed on a screen) to conduct the conversations with patients (asking similar questions as the neurologists asked in the Hallamshire dataset). The conversations between the IVA and the patients then will be passed to the automatic dementia detection system to predict a diagnostic label for the conversation.

We have been able to deploy the IVA in a memory clinic setting (at the Sheffield Royal Hallamshire Hospital) in three summers (2016, 2017 and 2018). This chapter focuses on the data collected during the 2016 summer (the dataset is called IVA2016). The chapter is structured as below:

Section 7.1 is a general introduction to virtual agents.

Section 7.3 describes our IVA, which was designed to collect data from people with dementia and other memory issues.

Section 7.4 contains the results achieved from the IVA2016 dataset using the dementia detection system.

Section 7.5 and **Section 7.6** present the discussion and the summary of this chapter respectively.

7.1 Introduction

The use of IVAs has recently become more prevalent in healthcare applications. An IVA is a talking head animation displayed on a screen which might be accompanied by other speech/video technologies such as Text To Speech (TTS), pre-recorded audio/video and ASR embedded in a form of Spoken Dialogue System (SDS) that conducts conversations with users or provide different services for them (e.g. motivating them to go for a walk).

Applications include use by people with mental health problems [Hayward et al., 2017; Huckvale et al., 2013; Leff et al., 2014; Rus-Calafell et al., 2014], MCI [Morandell et al., 2008], AD [Carrasco et al., 2008; Tran et al., 2016], and the HC [Cyarto et al., 2016]. Nakatani et al. [2018] developed a 3D virtual agent from a photo of a familiar face, such as a family member, to communicate with people with dementia and provide "person centred care". IVAs have been used for *detecting* dementia as well. Tanaka et al. [2017] designed an IVA with spoken dialogue for detecting the early signs of dementia. Although that system was based on standard cognitive tests (MMSE and Wechsler logical memory), in line with our findings, it demonstrated encouraging results for the use and acceptability of an IVA-based, automatic interactional system for patients with memory concerns.

In general, an interface based around *conversation* is often preferred over other modes of interaction with computers (keyboards or touch screens) as it is seen as more natural and easy to use. It is sometimes even preferred over interaction with human; for example, the disclosure of potentially embarrassing information to a computer may be easier than to a human being, especially if the talking head is perceived to be supported by Artificial Intelligent (AI) Rizzo et al. [2016].

7.2 Verbal fluency tests

A verbal fluency test is one of the standard cognitive tests used for assessing people at risk of developing dementia, and comes in two main varieties: *semantic* (naming from a category e.g. animal or fruit) and *phonemic* (naming words beginning with a letter e.g. "P"). Impaired verbal fluency is common amongst people with dementia. For example, people with AD show more deficiency in the 1-minute fluency semantic test comparing

to the 1-minute fluency phonemic test [Canning et al., 2004]. Forbes-McKay et al. [2005] reported that compared to HCs, people with AD i) produced fewer words, ii) tended to use words acquired earlier in life, i.e. words with a lower Age of Acquisition (AoA)[1], iii) use words with a higher occurrence frequency and as well as more typical examples (e.g. thinking of "lion" faster than "kangaroo").

Pakhomov and Hemmy [2014] claimed that the performance of the fluency semantic test is dependent on the efficiency of clustering the related items in a category by the examiner. They used a Latent Semantic Analysis (LSA) approach to automatically determine the category of the words and calculate the mean of semantic clusters for all words, as well as the mean of semantic clusters in the neighbouring words. However, they could not find a significant correlation between their automatic features and the manual scores. Later, they extended the features to the density of repeated words and semantic and lexical diversity. On a longitudinal study of people with dementia and HC, they found that the later features showed a much more significant decline in MCI and AD patients, while they almost stayed the same for HC Pakhomov et al. [2016].

Verbal fluency tests are routinely administered as part of diagnosing and automating this as well as the scoring would free up valuable time for the clinicians.

7.3 Using an IVA to elicit conversation

The initial objective of introducing the IVA was to assess the feasibility of eliciting conversations with people with memory problems. That is, the IVA acted as a neurologist (a *virtual doctor*) and asked similar questions to those asked in a real assessment situation.

The IVA software used for this study was based on the Botlibre[2] library. Only a single IVA was used for this experiment (a head of an adult male character with glasses). Based on feedback from end-users[3], we chose to replace the synthesized speech with recordings of human speech. The IVA had eye movements as well as lip syncing abilities (no other emotional behaviour like smiling, getting excited, etc.). The Botliber uses the image

[1]The age we normally learn a word for the first time.

[2]https://www.botlibre.com

[3]The South Yorkshire Dementia Research Advisory Group (SYDEM RAG) http://sydemrag.group.shef.ac.uk

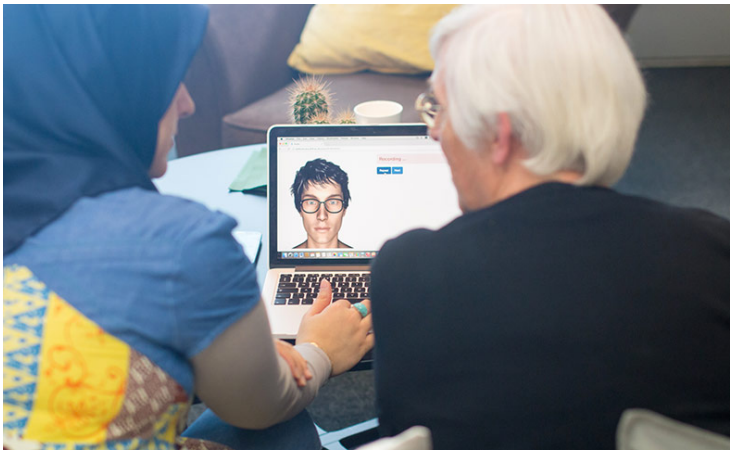


Figure 7.1: *The IVA setup: a laptop was located on a table displaying the IVA to the participants, while a hue camera as well as the laptop's built-in web-cam was video recording the session. Distant microphones on the table were used for audio recording.*

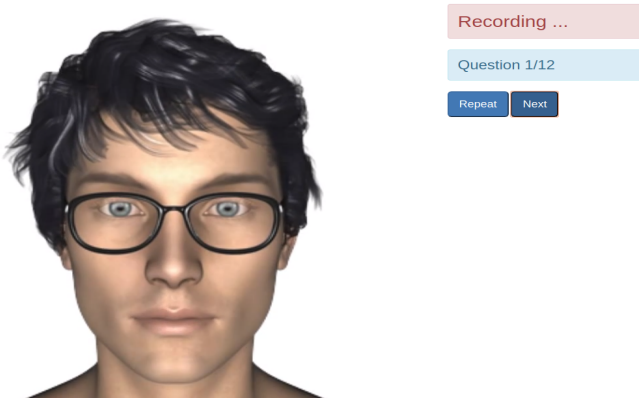


Figure 7.2: *The IVA acting as a neurologist. The web page plays a question and the patient can listen again by pressing the 'repeat' button (or 'space bar' key) or pressing the 'next' button (or 'enter' key) to move to the next question.*

replacing technique to simulate the eye movement and lip syncing. We did not also use any dialogue system techniques to provide feedback to the subjects based on their responds to the questions (i.e. it could not make the interactional conversation with the subjects).

The IVA asks a question when the participant clicks on a button. Since the participants were mostly elderly, who were less familiar with computers and the use of a PC mouse, as a further simplification, they were directed to use just two keys: 'enter' (play) and 'space bar' (repeat). A laptop was used to run the IVA application. The audio was recorded using a distant microphone, Tascam^{TM} DR-40, which was placed on the table, as well as

two other microphones close, attached to the subjects. The laptop's built-in camera was used for video capturing the participant's face (and/or that of the accompanying person) from the front, while another camera (Hue) was located on a table near to participants, and recorded the session from a side angle. This allowed us to capture extra movements of the patient which may not be captured by the front camera (see Figure 7.1). In the current study, we have not used the videos. Figure 7.2 shows a screen-shoot of the IVA.

The participants were asked a number of conversational questions and encouraged to take part in two 1-minute verbal fluency tests (*"name as many animals..."*, and *"name as many words starting with the letter P..."*. Data were collected in three summers: 2016, 2017, and 2018 (we refer to these as IVA2016, IVA2017 and IVA2018 datasets). Table 7.1 shows the questions and fluency test prompting. Note that these questions are very similar to those being asked by the neurologists in our initial data set (the Hal data set). The two verbal fluency tests are the standard screening tests known as the 'fluency semantic' and 'fluency phonemic' tests.

Note that we updated some of the questions/tests. For example, question Q1 changed slightly in 2018 since we were interested in adding a new group of participants, HC and they were not being recorded in the memory clinic, hence a more broad compound question was used. The question Q3 was divided into two separate questions in summer 2017, because the original question was a very long compound question which proved to be confusing for the participants in summer 2016. The phrase "Please give as much detail as you can" at the end of the question Q4 was omitted, since it triggered the participants to provide very long responses to the question. The test questions T1 and T2 were removed from the list in 2017 since the answers provided by the participants in 2016 were not very useful.

7.3.1 IVA datasets

The IVA has enabled us to evaluate our dementia detection system in a real clinical setting, and we have collected data during the summers of 2016, 2017 and 2018. The data was collected by three MSc students at the Department of Neurology, University of Sheffield based at the Royal Hallamshire Hospital.

Table 7.1: *Conversational questions/verbal fluency tests asked of the participants in three summers: 2016, 2017, and 2018.*

	2016	2017	2018	Question
Q1	X	X	-	Why have you come today and what are your expectations?
	-	-	X	Where have you come in from today and what are you hoping to find out?
Q2	X	X	X	Tell me what problems you have noticed with your memory?
Q3	X	-	-	Who is most worried about your memory, you or someone else? What did you do over last weekend, giving much details as you can?
	-	X	X	Who is most worried about your memory, you or someone else?
	-	X	X	What did you do over last weekend, giving much details as you can?
Q4	X	-	-	What has been in the news recently? Please give as much detail as you can.
	-	X	X	What has been in the news recently?
Q5	X	X	X	Tell me about the school you went to and how old were you when you left
Q6	X	X	X	Tell me what you did when you left school, what jobs did you do?
Q7	X	X	X	Tell me about your last job, give as much detail as you can.
Q8	X	X	X	Who manages your finances you or someone else? Has this changed recently?
T1	X	-	-	How well do you think your memory is performing compared to other people your age?, Please select one of the following options A: much better, B: slightly better, C: the same, D: slightly worse or E: much worse.
T2	X	-	-	How well do you think your memory is performing now compared to how it performed five to ten yeas ago? Please select one of the following options A: much better, B: slightly better, C: the same, D: slightly worse or E: much worse.
Phonemic test	X	X	X	Please name as many animals as you can in one minute, you can name any animal, you may now begin.
Semantic test	X	X	X	Please name as many words as you can that begin with the letter P. It can be any word beginning with P except for names of people such as Peter or names of countries such as Portugal. You have one minute and you may now begin.

A total number of 78 participants were recorded in the IVA2016/2017/2018 datasets: 24 in 2016, 21 in 2017, and 33 in 2018. Table 7.2 shows the number of the participants per year and diagnostic class. In addition to the two original diagnostic classes (FMD and ND), an additional class was introduced: the Mild Cognitive Impairment (MCI). They are patients who may develop dementia in the future, although a considerable number of people with the MCI might get better or stay the same (see **Chapter 2**). Some of the participants were diagnosed as having other memory difficulties and are labelled as the 'Rest' class in the table. Also the IVA2018 included a HC group. It is worth mentioning that the majority of the studies of automatic detection of dementia include either two diagnostic classes: ND (such as Alzheimer's Disease) vs. HC, or three: ND vs. MCI vs. HC (see **Chapter 3**). We will include the HC group in the next chapter (the final evaluation).

Table 7.2: *The number of participants in the IVA2016/2017/2018 datasets with the diagnostic classes. FMD: Functional Memory Disorder, ND: Neuro-degenerative Disorder, MCI: Mild Cognitive Impairment, HC: Healthy Control.*

	FMD	ND	MCI	HC	Rest	Total
IVA2016	6	6	6	0	6	24
IVA2017	0	6	8	0	7	21
IVA2018	5	7	4	15	2	33
Sum	11	19	18	15	15	78

Tables 7.3 and 7.4 show the demographic information of the participants of the IVA2016 dataset and the IVA2017/2018, respectively.

Table 7.3: *Demographic information of the participants in the IVA2016 dataset.*

	FMD (n=6)	ND (n=6)	MCI (n=6)
Age	55.7 (+/-8.94)	65.8 (+/-10.38)	63.3 (+/-8.96)
Female	16.7%	33.3%	33.3%

Table 7.4: *Demographic information of the participants in the IVA2017/2018 datasets.*

	FMD (n=5)	ND (n=13)	MCI (n=12)	HC (n=15)
Age	54.6 (+/-2.70)	71.8 (+/-6.99)	61.6 (+/-10.13)	69.5 (+/-7.95)
Female	100.0%	38.4%	33.3%	60.0%

7.4 Results

In this chapter, we only focus on the IVA2016 dataset, and the data collected in the IVA2017/2018 datasets will be used for final evaluation of our dementia detection system in **Chapter 8**.

The recordings of the 18 conversations between the IVA and the participants (the IVA2016) were given to our dementia detection system (6 FMD, 6 ND, and 6 MCI). The automatic transcript and segmentation had 11% DER and 59% WER. The k-fold, $k = 5$, cross validation is used for training the classifiers in this section.

The outputs from the manual transcript and the automatic transcript plus segmentation were then given to the feature extraction and the classifier components of the dementia detection system. Out of the 99 combined features introduced in **Chapter 6**, 27 feature were extracted from the speech of the neurologists. These features were removed from the combined features, since the IVA (who acted as a neurologist) played pre-recorded phrases to ask the same questions of the participants. Therefore a total number of 72 features were extracted by the feature extraction module of the dementia detection system to train the LR classifier to classify between the three patient groups: the FMD, the ND and the MCI.

Table 7.5: *Accuracy of the LR classifier to classify between different patient groups: FMD/ND/MCI, FMD/ND, FMD/MCI, and ND/MCI using the 72 combined features.*

Level of automation	FMD/ND/MCI	FMD/ND	FMD/MCI	ND/MCI
Manual Transcript	61.1%	**91.7%**	**58.3%**	83.3%
Auto transcript+segmentation	**66.7%**	83.3%	33.3%	**91.7%**

Table 7.5 shows the classification accuracy using the 72 features for the two levels of automation of transcript and segmentation: manual transcript, and automatic tran-

script+segmentation (i.e. using both the Diar and the ASR). The results are for the three-way classifier (FMD/ND/MCI) as well as the three binary classifications for the patient groups FMD/ND, FMD/MCI, and ND/MCI respectively. Since an extra class (MCI) has been introduces, this will make the classification task harder. Thus it is expected that the classifier accuracy drops considerably.

For the manual transcript, an accuracy of 61% was achieved by the three-way classifier. The accuracy of the FMD/ND and the ND/MCI classifiers were 92% and 83% respectively, while the FMD/MCI classifier achieved only 58% accuracy. This indicates that there were much overlaps between FMD and MCI groups and it was the hardest task in comparison to the other binary classifiers.

For the auto transcript+segmentation, comparing to the manual transcript results, the accuracy of the three-way classifier was slightly better than for the manual transcript (67% vs. 61%). The accuracy of the FMD/ND changed from 92% to 83%, and for the ND/MCI from 83% to 92%. However, the accuracy of the FMD/MCI classifier dropped considerably to 33% (a 24% decrease compared to the manual transcript). This indicates that there was a trade off between the accuracy of the binary classifiers, i.e. increasing accuracy of one of them resulted in decreasing the others.

7.4.1 Confusion matrix

Table 7.3 shows the confusion matrix for the 3-way classifier (Auto transcript+segmentation). Among the six samples of the FMD group, 4 (67%) were identified truly by the classifier, while one confused as the ND and one as the MCI. However, all the six ND were predicted correctly (100%, no confusion at all), while only two MCI were predicted truly (3 were identified as FMD and 1 as ND by mistake). This shows that identifying the MCI had the highest confusion among the three patient groups and there were mostly confused by the FMD patients. The high confusion between the MCI and the FMD can be seen also in the Table 7.5 (row 3, column 4), where the binary classifier only had 33% accuracy rate. On the whole identifying the ND group was the easier task followed by the FMD. The hardest group to identify correctly was the MCI group.

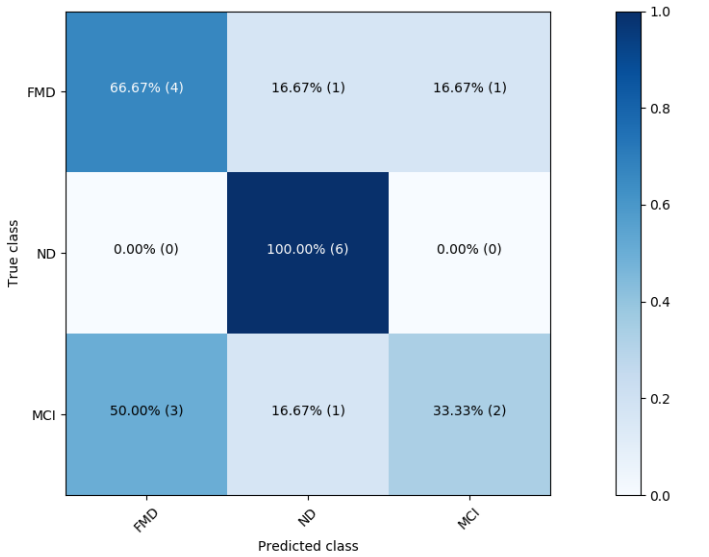


Figure 7.3: *Confusion matrix for the 3-way classifier (Auto transcript+segmentation).*

7.4.2 Feature selection

Using the RFE feature selection approach (we applied the k-fold cross validation (k=5), and the RFE was performed on the trained set.), which chooses first a feature with the highest contribution in classification and then recursively find the next feature with the highest contribution in the remaining feature set, until the desired number of features are selected. Using this approach, the top n features for the two levels of automation were chosen. The classification results are summarised in the Table 7.6. A similar approach as described in **Chapter 6** was employed to find the n.

Table 7.6: *Accuracy of the LR classifier to classify between different patient groups: FMD/ND/MCI, FMD/ND, FMD/MCI, and ND/MCI using top features for the three systems. NT: Number of top features.*

Level of automation	NT	FMD/ND/MCI	FMD/ND	FMD/MCI	ND/MCI
Manual Transcript	11	**66.7%**	83.3%	**83.3%**	**91.7%**
Auto transcript+segmentation	5	**66.7%**	**100.0%**	66.7%	25.0%

The top 11 features for the manual transcript system could improve the accuracy of the three-way classifier from 61% to 67%. This could also improve the accuracy of both the FMD/MCI (from 58% to 83%) and the ND/MCI (from 83% to 92%) classifiers. For the auto transcript+segmentation, using the 5 top features, the three-way classifier stayed

the same at 67%, while the accuracy of the FMD/ND classifier achieved 100%, and the accuracy of the FMD/MCI almost doubled (from 33% to 67%), however, the accuracy of the ND/MCI classifier drastically decreased from 92% to 25%.

This shows that feature selection by the RFE algorithm might find a subset of features which result in an overall (3-way) good classification rate, but at the price of having a worse accuracy for the binary classes. So more care has to be taken when carrying out-e.g. using different feature selection algorithm.

Table 7.7 shows the lists of the top 11 features for the manual transcript as well as the top 5 features for the auto transcript+segmentation. Surprisingly the word vector features had the highest ranks in the lists, followed by the acoustic features. There was only one lexical feature in the top features (average number of words for patient, PatAVAllWords). Also there was not any features associated with the accompanying person in the top feature lists. It is worth mentioning that only 3 out of 18 participants had APs, which might be the reason why none of the features extracted for the APs were in the top features.

Table 7.7: *Top features the RFE for the manual transcript and the Auto transcript+segmentation. WV_col1: The word vector features column 1. AC: acoustic features, LX: lexical features, and WV: word vector features.*

Rank	**Manual Transcript**	**Auto transcript+segmentation**
1	WV_col6(WV)	WV_col5(WV)
2	WV_col2(WV)	WV_col7(WV)
3	WV_col3(WV)	PatAVPauses(AC)
4	WV_col1(WV)	PatAvgH1-H2(AC)
5	WV_col5(WV)	PatAvgH1-A1(AC)
6	WV_col4(WV)	
7	PatAVAllWords(LX)	
8	PatAVTurnLength(AC)	
9	PatAvgH1-A2(AC)	
10	PatNoOfTurns(AC)	
11	PatAvgIntonation(AC)	

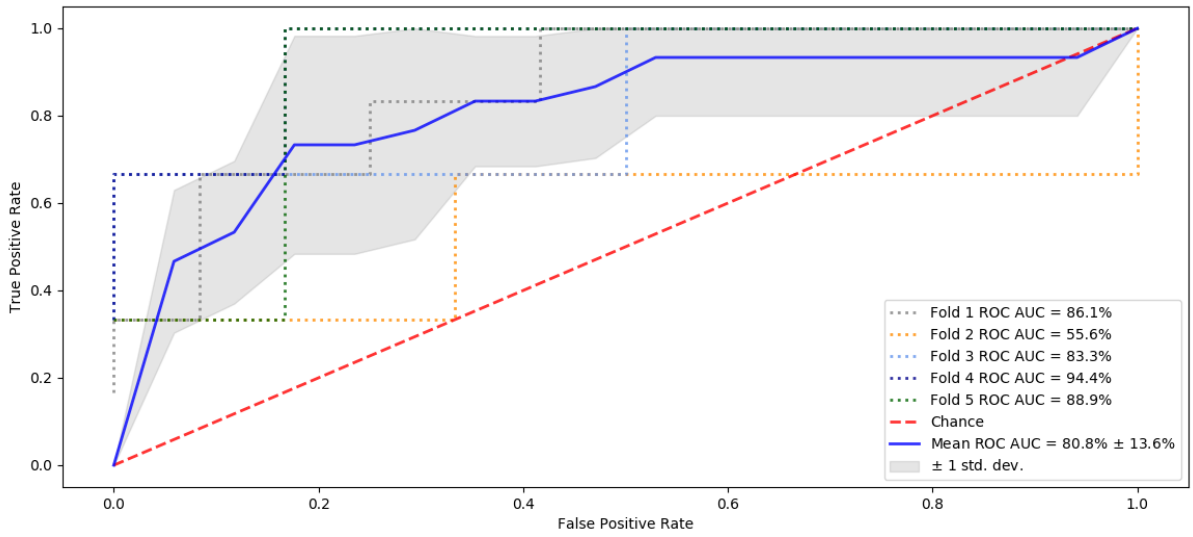


Figure 7.4: *The ROC curve for the FMD/ND/MCI classifier using the 72 features for Diar+ASR.*

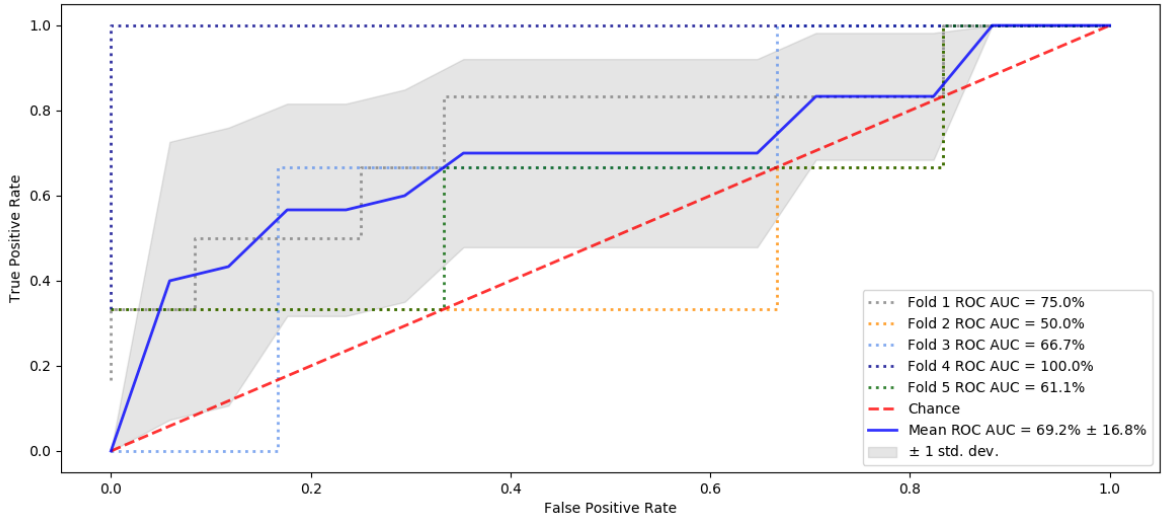


Figure 7.5: *The ROC curve for the FMD/ND/MCI classifier using the top 5 features for Diar+ASR.*

7.4.3 The Receiver Operating Characteristic curve

In this section, similarly to the previous chapter, we will use the ROC curve analysis in order to show how robust the three-way classifier for the transcript+segmentation was.

The ROC curves for the three-way classifier using the 72 features and the most significant (5) features are shown in Figures 7.4 and 7.5 respectively. For the former, the average ROC-AUC was 81% (k-fold, $k = 5$). For the latter, however, the average ROC-AUC dropped to 69%. This shows that the three-way classifier trained on the 72 features was much more robust than the classifier trained on the most significant (5) features despite the fact that both classifiers had 67% accuracy.

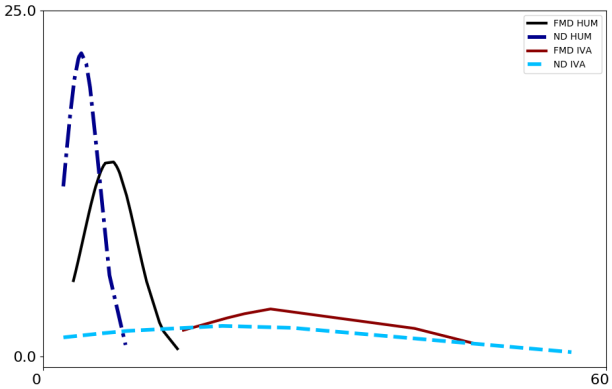
7.4.4 Comparing neurologist-led to IVA-led conversations

In this section, we compare the FMD/ND conversations in the IVA2016 dataset (IVA-led conversations, referred to as IVA in the following) with the human-led conversations in the Hallamshire dataset (HUM in the following). Out of the original 99 features, 27 neurologist-associated features were removed and using the k-fold, $k = 5$, the classification experiments repeated. The classifier accuracy for the two levels of automation of transcript and segmentation are listed in Table 7.8. To make it easier, we repeated the accuracy of the FMD/ND classifier for the IVA in the last column. Note that the HUM comprises of 15 FMD and 15 ND participants, and the IVA comprises of 6 FMD and 6 ND participants.

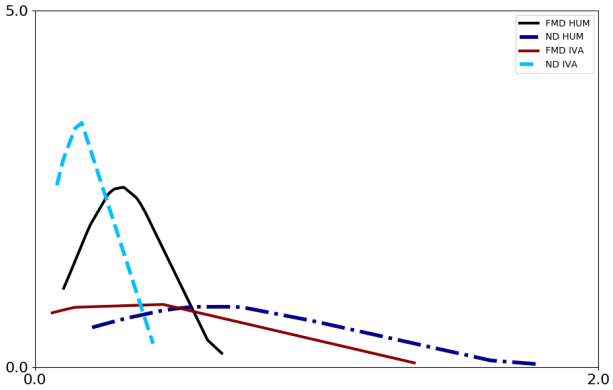
Table 7.8: *Accuracy of the FMD/ND classifier for the human-lead (HUM) and the IVA-led (IVA) conversations.*

Level of automation	HUM	IVA
Manual Transcript	76.7%	**91.7%**
Auto transcript+segmentation	**83.3%**	**83.3%**

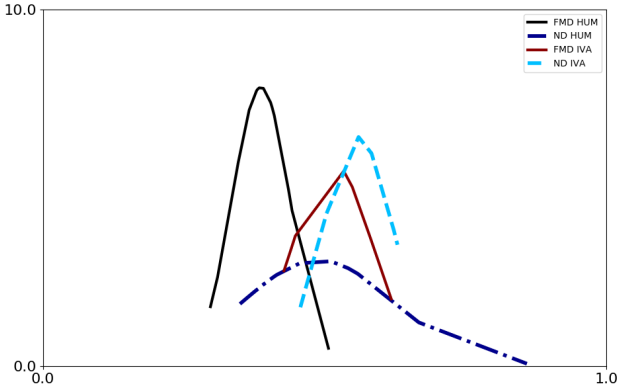
For the manual transcript, the accuracy achieved by the classifier for the HUM and the IVA conversations were 77% and 92% respectively. However for the auto transcript+segmentation, the accuracy of both classifiers were the same at 83%. Using the statistical test (Student's T test) no significant difference was found between the accuracy of the classifier obtained for the HUM and the IVA conversations.



(a)



(b)



(c)

Figure 7.6: (a) Distribution of the average turn length (in seconds). (b) Distribution of the average silence (in seconds). (c) Distribution of the average overall duration (in seconds).

Some differences were observed between the IVA and HUM conversations. Figure 7.6 shows three measures plotted for the ND and FMD groups for the IVA and HUM respectively. Looking at the distribution of the average length of the turns (Figure 7.6(a)), in both datasets, the patients speaking to the neurologist had shorter turns than when speaking to IVA. However, overall the IVA conversations had much longer turns, which is likely to be related to the fact that this prototype IVA provides no feedback to the patients in the form of nods, clarifying questions or back-channel noises to steer the conversation. As a result, some patients chose to give very lengthy responses to some of the questions.

The average silence plotted in Figure 7.6(b) shows a different picture. The least silence is observed for the ND-IVA group and the most for the ND-HUM group. This may be a result of the neurologists being instructed to wait much longer than would normally be expected for the patients to provide an answer. When working with the IVA, the patients always had the option of clicking 'next' and moving the IVA on to the next question. This suggests that many chose to take this option quite readily when they were unable to give a satisfactory answer.

Finally, Figure 7.6(c) shows the average total duration of the conversation. Despite the average turn length (a), appearing to be quite a discriminative measure, this is less clearly so.

7.5 Discussion

The accuracy of the three-way classifier using the manual transcript and the automatic transcript and segmentation were 61% and 67% respectively, this indicates that despite the errors caused by the Diar and the ASR, the 72 extracted features were robust when classifying between the three classes with an acceptable level of accuracy (the chance level accuracy is 33%), although, the discrimination for the FMD/MCI was harder (the accuracy dropped from 58% to 33%) due to the high amount of overlaps between the features' distribution for the two patient groups (i.e. the two groups do not share the common issues seen in the ND patients, like struggling to remember, long pauses, not able to answer all parts of questions, etc.). In clinical situations, similarly discriminating

between FMD and MCI patients is not an easy task.

The most significant features for the manual transcript could improve the accuracy of the three-way classifier from 61% to 67% while the accuracy of the binary classifiers FMD/MCI and ND/MCI improved remarkably. However, for the automatic transcript and segmentation, the most significant features resulted in the same accuracy for the three-way classifier. The FMD/ND classifier achieved 100% accuracy while the accuracy of the other two binary classifiers decreased considerably. This indicates that the classifiers trained on the most significant features for the automated system were not as robust as the classifiers trained on the 72 features.

The features extracted for the APs were not among the most important features. It might be due to the fact that in the experiment, only 3 out of 18 participants had APs and all of them were in the ND group (i.e. 50% of ND groups compared to 0% of FMD and 0% of MCI). Since there were not many APs involved in the experiment, their features could not be in the top features.

Looking at the confusion matrix, all the ND patients were predicted correctly, while 67% of FMD were identified truly and only 33% of MCI could be identified without confusion. Most of the confusion were between the MCI and the FMD.

Comparing the ROC-AUC of the three-way classifiers for the automatic transcript and segmentation using the 72 features and the most significant (5) features confirms that the former classifier was more robust than the latter (with 82% ROC-AUC vs. 69% ROC-AUC).

In a real situation, doctors are faced with assessing many different patient groups and due to the high degree of overlap between the symptoms of the different memory disorders, making a final decision about a patient's diagnosis is a challenging and complex task. On the whole, adding the MCI group caused the classification to become harder, which decreased the accuracy of the classifier. The binary classifier MCI/FMD was the hardest classification of the three binary classifications. Diagnosing between the two groups is hard for clinicians as well.

Comparing the results of the IVA2016 dataset (the IVA-led conversations) to the Hallamshire dataset (human-led conversations), we managed to somewhat improve the

accuracy in the IVA-led conversations using the manual transcript. However, using the automatic transcription and segmentation the same accuracy was achieved for both datasets. There were differences between the two types of conversations in the length of turns, silences and duration, however, these differences were mostly due the fact that the participants knew they are talking to a virtual agent and they could themselves control the conversation by clicking the 'next' button.

7.6 Summary

In this chapter, we introduced an IVA to our dementia detection system. The IVA asked similar questions of the patients as the questions in the Hallamshire dataset. Using the IVA we collected data during three summers (2016, 2017 and 2018).

In this chapter we focused on the IVA2016 dataset containing a total number of 18 conversations from three different patient groups (FMD, ND and MCI). The three-way classifier accuracy was 67% for the fully automated system (automatic transcription and segmentation) using the 72 features introduced in the previous chapter. Although the accuracy of the two binary classifiers the FMD/ND, and the ND/MCI achieved 83% and 92% respectively.

Using the most significant (5) features the accuracy of the three-way classifier did not improve, however, the accuracy of the FMD/ND and the FMD/MCI increased from 83% and 33% to 100% and 67% respectively, whilst the accuracy of the ND/MCI dropped drastically from 92% to 25%. This confirms that there were trade offs between the accuracy of the three binary classifiers, increasing one classifier reduced the accuracy of the others.

The ROC curve analysis also confirmed the difficulty of the three-way classification.

Comparing the IVA-led conversations (the IVA2016) to the human-led conversations (the Hallamshire dataset) using the 72 features on manual transcripts, the binary classifier FMD/ND of the IVA-led conversations achieved 92% accuracy, which was 15% better than the human-led conversations with 77% accuracy, however, both achieved the same accuracy on the automatic transcript and segmentation (83%). Despite some differences between the IVA-led and the human-led conversations (e.g. in the length of turns and silences), the IVA-led conversations could be used successfully to discriminate between the two patient groups similar to the human-led conversations.

Chapter 8

Final evaluation

Contents

8.1 Introduction 161

8.2 Final results 162

8.2.1 Effect of adding the healthy control group 162

8.2.2 Processing the verbal fluency tests 164

8.2.3 Combining the conversations with the verbal fluency tests . . . 165

8.2.4 Combining all the IVA datasets 167

8.2.5 Feature selection (RFE) 169

8.2.6 Feature selection (statistically significant) 171

8.2.7 F1-measure 173

8.3 Discussion 175

8.4 Summary 179

In this chapter the conversations collected by the IVA during the summer 2017 and 2018 will be used for the final evaluation of the dementia detection system. The datasets include 8 conversational questions asked by the IVA from the participants, as well as three verbal fluency tests. In addition to the 72 features extracted from the conversational questions, some additional features will be extracted automatically from the first two cognitive test questions. Comparing to the previous chapter, we will have another patient group HC (4 patient groups altogether: ND, FMD, MCI, and HC). This will make the classification task harder but it will reflect a more realistic situations. The chapter is organised as below:

Section 8.1 is an introduction to the chapter.

Section 8.2 presents the results of training a number of classifiers using the conversations and the verbal fluency tests of the IVA datasets.

Section 8.3 and **Section 8.4** include the discussion and the summary of this chapter respectively.

8.1 Introduction

In the previous chapters, our dementia detection system was introduced as well as its individual components. Despite the errors caused by the diarisation and the ASR modules, the classifier trained on the features extracted from the automated transcript could classify with a high accuracy for the two patient groups (FMD vs. ND, see **Chapter 6**) in the Hal dataset. Then we introduced the IVA for conducting the conversations with the participants as well as administering a few standard verbal fluency tests.

As we introduced the MCI as an additional patient group to the classifier, the accuracy decreased considerably due to difficulties of the three-way classification task. More importantly, the complete systems based on the manual transcript and the automated transcript and segmentation, both gave the same accuracy for the three-way and the FMD/ND classifiers. However, this was different for the two other binary classifiers (FMD/MCI and ND/MCI) where a decline in accuracy was observed for the former classifier, but there was an increase for the latter.

In real clinical conditions, there might be a number of patients referred that do not have any dementia-related memory issues. In the literature, many studies about detecting dementia include healthy controls (HCs) which represents this group of referrals. The IVA2017/18 datasets includes HCs. Therefore as a final evaluation of our dementia detection system we will use the IVA2017/18 with four diagnostic groups: FMD, ND, MCI and HC.

The four-way classification is likely to be much harder than the three-way and the binary classification. Thus the classifier would make more mistakes and the overall accuracy would drop drastically.

In addition to processing the conversations of the datasets, as a further step, the two verbal fluency tests (animal naming, and words beginning with letter 'P') will be automatically scored. We examine two ways of using these scores to improve the accuracy of the system, namely: as new features to train the classifier, or by adding them to the 72 features introduced in the previous chapter.

8.2 Final results

So far we mostly used the accuracy of the classifier as a measure to evaluate the performance of our dementia detection system. However as we mentioned in **Chapter 6**, in addition to the accuracy, the high sensitivity and high specificity of a classifier are important. The ROC analysis shows how the classifier sensitivity and specificity change for different parameters and settings of a classifier, i.e. the higher ROC-AUC the better sensitivity and specificity. It is possible to have a classifier with a high accuracy rate but a low ROC-AUC and vice verse. However, it is generally acceptable to rely on a classifier with both high accuracy and high ROC-AUC.

Therefore both the 'accuracy' and the 'ROC-AUC' will be used as the evaluation measures of the final dementia detection system. Similar to the previous chapter, the k-fold ($k = 5$) cross validation approach will be used for training the classifiers.

8.2.1 Effect of adding the healthy control group

As mentioned in **Section 7.3**, the IVA2017/18 datasets include an additional group, HC (see Table 7.4). A total number of 45 (FMD:5, ND:13, MCI:12, and HC:15) conversations were passed to the dementia detection system to extract the 72 features and train a four-way classifier (FMD/ND/MCI/HC). In addition to this classifier, 10 more classifiers were trained with the subset of the participant groups: four three-way classifiers (FMD/ND/MCI, FMD/ND/HC, FMD/MCI/HC and ND/MCI/HC), and six binary classifiers (FMD/ND, FMD/MCI, ND/MCI, FMD/HC, ND/HC, and MCI/HC).

Figure 8.1 shows the accuracy of the 11 classifiers (the green bars) as well as their corresponding ROC-AUC (the grey bars).

The accuracy of the four-way classifier (with the chance level of 25%) was 53% (+/-10% errors) while its ROC-AUC was 74% (+/-7% errors). This indicates that we trained a robust classifier despite the difficulties of the four-way classification task.

Amongst the four three-way classifiers, the classifiers including HC had better results with 73% (+/-11%), 66% (+/-12%), and 63% (+/-11%) for FMD/ND/HC, FMD/MCI/HC, and ND/MCI/HC respectively. They also all had a high ROC-AUC

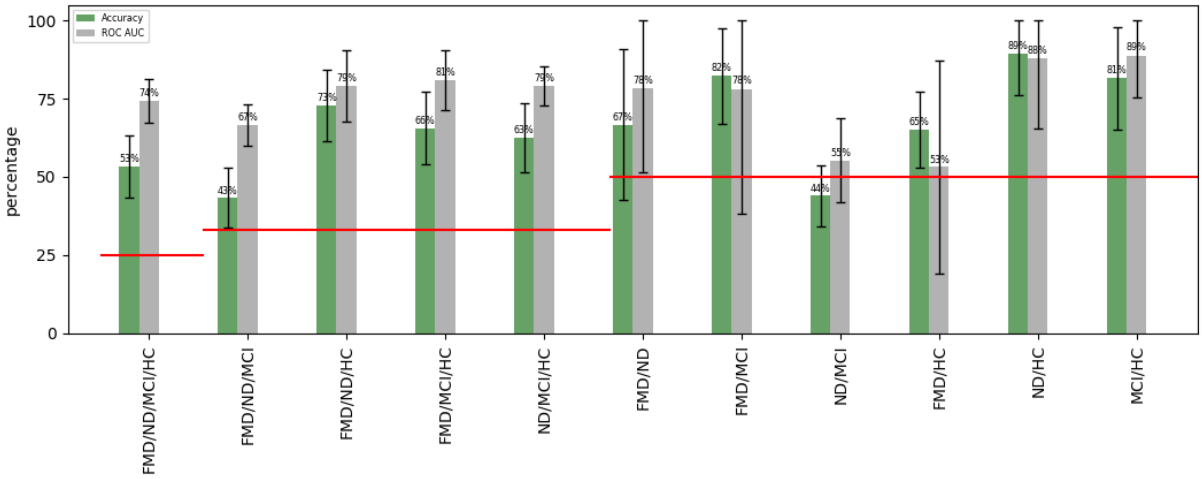


Figure 8.1: *Accuracy and ROC-AUC of the 11 LR classifiers for the conversations of the IVA2017/18 datasets with error bars (red lines: chance levels).*

(over 79% with less than 11% errors). HC is clearly the easiest group to distinguish from the rest. However, the accuracy of the FMD/ND/MCI classifier was low with 43% (+/-10%) (ROC-AUC 67% (+/-7%)). This shows the overlap between the features extracted for these three groups were high, which made it harder for the classifier to identify them correctly. Comparing these to the results gained by the IVA2016 dataset (61% accuracy, 81% ROC-AUC, see Table 7.5 and Figure 7.4), here both the accuracy and the ROC-AUC were considerably lower. This could be due to the imbalanced number of participants and contributions of each gender in the IVA2017/18 datasets comparing to the IVA2016 dataset. There were 6 FMD, 6 ND and 6 MCI in IVA2016 dataset, while the number of FMD, ND and MCI in the IVA2017/18 were 5, 13 and 12 respectively. Also only 16.7% of the FMD participants in the IVA2016 dataset were female, while 100% of the FMD participants were female in the IVA2017/18 datasets (see Table 7.3 and 7.4).

Between the six binary classifiers, the ND/HC classifier had the best classification accuracy (89% (+/-13%)) with a high ROC-AUC (88%, but with also a high error range of 22%), followed by the MCI/HC (accuracy: 81% (+/-16%) ROC-AUC: 89% (+/-13%)). However the lowest accuracy was achieved by the ND/MCI with 44% (+/-10%). This indicates that discrimination of the MCI patients from the ND patients was the hardest task done by the classifiers. Comparing the accuracy of the FMD/ND classifier

with the accuracy of the Hal dataset, the accuracy achieved here was not good (67% vs. 77% see **Section 7.4.4**) with over a 24% error range. The FMD/MCI classifier had the worse error range of 40%[1] for its average ROC-AUC of 78% (i.e. for some folds ROC-AUC was as little as 38%), despite having a relatively high accuracy of 82%.

8.2.2 Processing the verbal fluency tests

In addition to the conversational questions, the IVA2017/18 included the verbal fluency tests ('fluency semantic test' and 'fluency phonemic test' see Table 7.1). The former test shows the ability of remembering words from our semantic and episodic memory, while the latter assesses our phonological awareness skill (reading ability). People with dementia may struggle with their semantic memory and reading ability. In counting the names, the repeated and non-relevant names should be omitted. The count below a threshold (e.g. 14 names in a minute) may indicate an issue with memory.

The number of correctly produced names as well as the average and the standard deviation of the AoA for the words are produced automatically (from the outputs of the ASR). The language model of the ASR (n-gram language model) used here was trained on a general list of animals and words beginning with the letter 'P'. Table 8.1 describes details of the features extracted from the verbal fluency tests.

The six verbal fluency tests' features extracted from the IVA2017/18 datasets were given to the 11 LR classifiers. Figure 8.2 shows the accuracy and the ROC-AUC for the classifiers trained on the features.

The accuracy of the four-way classifier was 43% (+/-10%), which was 10% lower than the four-way classifier trained on the conversations of the IVA2017/18 datasets, however, the ROC-AUC was above 70% (+/-9%), which shows a relatively robust classification.

The four three-way classifiers all achieved the accuracy rates between 51% and 58% and the ROC-AUC between 67% and 73%, although the error range for accuracy of FMD/ND/HC was high (24%). In particular the FMD/ND/MCI classifier achieved better results in comparison to the classifier for the conversations of the IVA2017/18 datasets

[1]when the error range of a ROC-AUC of a classifier is high, it is hard to claim that the classifier is robust despite having a high average value for ROC-AUC.

Table 8.1: *verbal fluency tests' features.*

No.	Feature	Description
1	PatSemCount	Number of unique animals correctly uttered in the fluency semantic test.
2	PatSemAVGAoA	Average AoA for the fluency semantic test.
3	PatSemSTDAoA	Standard deviation of the AoA for the fluency semantic test.
4	PatPhnCount	Number of unique words correctly uttered in the fluency phonemic test.
5	PatPhnAVGAoA	Average AoA for the fluency phonemic test.
6	PatPhnSTDAoA	Standard deviation of the AoA for the fluency phonemic test.

(accuracy: 53% vs. 43% , and ROC-AUC: 73% vs. 67%).

The best accuracy for the binary classifiers was achieved by the FMD/ND classifier with 78% (but with a 25% error range) as well as the highest ROC-AUC (91% (+/-13%)). This was better than the classifier for the conversations, and comparable with the accuracy of the Hal dataset (78% vs. 77%). However the lowest ROC-AUC (22%) with a high error range of 27% was achieved by the FMD/HC classifier as well as the lowest accuracy (53% (+/-13%)). This indicates that the verbal fluency tests were not very successful in distinguishing between the FMD and the HC groups. Also the FMD/MCI classifier with a high accuracy of 77% (+/-12%) had the worse error range of 37% for ROC-AUC. Thus distinguishing between both the FMD and the MCI participants and the FMD and the HC participants were not easy tasks.

8.2.3 Combining the conversations with the verbal fluency tests

Since the results of the classifiers trained with the six cognitive fluency test features were comparable with the results of the classifiers trained on the conversations, we combined the fluency features with the conversation features (78 features in total). Figure 8.3 shows the accuracy and the ROC-AUC of the classifiers trained on the combined 78 features.

The overall accuracy and ROC-AUC of the classifiers were considerably higher. The four-way classifier achieved 59% (+/-10%) accuracy, which was 6% better than the classi-

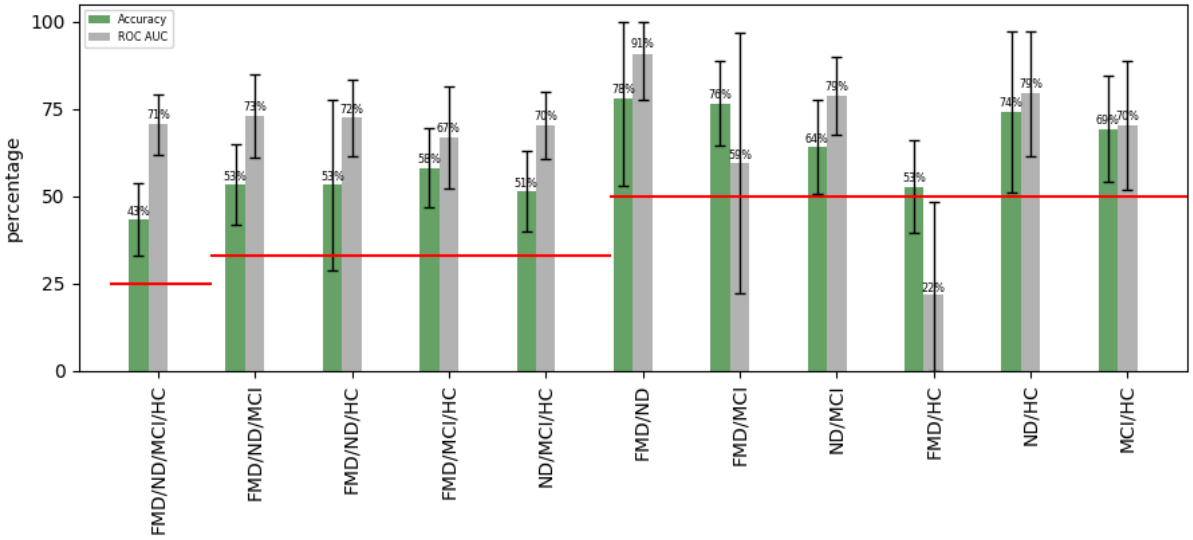


Figure 8.2: *Accuracy and ROC-AUC of the 11 LR classifiers using the verbal fluency tests' features of the IVA2017/18 datasets (red lines: chance levels).*

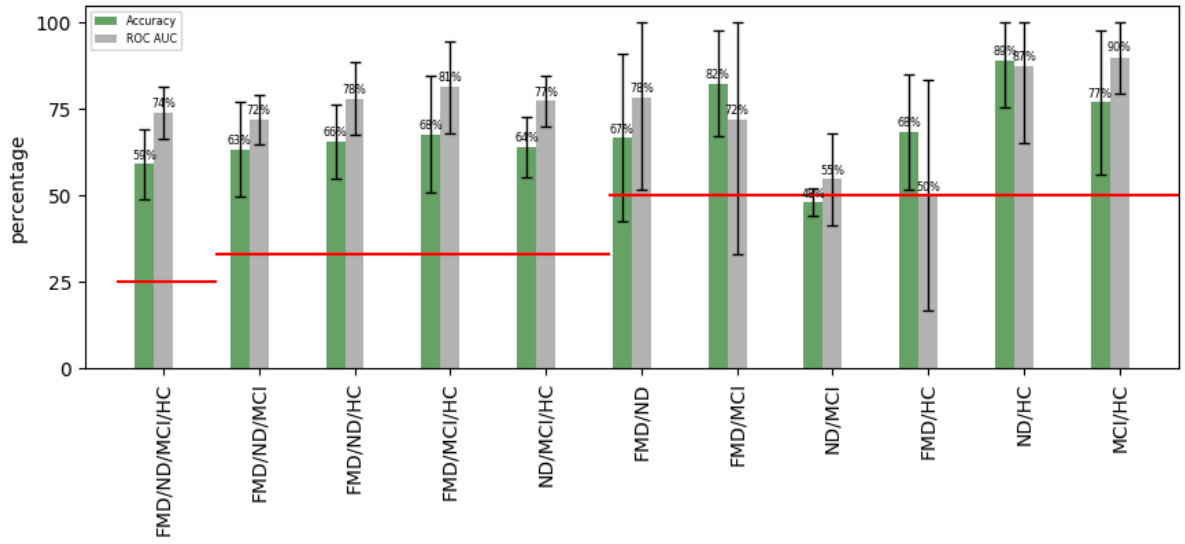


Figure 8.3: *Accuracy and ROC-AUC of the 11 LR classifiers using both the conversations and the verbal fluency tests of the IVA2017/18 datasets (red lines: chance levels).*

fier trained on the conversations only. The ROC-AUC for both the classifiers were almost the same (74%).

Similarly, the tree-way classifiers included the HC group had higher accuracy and ROC-AUC (accuracy: between 64% and 68% (error range between 9% and 17%), ROC-AUC: between 77% and 81% (error range between 7% and 13%)). The FMD/ND/MCI classifier achieved 63% accuracy (20% better than for the conversation-only) and 72% ROC-AUC (5% better than for the conversation-only).

The best accuracy for the binary classifiers was achieved by the ND/HC classifier with 89% accuracy and 87% ROC-AUC (almost identical to the classifier of the conversation-only). The FMD/ND achieved the same accuracy and ROC-AUC. The lowest accuracy was gained by the ND/MCI with 48% and the lowest ROC-AUC by the FMD/HC with 50% and 33% of error range.

8.2.4 Combining all the IVA datasets

Finally, we combined the IVA2016 and IVA2017/18 datasets and extracted the 78 features. Note that there was a fluency test missing from the IVA2016 (one FMD participant), therefore the combined datasets included 61 samples (FMD=10, ND=19, MCI=18, and HC=14). For training the classifiers the k-fold ($k = 10$) cross validation approach was applied. Figure 8.4 shows the accuracy and the ROC-AUC for the 11 classifiers trained on the features extracted from the combined datasets IVA2016/17/18.

Generally, comparing to the results of IVA2017/18 datasets, the accuracy of all classifiers were slightly worse. Note that there were not any HC participants in the IVA2016 dataset, while 15 out of 45 participants (33%) of the IVA2017/18 datasets were HC. However, in the combined datasets, there were fewer percentage (15 out of 61 or 25%) of the HC participants, who can be identified much easier than the other clinical categories.

The four-way classifier achieved 48% accuracy (5% less than for the IVA2017/18 datasets) but also with a higher error range of 25%, also with a 70% ROC-AUC (4% less than for the IVA2017/18 datasets).

For the three-way classifiers, the accuracy of the FMD/ND/MCI classifier decreased from 63% to 49%, and its ROC-AUC dropped from 72% to 70%. The accuracy of the

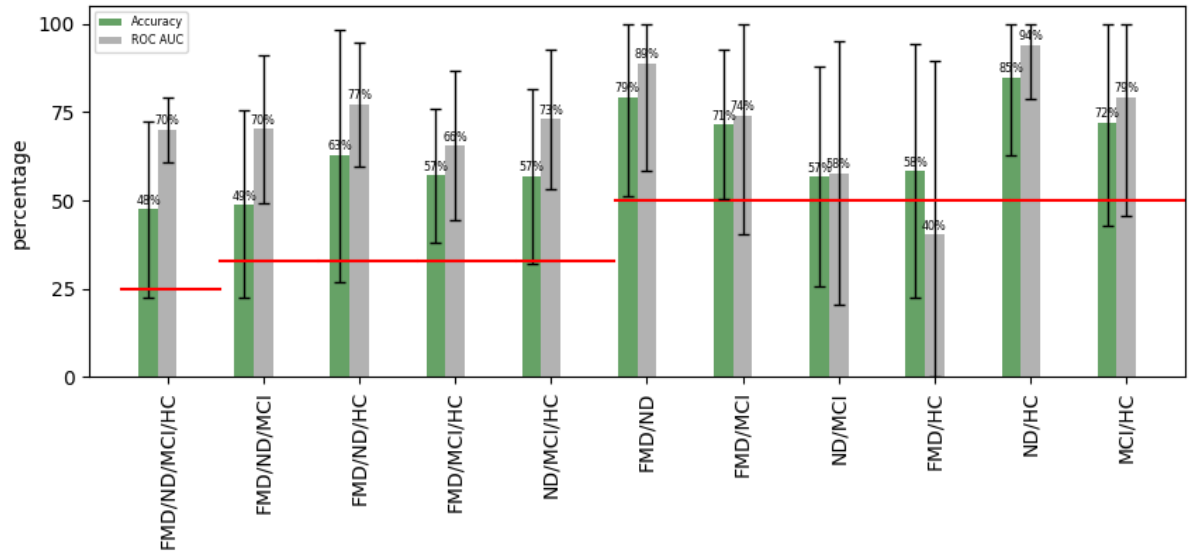


Figure 8.4: *Accuracy and ROC-AUC of the 11 LR classifiers for the IVA2016/17/18 datasets (78 features) (red lines: chance levels).*

classifiers including the HC also declined considerably. The error ranges are also much higher than for the IVA2017/18 datasets.

Amongst the 6 binary classifiers the ND/HC and the FMD/ND classifiers achieved the highest accuracy, 85% and 79% respectively. The accuracy achieved by the FMD/ND achieved a slightly better accuracy comparing to the classifier for the Hal dataset (79% vs. 77%). The worse error range of ROC-AUC belonged to FMD/HC which again confirms the difficulty of training a robust classifier distinguishing between these two clinical categories.

Figure 8.5 shows the confusion matrix for the four-way classifier (IVA2016/17/18 datasets). The main diagonal of the matrix shows the percentage of the correct classification, and the entries outside of the main diagonal shows the confusion between the predicted class and the true class. As it was expected, most of the HC recordings (71%) were identified correctly. The rest were confused with the FMD class (29%).

For the ND class, 58% of the patients were identified correctly, however, 32% were misclassified as MCI and 11% as FMD. Only 33% of MCI patients are classified correctly, while confusion with ND was 33% and with HC 28%. The FMD patients are mostly confused by HCs with 60% (the highest confusion) and only 20% of FMD patients were

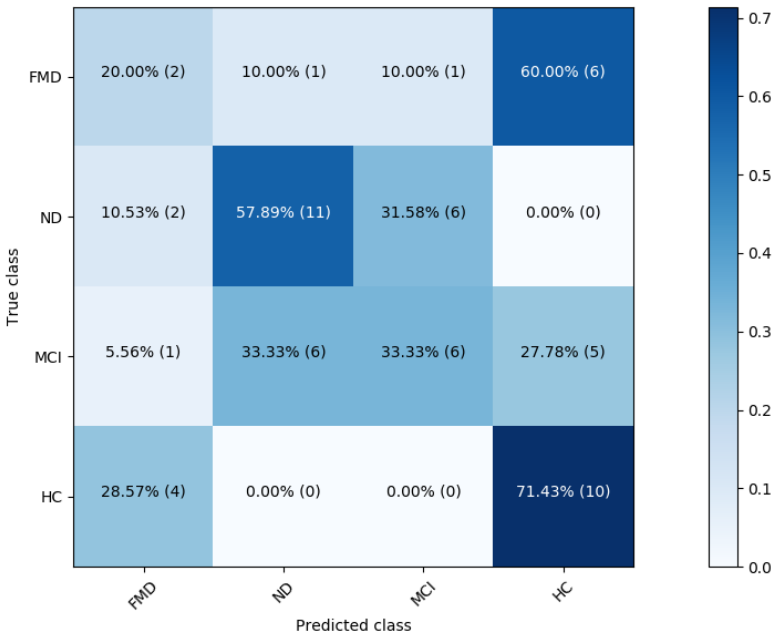


Figure 8.5: *Confusion matrix for the four-way classifier (IVA2016/17/18 datasets).*

correctly identified. This indicates that identifying HC was the easiest task by the classifier followed by the ND group. The distributions of the features extracted from the FMD group might be close to the distributions of the features extracted from the HC group and also there were similarities between the MCI and the ND groups, as we know that in clinics the MCI and FMD patients have common symptoms with the ND patients.

8.2.5 Feature selection (RFE)

Using the RFE feature selection methodology (on the train set) and the approach introduced in **Section 6.2.3**, the 22 most significant features were selected out of the 78 features for the IVA2016/17/18 datasets.

The accuracy of the four-way classifier was 62% (+/-21%) and the ROC-AUC 82% (+/-15%). For this classification task, the trained classifier could be considered as a robust classifier. Comparing to the classifier trained on 78 features, both the accuracy and the ROC-AUC improved considerably.

All the four three-way classifiers achieved similar or higher accuracy and ROC-AUC than those trained by all the features. More importantly the FMD/ND/MCI achieved

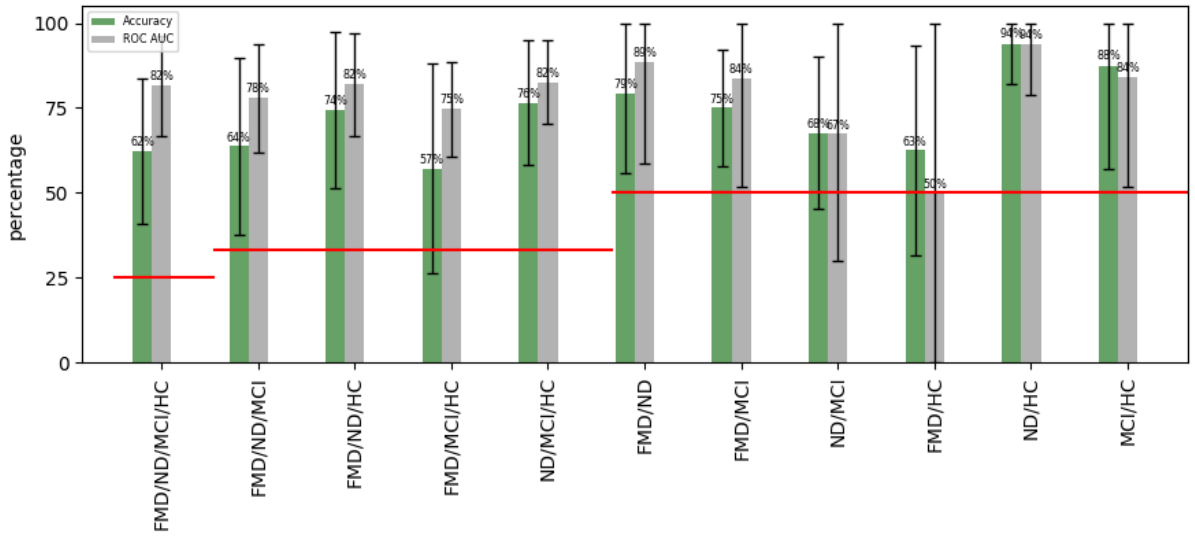


Figure 8.6: *Accuracy and ROC-AUC of the 11 LR classifiers for the IVA2016/17/18 datasets (the most significant features (22)) (red lines: chance levels).*

64% accuracy but with 26% of error range.

The accuracy and ROC-AUC of the six binary classifiers remained the same or improved significantly. The ND/HC classifier had the best accuracy of 94% (+/-12%), followed by the MCI/HC with 88% (+/-30%) accuracy (with high ROC-AUC of 94% (+/-15%) and 84% (+/-32%) respectively). The FMD/ND classifier achieved 79% (+/-24%) accuracy and 89% (+/-30%) ROC-AUC, which was better than the classifier trained on the IVA2017/18 datasets and the Hal dataset. Ironically the worse error bar of ROC-AUC was for FMD/HC.

Figure 8.7 shows the confusion matrix for the four-way classifier using the most significant features. Over 86% of HC and 74% of ND patient were classified correctly, while the rate of correct classification for MCI and FMD were 44% and 40% respectively. Again the highest confusion was between FMD and HC with 40% confusion. The confusion matrix confirms that identifying ND from HC was the easiest task by the classifier, while identifying FMD and MCI were the hardest (more confusing) tasks.

Table 8.2 lists 22 most significant features using the RFE approach (on IVA2016/17/18 dataset on the train set). Among these 22 features, 7 were acoustic, 6 were word vectors, 4 were fluency tests, 3 were lexical and 2 were semantic. Therefore the acoustic features and

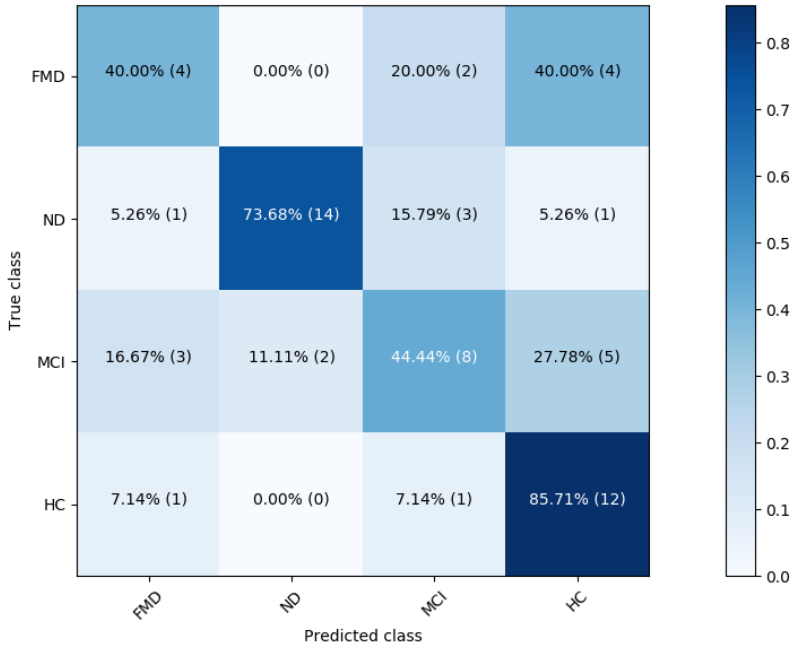


Figure 8.7: *Confusion matrix for the four-way classifier (22 most significant features using the RFE approach (IVA2016/17/18 datasets)).*

word vectors were the two important feature types, while the semantic features were the least important features. From the two verbal fluency tests, all three features related to the fluency semantic test were in this set of most features, which indicates that the features extracted from the fluency semantic tests had more contribution in the classification task.

Features in bold were seen in Table 3.7 (where only CA-inspired features used to classify the Hal dataset, 15 FMD vs 15 ND). So 6 out 10 top features selected by the RFE in Chapter 3 were amongst the top 22 features gained by the RFE on IVA2016/17/18 dataset (2 acoustic, 2 lexical and 2 semantic features).

8.2.6 Feature selection (statistically significant)

Similar to Section 3.3.1 in parallel to the RFE feature selection approach, we used the statistic tests (Student's t-test for normal and Mann-Whitney U test for non-normal features) to identify the statistically significant features among the 78 final features. Table 8.3 shows the 24 statistically significant features using the statistic tests. 10 out of 22 features in Table 8.2 can be seen in this table as well (marked with *), however, considerable

Table 8.2: *The 22 significant features using the RFE approach. Features in bold were also in Table 3.7 (top 10 features on Hal dataset using only CA-inspired features).*

Rank	Feature	Feature type
1	ApsAvgSil	Acoustic
2	PatAVPauses	Acoustic
3	PatAvgSil	Acoustic
4	PatSemSTDAoA	Fluency semantic test
5	ApsAVUniqueWords	Lexical
6	APsAVTurnLength	Acoustic
7	**PatAVFillers**	Semantic
8	PatSemCount	Fluency semantic test
9	WV_col5	Word vector
10	PatPhnAVGAoA	Fluency phonemic test
11	PatSemAVGAoA	Fluency semantic test
12	WV_col4	Word vector
13	WV_col7	Word vector
14	**APsNoOfTurns**	Acoustic
15	WV_col1	Word vector
16	WV_col3	Word vector
17	**PatFailureExampleEmptyWords**	Semantic
18	**PatAVUniqueWords**	Lexical
19	WV_col2	Word vector
20	**PatAVTurnLength**	Acoustic
21	**PatAVAllWords**	Lexical
22	PatNoOfTurns	Acoustic

number of these 24 features were for the APs and mostly they were lexical features. 15 out of 24 features were lexical features, 5 were acoustic features, 2 were semantic features, 1 is word vector feature and 1 is fluency test feature. Therefore, the features selected by the statistic tests were different than the 22 features selected by the RFE and the classifier accuracy gained by these features is different as well. In fact the accuracy of the four-way classifier using the 24 statistically significant was 57%, which is 5% less than the accuracy of the classifier trained by the 22 features (RFE approach).

Figure 8.8 shows the confusion matrix for this classification. Similar to confusion matrix of the 22 features selected by the RFE approach, 86% (12 out of 14) of HC and 74% (14 out of 19) ND were predicted truly, however, only 33% (6 out of 18) of MCI and 30% (3 out of 10) of FMD were identified correctly. The maximum confusion were for 50% (5 out of 10) of FMD patients who were predicted as HC by mistake (this was 40%

Table 8.3: *The 24 statistically significant features using the normality tests (Shapiro-Wilk and D'Agostino) and then parametric (Student's t-test) for normal (norm.) features and non-parametric (Mann-Whitney U test) for non-normal (non-norm.) features. Features in bold were also in Table 3.8 (top 10 features on Hal dataset using the statistic tests. Features with '*' were in Table 8.2.*

Rank	Feature	Feature type
1	PatAvgNoun	Lexical
2	PatAvgCardinal	Lexical
3	ApsAVUniqueWords*	Lexical
4	ApsAvgSil*	Acoustic
5	PatSemSTDAoA*	Fluency semantic test
6	ApsAVTurnLength*	Acoustic
7	WV_col5*	Word vector
8	**PatAVFillers***	Semantic
9	**APsNoOfTurns***	Acoustic
10	**PatAVEmptyWords**	Semantic
11	**PatAVUniqueWords***	Lexical
12	**PatAVAllWords***	Lexical
13	**PatAVTurnLength***	Acoustic
14	PatNoOfTurns*	Acoustic
15	PatAvgOtherPOS	Lexical
16	ApsAvgVerb	Lexical
17	PatAvgPreposition	Lexical
18	APsAvgPronoun	Lexical
19	ApsAvgPreposition	Lexical
20	PatAvgWhword	Lexical
21	ApsAvgPreposition	Lexical
22	APsAvgConjunction	Lexical
23	APsAvgAdjective	Lexical
24	APsAvgDeterminer	Lexical

in Figure 8.7). 28% (5 out of 18) of MCI were confused by ND patients.

8.2.7 F1-measure

In addition to the accuracy of a classifier, some studies report the precision (the ratio of the true-positives to the total predictions), the recall (the ratio of the true-positives to the total true positives) and the F1-measure (combining the precision and the recall[1]) for the classifier. Having a high F1-measure indicates a stronger classifier (with both good

[1] $F1 = 2\frac{precision.recall}{precision+recall}$

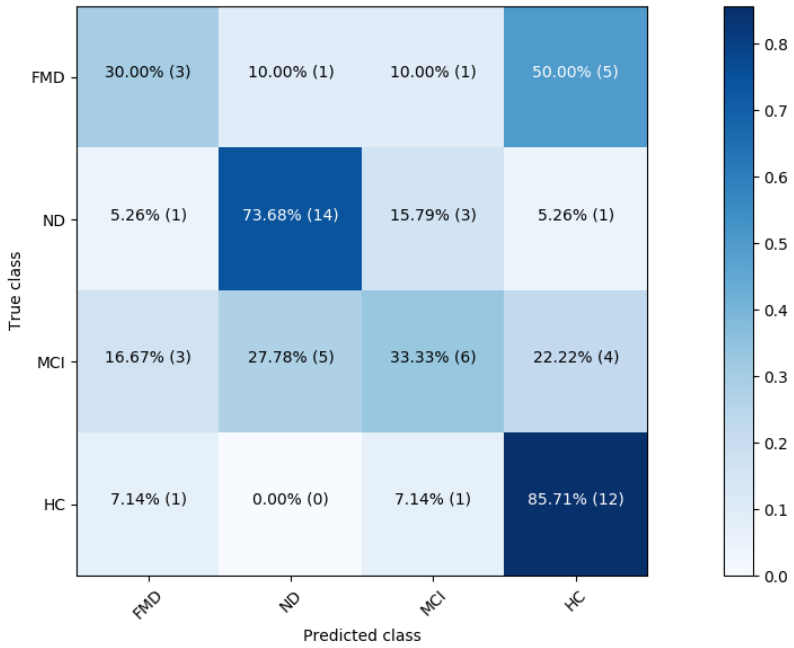


Figure 8.8: *Confusion matrix for the four-way classifier (24 most statistically significant features (IVA2016/17/18 datasets)).*

Table 8.4: *Unweighted average precision, recall and F1-measure for the four-way classifiers. conv.: conversations, fl.tst.: fluency tests*

Dataset	**Features**	**Accuracy**	**Precision**	**Recall**	**F1-measure**
IVA2016/17/18 conv.+fl.tst.	78	47.6%	44.3%	45.7%	44.1%
IVA2016/17/18 conv.+fl.tst.	22(RFE)	**62.3%**	**60.9%**	**61.0%**	**59.7%**
IVA2016/17/18 conv.+fl.tst.	24(stats.)	57.0%	54.1%	55.7%	53.3%

precision and recall). Since in our experiments there were imbalanced number of samples in each patient group we can use unweighted average F1-measure to show the performance of the classifier.

For the four-way classifier using the 78 features, the 22 most significant features (RFE) and the 24 most statistically significant features we calculated the unweighted average precision, recall and F1-measure (Table 8.4). As can be seen, the unweighted F1-measures for the classifier using all 78 features was around 44% (accuracy around 48%). Using the 22 features (RFE) the unweighted F1-measure increased to around 60% and using the 24 statistically significant features the unweighted F1-measure gained was 53%. Therefore, in general the performance of the four-way classifier using the 22 features resulting from the RFE performed better than all features and the 24 statistically significant features.

Comparing our final results (blind evaluation) to the results reported by [Weiner et al., 2018] (who used a similar pipeline as our dementia detection system), they achieved 49% UAR using a three-way classifier (AD, HC and AACD groups) using 80 samples with transcriptions. As can be seen from Table 8.4, our four-way classifier using all 78 features gained around 46% UAR (UAR 61% using the 22 top features), while we had only 61 samples to train the classifier. They also used 188 un-transcribed samples and gained 65% UAR, but we cannot compare our results with that since they had much more data to train the classifier and fewer classes to identify (three vs. four).

8.3 Discussion

In this chapter we showed that the automatic dementia detection system can classify between the four diagnostic classes with a relatively good accuracy and ROC-AUC. The classifiers trained on the most significant (22) features could classify between 57% and 94% accuracy across the 11 classifiers (ROC-AUC between 50% and 94%). The four-way classifier achieved between 62% and 82% ROC-AUC. In the previous chapters (**Chapter 6** and **7**) we showed that the differences between the results achieved by the classifiers on the features extracted from the manual transcript were not much different than the classifiers trained on the automatic transcript and segmentation. Now the question is why the classifiers can identify almost the same way despite having the errors caused by the automated segmentation (the diarisation module) and the automated transcript (the ASR module).

The most important diarisation error was caused by the speaker error. This will introduce uncertainty between the segments allocated to the patients and the accompanying persons, i.e. some of the segments were wrongly associated to the speakers. Consequently some of the features extracted for the patients were mixed up wrongly with the features extracted for the accompanying person. This can affect the results of the classifier, however, since there were at most two speakers to be recognised by the diarisation module, the effects of the error may not be too significant.

Most importantly were the errors from the ASR module. These errors include three

sub categories of errors: insertion, deletion and substitution errors. Figures 8.9, 8.10 and 8.11 show the top 50 words with the highest insertion, deletion and substitution errors for the automated system. Looking at the three figures, we found out that all lists include the 'function words' in English ('I', 'and', 'you', 'a', 'the', etc.), i.e. most errors caused by the ASR were related to the very common words in English. These function words, although very important in an English sentence, do no necessarily affect all the features extracted by our system. Especially the semantic and the word vector features ignore these function words. These words were normally considered as the stop list words which were omitted from the other words during the pre-processing in almost all common NLP text processing step. This can justify why the acoustic, semantic and the word vector features remained as some of the most important features in the feature selection process. Also looking at the top 50 words with the highest substitution errors, we can see that the substituted words were phonetically very close together, examples: 'er' and 'um', 'no' and 'know', 'and' and 'um', 'being' and 'been'. It is worth mentioning that considerable amount of the total words were function words (e.g. 53% of IVA2016 dataset). We re-calculated the WER ignoring the function words, however the number of errors did not decrease.

In addition to the conversations, the verbal fluency tests contained important information which can be used for classification. The features extracted from the tests were amongst the most significant features, which indicates how important these features were for discriminating the diagnostic categorises. Of these two verbal fluency tests, the fluency semantic test was much more discriminate than the fluency phonemic test. One of the major reason for this is the difficulty of automatic recognition of the words beginning with the letter 'P' compared to the animal names. Mixing up the P-words causes errors in counting as well as calculating the precise age of acquisition for these words.

There were many acoustic, word vector and fluency test features in the 22 features selected by the RFE approach which shows the importance of these feature types. However, the 24 statistically significant features, were mostly lexical features and due to the ASR errors the extracted lexical features could not not reliable. This might be the reason why these statistically important features were resulted in a slightly lower unweighted average f1-measure of the four-way classifier compared to the features gained by the RFE.

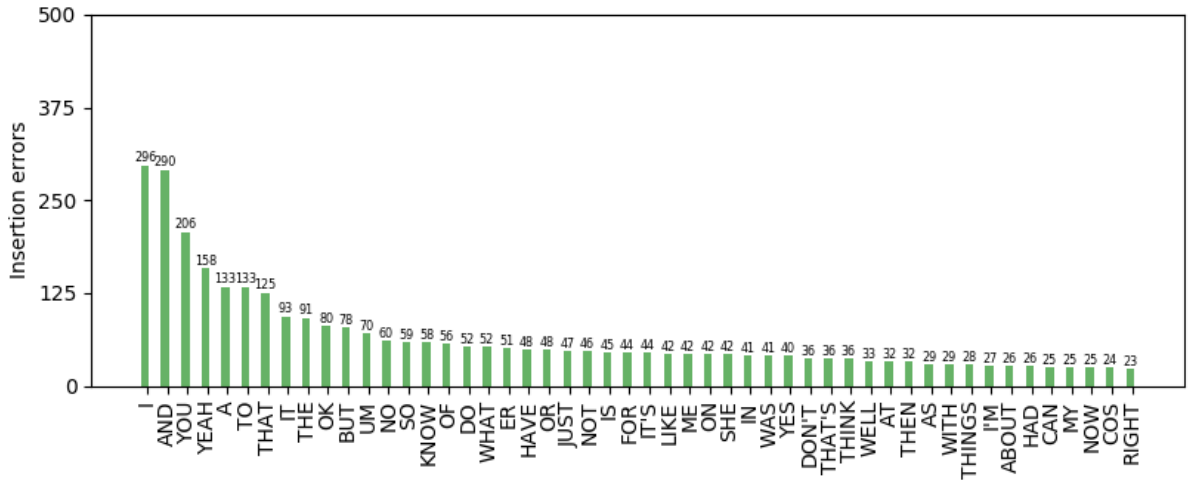


Figure 8.9: *The top 50 words with the highest insertion errors.*

The confusion matrix for the four-way classifier using all the 78 features, 22 features selected by the RFE approach and 24 statistically important features, all generally showed that the HC patient groups and the ND groups had the least confusion with the other patients group. However, most FMD subjects were confused as HC, and most MCI subjects were confused as ND. These confusions reflect the common symptoms between these patient groups. As we know that patients with MCI shares symptoms with ND and HC also FMD patients have similar memory complain as ND patients but they might have features close to HC.

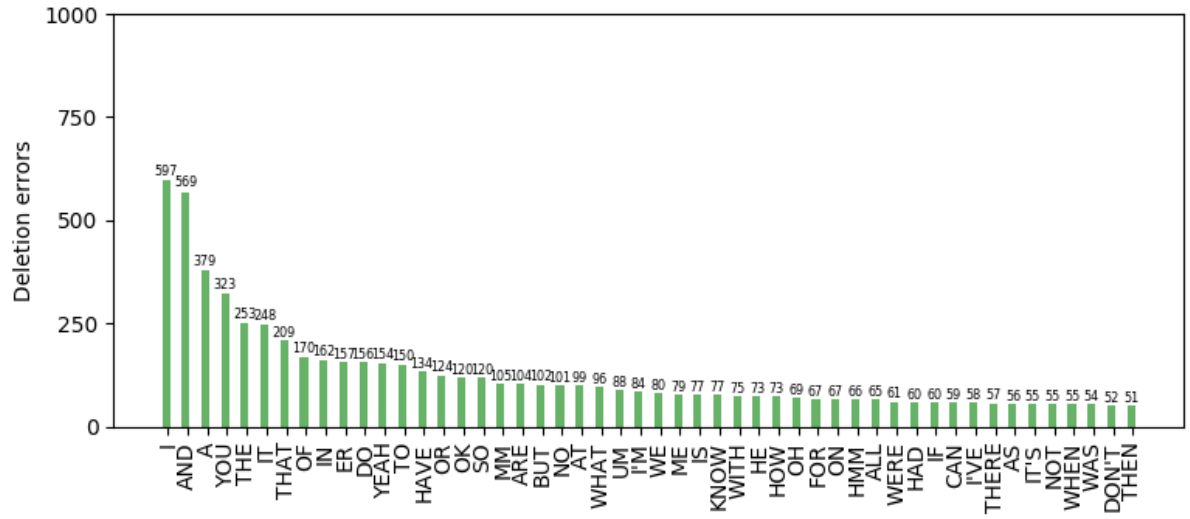


Figure 8.10: *The top 50 words with the highest deletion errors.*

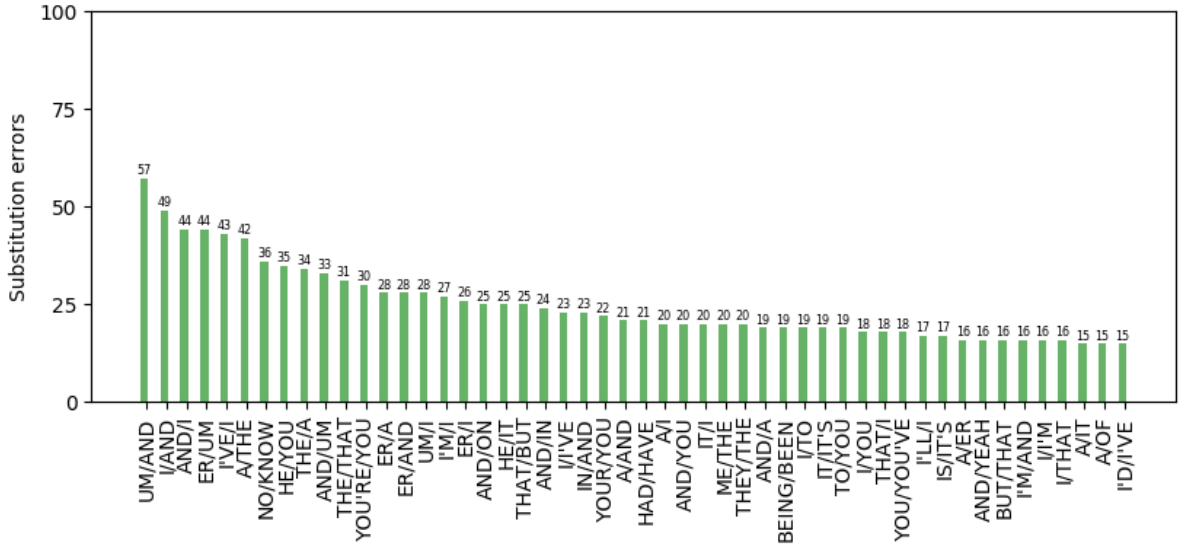


Figure 8.11: *The top 50 words with the highest substitution errors.*

8.4 Summary

In this chapter, we used the conversations and the verbal fluency tests collected in the IVA2017/18 datasets to evaluate the overall dementia detection system. The measures used for the final evaluation of the system were the accuracy of the classifiers as well as the ROC-AUC (robustness of the classifier with regards to both the sensitivity and the specificity).

The datasets included four diagnostic classes: FMD, ND, MCI, and HC. The best results were achieved when we mixed the conversations and the tests of the IVA2016 dataset with the IVA2017/18 datasets and selected the 22 most significant features (RFE approach). Despite having the ASR and diarisation errors, the accuracy of the four-way classifier was 62% and the ROC-AUC 82%. Considering the difficulties of a four-way classification task, the results seem promising and reflect much more realistically the conditions of the referrals to the memory clinics and difficulties of making diagnostic decisions by the neurologists. Specially we showed that the acoustic features, the word vector and verbal fluency tests features were mostly in the list of the most significant features gained by the RFE feature selection approach.

Chapter 9

Conclusions and further work

Contents

9.1 Conclusions 182

9.1.1 Feasibility of developing an automatic system to identify dementia 183

9.1.2 Techniques and methodologies required for the system 184

9.1.3 Providing diagnostic information for neurologists 185

9.1.4 Keeping track of dementia over time 186

9.1.5 Final evaluation on a real clinical setting 186

9.2 Future work 187

9.2.1 Improving the components of the system 187

9.2.2 Dealing with other challenges of conversations 187

9.2.3 Extracting other types of features 188

9.2.4 Investigating other cognitive tests 188

9.2.5 Improving the Intelligent Virtual Agent (IVA) 188

9.2.6 Longitudinal applications 189

9.3 Concluding remarks 189

9.1 Conclusions

After introducing the project in **Chapter 1**, the thesis started with a background overview of dementia in **Chapter 2**. Dementia is one of the leading causes of death in the UK, accounting for 12.7% of all deaths, and the number of people developing dementia is predicted to rise to one million by 2021 [Alzheimer's society, 2018]. According to the UK governments' report [Department of Health, 2018], the cost of dementia on society is estimated to be 26 billion a year. Dementia is a disorder of the brain, caused by a number of different pathological processes including Alzheimer's Disease (AD). Dementia predominantly affects the neuropsychological domains of learning and memory but speech and language are also affected. As they lose their speech and communication ability, they get gradually isolated and passive and and forget their very basic human needs. Dementia ultimately leads to death.

Treatments are most effective in the early stages of neuro-degenerative disorders before dementia has developed and irreversible brain damage has occurred (there is no cure, they mostly delay the progression of the decease). However, it is difficult to identify people in the early stages of neuro-degeneration because the symptoms of disorders causing dementia overlap with memory concerns associated with normal ageing, depression or excessive anxiety about cognitive function. Current tests capable of identifying people at high risk of developing dementia (e.g. Positron Emission Tomography (PET) and amyloid analysis of the CerebroSpinal Fluid (CSF)) are expensive and invasive. The currently available tests for stratifying or screening people with cognitive complaints, based on pen-and-paper testing, lack sensitivity or specificity, especially early in the disease process. In addition, they require time and human resources.

It is evident that the qualitative Conversation Analysis (CA) of the neurologist-patient interactions can help in identifying clues to distinguish between patients with memory problems due to an emerging Neuro-degenerative Disorders (ND) and non-progressive memory difficulties. However, applying such a manual process is prohibitively expensive and time consuming, requires human expertise and resources and not feasible for large-scale use.

There is, therefore, an urgent medical need for a reliable, repeatable, non-invasive, easy to use, and low-cost tool for identifying people at risk of developing dementia. This would ensure quicker access to specialist assessment and treatment for those found to be at high risk and more rapid reassurance for those at low risk of developing dementia. An ideal tool would allow re-testing for those at intermediate risk or whose performance fluctuates, and would be usable without requiring assessor expertise for example in people's own homes.

This thesis was based on deploying the latest technologies in speech, text and machine learning to automatically analyse conversations between patients and neurologists to detect the early signs of Neuro-degenerative Disorders (ND). Automatic analysis of conversation is a challenging task requiring especially automatic segmentation of the audio streams and then transcribing the utterances of the segments. As we developed the dementia detection system, we attempted to explore and find the answers to a number of fundamental research questions listed in **Chapter 1**.

9.1.1 Feasibility of developing an automatic system to identify dementia

The first research question was about the feasibility of developing such an automatic tool to identify the early signs of dementia. In order to answer the question we started by developing an initial prototype of the automatic dementia detection system. **Chapter 3** was dedicated to introducing our suggested system. In this chapter we concentrated mostly on the feature extraction module of the system (i.e. automatically extracting features from the conversations). Then these features were passed to a classifier. We introduced a number of CA features inspired by the features identified by Elsey et al. [2015]. The best classifier accuracy trained on these features was 97% (using the Perceptron classifier), however, two out of the features were visual-conceptual, and as we were concentrating on audio processing techniques, these features were omitted from the list of the CA-inspired features.

Chapter 3 showed the feasibility of developing an automatic dementia detection system by processing the conversations between the patients, the neurologists, and the accompanying persons if they were present. The initial promising results confirmed that

the introduced quantitative features were as informative as the qualitative features identified by human experts (the accuracy of the classifiers were close to the results reported by Elsey et al. [2015]). In addition, we discovered the importance of extracting features from all the participants in a conversation, since some of the features extracted from the neurologist and the accompanying persons were found to be contributing positively and significantly in the classification task (they were in the list of the most significant features).

9.1.2 Techniques and methodologies required for the system

The second research question was about the type of speech, text and machine learning techniques that was required for developing a fully automatic system. The automated dementia detection system, which was introduced in **Chapter 3**, included a speaker diarisation unit, an Automatic Speech Recognition (ASR), a feature extraction module and a final classifier. Gradually in the following chapters, we introduced more automation to the system by adding automatic modules to the system. We nominated six different classifiers, all showed a high accuracy rates (over 91%) to classify between the two patient groups (Functional Memory Disorder (FMD) and ND), however, the statistic tests did not show any significant differences between the performance of the classifier. For our dementia detection, however, we needed to choose a single classifier. The Logistic Regression (LR) classifier was nominated for the system (in a number of experiments we found this shows slightly higher accuracy).

Chapter 4 concentrated on the ASR module (automatic transcription) of the system. We trained a number of baseline ASRs (HMM-based) as well as the final ASR (DNN-based) using the Kaldi toolkit. Training ASRs to transcribe natural conversation is a challenging task, especially when the conversations were not recorded by good quality microphones and in a quiet environment with an acceptable level of noise. Despite all these issues, we managed to train a relatively good ASR (in comparison to published performance levels on similar spontaneous speech recognition tasks).

Chapter 5 presented the diarisation module (automatic segmentation) of the system. We started with a baseline diarisation module, however, the best results on the

Hallamshire (Hal) dataset (our original dataset) was achieved by the diarisation module trained using the Kaldi toolkit.

The initial CA-inspired features included acoustic, lexical and semantic features (we removed the visual-conceptual features due to time limitation of this work), however, **Chapter 6** explored additional features including the extended acoustic, the extended lexical and the word vector features, and the total number of features arrived at was 99. In this combination, the acoustic and the word vector features were more important than the semantic and the lexical features. The lexical features, and to some extent the semantic features, were directly dependent on the words produced by the ASR. Therefore the ASR errors could compromise the role of these features in discriminating dementia.

As we introduced the IVA in **Chapter 7** we removed the features extracted from the neurologist (since the agent always asks the same questions). The new data collection using the IVA enabled us to also recruit patients with the third diagnosis, Mild Cognitive Impairment (MCI) which made the classification task more challenging (the accuracy of the classifier dropped from around 90% for two classes to 67% for three classes). Finally in **Chapter 8** we used the latest collected data which included the forth group, Healthy Control (HC) to the classification task and to boost the classifier, the fluency test features were added to the features. The classifiers results, confirmed the importance of these features as well as the acoustic and word vector features.

The experiments reported in **Chapters 4**, **5**, and **6** allowed us to identify different methodologies and techniques required for completing the automatic dementia detection system, hence addressing the second research question. A relatively standard pipeline was sufficient to prove the functionality of the automatic dementia detection system, however, great care needed to handle the challenges and complexities inherent in the automatic processing of spontaneous speech found in these types of conversations.

9.1.3 Providing diagnostic information for neurologists

The third research question was focused on improving the diagnostic information that neurologists might have access to. Throughout the project, we worked closely with the neurologists and neuroscientists to ensure that the developed dementia detection system

might one day be integrated in the current diagnostic pathways (e.g., as part of early stratification), or be used for general population screening. In addition to diagnostic categories that the system might output, some explorative work was done to find ways to communicate the results of such a test to clinicians. They are mostly interested in human readable information (quantitative and qualitative features) which might be achieved by converting current feature values to scores (e.g., between 1 and 10). Further work is needed though to fully understand what more would be needed.

9.1.4 Keeping track of dementia over time

The last research question was about how to collect more data and tracking the progression of the patients' symptoms over time. In order to answer this question we used the IVA to collect data in the memory clinic during three summers (2016, 2017 and 2018). The IVA enables us to not only elicit conversations with patients, but also to administer two standard cognitive tests (the fluency semantic and fluency phonemic tests) and automatically score these tests. We showed that despite having different lengths of turns and pauses, the IVA-led conversations achieved comparable results to the neurologist-led conversations. The IVA allowed us to add to our data collection throughout three year, however, only two of the participants were able to return for a second time to take the IVA test. Therefore, we were not able to evaluate how repeatable the test is. This is something we will be exploring in the future, and the accuracy achieved for the one-shot tests is hopefully an indication that a good degree of sensitivity to a decline in cognition can be measured.

9.1.5 Final evaluation on a real clinical setting

Finally, **Chapter** 8 included the final evaluation of the system, using the conversation and the verbal fluency tests and using the full system as well as the full set of diagnostic classes (combining the IVA2017/18 datasets with the IVA2016 dataset). The final dataset contained the additional diagnostic class of the HC and hence reflecting more realistically situations that neurologists face in memory clinics, where some HCs are also present. Despite the errors caused by the diarisation and the ASR, the four-way classifier and 10

sub classifiers all had reasonably good classification accuracy and Receiver Operating Characteristic (ROC)-Area Under Curve (AUC) using the most significant features.

The most important features were the CA-inspired and the word vector features which are not dependent on the errors caused by the diarisation and the ASR. On the other hand, the lexical-only and the acoustic-only features were not as significant of the other feature types, since the ASR errors makes it difficult for them to contribute sufficient discriminant information in the classification task. The cognitive tests (especially the Semantic Test) were in the list of the most significant features as well.

In summary, we have developed and evaluated, in a real clinical setting, a system for detecting early signs of dementia based on a number of text, speech and machine learning technologies. The results achieved in the final evaluation chapter shows great promise, and that potentially this system can help to detect the early signs of dementia.

9.2 Future work

9.2.1 Improving the components of the system

Working with medical datasets has its own issues such as having limited access to data, mostly to do with ethical issues of sharing data, and a limited number of recordings to work with. We are hoping to continue to record conversations with the IVA-based system. Access to more data would allow to improve the different components of the dementia detection system (training a better language and acoustic models for the ASR, improving the diarisation module and training a better classifier (e.g. Deep Neural Network (DNN) based)), which in return should result in better overall performance of the system in terms of accuracy and robustness. In particular, having access to a critical mass of recordings of the more rare types of dementia, such as Fronto-Temporal Dementia (FTD), would improve the versatility of the tool.

9.2.2 Dealing with other challenges of conversations

In **Chapter 4** we listed the major challenges of dealing with spontaneous speech. In training the ASR we tried to deal with some of these issues, however, there is plenty of room for improvement. Finding solutions for the mentioned problems is not easy. For

instance, in order to deal with the overlapping segments we removed very short segments, however, we might lose the words uttered by the speakers there. Investigating other approaches to recover at least parts of the overlapping segments in a conversation, and deal in a different way with the turn-taking phenomenon, would be beneficial. Disfluencies were one of the other challenges to the ASR in spontaneous speech. In our approach, we recognised them similar to the other words in utterance, however, it is possible to train ASR to detect disfluencies and/or cope with them more effectively.

9.2.3 Extracting other types of features

There are other types of features we could explore in this study, but due to time limitation we did not investigate them. For instance, we could detect the emotions of speakers in their utterances. There might be informative emotional features which can contribute to identifying dementia. Detecting disfluencies is another example of feature types which need further investigations.

9.2.4 Investigating other cognitive tests

Although the main focus of this study was processing audio recordings of conversation, we found that the other standard cognitive tests (e.g. semantic and phonemic tests) also could be administrated by the IVA. The automatic scoring of the tests could improve the overall accuracy and robustness of the classifier. There are other well-known standard cognitive tests like the 'Cookie Theft' picture descriptions which can be automatically scored and investigated with respect to how they could be useful in identifying early signs of dementia.

9.2.5 Improving the IVA

Due to the limitations of this work, we used a relatively simple IVA to act as a neurologist and conduct conversations with patients. Obviously, the quality of the IVA, especially in terms of intelligibility and interactional capabilities could be improved, for instance by applying the Spoken Dialogue System (SDS) techniques. More intelligent IVA could conduct much more flexible and human-like conversations with patients. This can be

investigated as a further study.

9.2.6 Longitudinal applications

Although we have demonstrated a system which potentially can help in diagnosing dementia, we did not explore how this type of system can be used in longitudinal applications. The development of a fully-automatic test suitable for home-based tracking and monitoring is another research area which would warrant further investigation. Also, in developing such a system, investigating how clinicians could best access the information of the patients as part of their diagnostic work is an open research question.

9.3 Concluding remarks

This thesis introduced a unique and novel approach to identify people in risk of developing dementia based on the fully automatic analysis of people's conversational ability. We explored and investigated a number of techniques and methodologies in machine learning, speech technology and natural language processing. The final evaluation of the system on a real clinical setting revealed promising results which confirms a potential bright future of developing a low-cost, repeatable, non-invasive, and less stressful alternative to current cognitive assessments.

References

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1533–1545. 68, 85

Al-Hameed, S., Benaissa, M., and Christensen, H. (2016). Simple and robust audio-based detection of biomarkers for Alzheimers disease. *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 32–36. 39

Al-Hameed, S., Benaissa, M., and Christensen, H. (2017). Detecting and predicting Alzheimer's disease severity in longitudinal acoustic data. In *Proc International Conference on Bioinformatics Research and Applications*, pages 57–61. ACM. 39

Al-Hameed, S., Benaissa, M., Christensen, H., Mirheidari, B., Blackburn, D., and Reuber, M. (2018). Using acoustic measures to assess cognitive interactional capability in patients presenting with memory problems. *Submitted to PLOS ONE*. 11, 136, 137

Alberdi, A., Aztiria, A., and Basarab, A. (2016). On the early diagnosis of Alzheimer's disease from multimodal signals: A survey. *Artificial intelligence in medicine*, 71:1–29. 16, 19, 20

Alexandra, C., Graff, D., and Zipperlen, G. (1997). Callhome american english speech ldc97s42. https://catalog.ldc.upenn.edu/LDC97S42 [Online; accessed 7 November 2018]. 83, 108

Ali, A., Bell, P., Glass, J., Messaoui, Y., Mubarak, H., Renals, S., and Zhang, Y. (2016). The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. In *Proc Spoken Language Technology (SLT)*, pages 279–284. IEEE. 106

Ali, A., Vogel, S., and Renals, S. (2017). Speech recognition challenge in the wild: Arabic mgb-3. In *Proc Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322. IEEE. 106

Alzheimer's Disease International (2018). Media quick facts - dementia. https://www.alz.co.uk/media/quick-facts [Online; accessed 2 June 2018]. 2

Alzheimer's Society (2015a). Alzheimer's society :dementia uk update. https://www.alzheimers.org.uk/dementiauk [Online; accessed 7 November 2018]. 15

Alzheimer's Society (2015b). What is dementia with lewy bodies (dlb)? https://www.alzheimers.org.uk/site/scripts/documents_info.php?documentID=113 [Online; accessed 7 November 2018]. 15

Alzheimer's Society (2015c). What is frontotemporal dementia? https://www.alzheimers.org.uk/site/scripts/documents_info.php?documentID=167 [Online; accessed 7 November 2018]. 15

Alzheimer's Society (2015d). What is vascular dementia? https://www.alzheimers.org.uk/site/scripts/documents_info.php?documentID=161 [Online; accessed 7 November 2018]. 15

Alzheimer's society (2018). Rise in deaths due to dementia. https://www.alzheimers.org.uk/news/2018-10-23/rise-deaths-due-dementia [Online; accessed 5 November 2018]. 182

Appell, J., Kertesz, A., and Fisman, M. (1982). A study of language functioning in Alzheimer patients. *Brain and Language*, 17(1):73–91. 2

Asgari, M., Kaye, J., and Dodge, H. (2017). Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(2):219–228. 39

Audhkhasi, K., Ramabhadran, B., Saon, G., Picheny, M., and Nahamoo, D. (2017). Direct acoustics-to-word models for english conversational speech recognition. *arXiv preprint arXiv:1703.07754*. 126

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., and Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 4945–4949. IEEE. 80

Barker, J., Vincent, E., Ma, N., Christensen, H., and Green, P. (2013). The pascal chime speech separation and recognition challenge. *Computer Speech & Language*, 27(3):621–633. 83, 86

Barker, J., Watanabe, S., Vincent, E., and Trmal, J. (2018). The fifth'chime'speech separation and recognition challenge: Dataset, task and baselines. *arXiv preprint arXiv:1803.10609*. 84

Bayles, K. A. and Kaszniak, A. W. (1987). *Communication and cognition in normal aging and dementia*. Taylor & Francis Ltd London. 2, 15, 16, 17, 18

Bell, P., Gales, M. J., Hain, T., Kilgour, J., Lanchantin, P., Liu, X., McParland, A., Renals, S., Saz, O., Wester, M., et al. (2015a). The mgb challenge: Evaluating multi-genre broadcast media recognition. In *Proc Automatic Speech Recognition and Understanding (ASRU)*, pages 687–693. IEEE. 106, 109

Bell, P., Lai, C., Llewellyn, C., Birch, A., and Sinclair, M. (2015b). A system for automatic broadcast news summarisation, geolocation and translation. In *Proc 6th Conference of the International Speech Communication Association*. 70

Bell, S., Harkness, K., Dickson, J. M., and Blackburn, D. (2015c). A diagnosis for 55: what is the cost of government initiatives in dementia case finding. *Age and Ageing*, 44:344–345. 55

Bengio, Y. (2009). *Learning deep architectures for AI*, volume 2. Now Publishers Inc. xix, 68, 72, 74

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. OReilly Media Inc. 45

Blackburn, D., Mirheidari, B., Rutten, C., Mayer, I., Walker, T., Christensen, H., and Rueber, M. (2017a). Po029 an avatar aid in memory clinic. 10

Blackburn, D., Reuber, M., Christensen, H., Mayer, I., Rutten, C., Venneri, A., and Mirheidari, B. (2017b). An avatar to screen for cognitive impairment. *Journal of the Neurological Sciences*, 381:319. 10

Blackburn, D. J., Wakefield, S., Shanks, M. F., Harkness, K., Reuber, M., and 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(1):71–81. 19

Blanchard, N., Donnelly, P. J., Olney, A. M., Samei, B., Ward, B., Sun, X., Kelly, S., Nystrand, M., and DMello, S. K. (2016). Identifying teacher questions using automatic speech recognition in classrooms. In *Proc 17th Meeting of the Special Interest Group on Discourse and Dialogue*, page 191. 70

Blanken, G., Dittmann, J., Haas, J. C., and Wallesch, C. W. (1987). Spontaneous speech in senile dementia and aphasia: implications for a neurolinguistic model of language production. *Cognition*, 27:247–274. 24, 123

Boakye, K., Trueba-Hornero, B., Vinyals, O., and Friedland, G. (2008). Overlapped speech detection for improved speaker diarization in multiparty meetings. In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 4353–4356. IEEE. 107

Boersma, P. P. G. et al. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5. 121

Bordawekar, R. and Shmueli, O. (2017). Using word embedding to enable semantic queries in relational databases. In *Proc 1st Workshop on Data Management for End-to-End Machine Learning*, page 5. ACM. 126

Bourlard, H. A. and Morgan, N. (2012). *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media. 74

Boustani, M., Callahan, C., Unverzagt, F., Austrom, M., Perkins, A., Fultz, B., Hui, S.,

and Hendrie, H. (2005). Implementing a screening and diagnosis program for dementia in primary care. *J Gen Intern Med*, 20(7):572–577. 55

Bucks, R. S., Singh, S., Cuerden, J. M., and Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology.*, 14:71–91. 24, 123

Canning, S. D., Leach, L., Stuss, D., Ngo, L., and Black, S. (2004). Diagnostic utility of abbreviated fluency measures in Alzheimer disease and vascular dementia. *Neurology*, 62(4):556–562. 142

Carrasco, E., Epelde, G., Moreno, A., Ortiz, A., Garcia, I., Buiza, C., Urdaneta, E., Etxaniz, A., González, M. F., and Arruti, A. (2008). Natural interaction between avatars and persons with alzheimers disease. In *Proc International Conference on Computers for Handicapped Persons*, pages 38–45. Springer. 141

Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE. 70

Chen, I., Siniscalchi, S. M., and Lee, C. (2014). Attribute based lattice re-scoring in spontaneous speech recognition. In *Pro International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 3349–3353. IEEE. 85

Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics. 69

Chen, S. S. and Gopalakrishnan, P. S. (1998). Clustering via The Bayesian Information Criterion With Applications In Speech Recognition. *Acoustics, Speech and Signal Processing*, 2:645–648. 104

Cheng, G., Peddinti, V., Povey, D., Manohar, V., Khudanpur, S., and Yan, Y. (2017). An exploration of dropout with lstms. In *Proc INTERSPEECH*. ISCA. 74, 90

Cho, E., Kilgour, K., Niehues, J., and Waibel, A. (2015). Combination of nn and crf models for joint detection of punctuation and disfluencies. In *Proc 6th Conference of the International Speech Communication Association*. 70

Christodoulides, G., Avanzi, M., et al. (2015). Automatic detection and annotation of disfluencies in spoken french corpora. In *Proc 6th Conference of the International Speech Communication Association*. 70

Cieri, C., Miller, D., and Walker, K. (2004). The fisher corpus: a resource for the next generations of speech-to-text. In *Proc 4th International Conference on Language Resources and Evaluation (LREC)*. 83, 91, 108

Coen, R. F., Robertson, D. A., Kenny, R. A., and King-Kallimanis, B. L. (2016). Strengths and limitations of the moca for assessing cognitive functioning: findings from a large representative sample of irish older adults. *Journal of geriatric psychiatry and neurology*, 29(1):18–24. 21

Consortium, L. D. et al. (1994). Csr-ii (wsj1) complete. *Linguistic Data Consortium, Philadelphia, vol. LDC94S13A*. 83

Cui, X., Bernard, A., and Alwan, A. (2003). A noise-robust asr back-end technique based on weighted viterbi recognition. In *Eighth European Conference on Speech Communication and Technology*. 66

Cyarto, E. V., Batchelor, F., Baker, S., and Dow, B. (2016). Active ageing with avatars: a virtual exercise class for older adults. In *Proc 28th Australian Conference on Computer-Human Interaction*, pages 302–309. ACM. 141

Dat, T. H., Dennis, J., Ren, L. Y., and Terence, N. W. Z. (2016). A comparative study of multi-channel processing methods for noisy automatic speech recognition in urban environments. In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 6465–6469. IEEE. 71

De Renzi, E. and Faglioni, P. (1978). Normative data and screening power of a shortened version of the token test. *Cortex*, 14(1):41–49. 42

Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., and Dumouchel, P. (2009). Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *Tenth Annual conference of the international speech communication association*. 90

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798. 107

Deng, L. and Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech and Language Processing*, 21(5):1060–1089. 67, 68

Department of Health (2018). Dementia: applying all our health. https://www.gov.uk/government/publications/dementia-applying-all-our-health/dementia-applying-all-our-health [Online; accessed 2 June 2018]. 2, 15, 182

Dhaka, A. K. and Salvi, G. (2017). Sparse autoencoder based semi-supervised learning for phone classification with limited annotations. In *Proc Grounding Language Understanding (GLU)*, pages 22–26. 81

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923. 49

Droppo, J., Deng, L., and Acero, A. (2001). Evaluation of the splice algorithm on the aurora2 database. In *Proc 7th European Conference on Speech Communication and Technology*. 66

Eide, E. and Gish, H. (1996). A parametric approach to vocal tract length normalization. In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 1, pages 346–348. IEEE. 66

Elsey, C., Drew, P., Jones, D., Blackburn, D., Wakefield, S., Harkness, K., Venneri, A., and 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. xxiii, 4, 5, 20, 27, 28, 30, 34, 40, 42, 44, 46, 47, 56, 57, 183, 184

Enarvi, S. and Kurimo, M. (2016). Theanolm-an extensible toolkit for neural network language modeling. *arXiv preprint arXiv:1605.00942*. 85

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM. 122

Fernández-Matarrubia, M., Matías-Guiu, J., Moreno-Ramos, T., and Matías-Guiu, J. (2014). Behavioural variant frontotemporal dementia: Clinical and therapeutic approaches. *Neurología (English Edition)*, 29(8):464–472. 15

Ferras, M. and Bourlard, S. M. H. (2016). Speaker diarization and linking of meeting data. *IEEE/ACM Transactions on Audio Speech and language processing*, 24(11):1935–45. 108

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*, 12(3):198–98. 21, 42

Forbes-McKay, K. E., Ellis, A. W., Shanks, M. F., and 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(11):1625–1632. 142

Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278. 69

Fransen, J., Pye, D., Robinson, T., Woodland, P., and Young, S. (1994). Wsjcam0 corpus and recording description. *Cambridge University Engineering Department (CUED) Speech Group, Trumpington Street, Cambridge CB2 1PZ, UK, Tech. Rep.* 83

Fraser, K. C. and Hirst, G. (2016). Detecting semantic changes in Alzheimers disease with vector space models. In *Proc Processing of Linguistic and Extra-Linguistic Data from*

People with Various Forms of Cognitive/Psychiatric Impairments. Linköping University Electronic Press. 39

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2015). Linguistic Features Identify Alzheimer s Disease in Narrative Speech. *Journal of Alzheimer's Disease*, 49:407–22. 38

Gales, M. J. et al. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language*, 12(2):75–98. 66

Gao, S., Zhang, H., and Gao, K. (2017). Text understanding with a hybrid neural network based learning. In *Proc International Conference of Pioneering Computer Scientists, Engineers and Educators*, pages 115–125. Springer. 126

Garcia-Romero, D., Snyder, D., Sell, G., Povey, D., and McCree, A. (2017). Speaker diarization using deep neural network embeddings. In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 4930–4934. IEEE. 108

Garofolo, J., Graff, D., Paul, D., and Pallett, D. (1993). Csr-i (wsj0) complete ldc93s6a. *Web Download. Philadelphia: Linguistic Data Consortium*. 83

Garofolo, J. S. (1993). Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*. 82

Gers, F. A., Schmidhuber, J., and Cummins, F. (1999). Learning to forget: Continual prediction with lstm. *IET*. 76

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD telephone speech corpus for research and development. In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 1, pages 517–520. 82, 91

Gonzalez-Moreira, E., Torres-Boza, D., Kairuz, H. A., Ferrer, C., Garcia-Zamora, M., Espinoza-Cuadros, F., and Hernandez-Gomez, L. A. (2015). Automatic prosodic analysis to identify mild dementia. *BioMed research international*, 2015. 17

Goodglass, H., Kaplan, E., and Weintraub, S. (1983). *Boston naming test*. Lea & Febiger. 22

Gordon, M. and Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29(4):383–406. 122

Gosztolya, G., Tóth, L., Grósz, T., Vincze, V., Hoffmann, I., Szatlóczki, G., Pákáski, M., and Kálmán, J. (2016). Detecting mild cognitive impairment from spontaneous speech by correlation-based phonetic feature selection. In *Proc INTERSPEECH*, pages 107–111. ISCA. 36

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc 23rd International Conference on Machine Learning*, pages 369–376. ACM. 78

Graves, A. and Jaitly, N. (2014a). Towards end-to-end speech recognition with recurrent neural networks. In *Proc International Conference on Machine Learning*, pages 1764–1772. 68, 86

Graves, A. and Jaitly, N. (2014b). Towards end-to-end speech recognition with recurrent neural networks. In *Proc International Conference on Machine Learning*, pages 1764–1772. 78, 80

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 6645–6649. IEEE. 85

Gravier, G., Betser, M., and Ben, M. (2010). AudioSeg: Audio Segmentation Toolkit, release 1.2. *IRISA*. 105

Greenberg, C. S., Bansé, D., Doddington, G. R., Garcia-Romero, D., Godfrey, J. J., Kinnunen, T., Martin, A. F., McCree, A., Przybocki, M., and Reynolds, D. A. (2014). The nist 2014 speaker recognition i-vector machine learning challenge. In *Odyssey: The Speaker and Language Recognition Workshop*, pages 224–230. 105

Greenberg, C. S., Stanford, V. M., Martin, A. F., Yadagiri, M., Doddington, G. R., Godfrey, J. J., and Hernandez-Cordero, J. (2013). The 2012 nist speaker recognition evaluation. In *Proc INTERSPEECH*, pages 1971–1975. ISCA. 105

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182. 49

Hamilton, H. E. (1994). *Conversations with an Alzheimers patient: An interactional sociolinguistic study*. Cambridge, England: Cambridge University Press. 2, 15, 16, 17, 18, 26, 47

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162. 125

Harry, A. and Crowe, S. F. (2014). Is the boston naming test still fit for purpose? *The Clinical Neuropsychologist*, 28(3):486–504. 22

Hayward, M., Jones, A.-M., Bogen-Johnston, L., Thomas, N., and Strauss, C. (2017). Relating therapy for distressing auditory hallucinations: A pilot randomized controlled trial. *Schizophrenia research*, 183:137–142. 141

Heittola, T., Mesaros, A., Virtanen, T., and Gabbouj, M. (2013). Supervised model training for overlapping sound events based on unsupervised source separation. In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 8677–8681. 71

Hessler, J., Brnner, M., Etgen, T., Ander, K.-H., Frstl, H., Poppert, H., Sander, D., and Bickel, H. (2014). Suitability of the 6CIT as a screening test for dementia in primary care patients. *Aging Ment Health*, 18(4):515–520. 55

Hilger, F. and Ney, H. (2006). Quantile based histogram equalization for noise robust large vocabulary speech recognition. *IEEE Transactions on Audio Speech and Language Processing*, 14(3):845–854. 66

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., , and Kingsbury, B. (2012). Deep neural networks for

acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97. 68, 72, 73, 74, 85

Hinton, G., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554. 68, 74, 84

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780. 76

Holmes, J. and Holmes, W. (2001). *Speech Synthesis and Recognition*. UK:Taylor & Francis, second edition. 84

Hori, T., Hori, C., Watanabe, S., and Hershey, J. R. (2016). Minimum word error training of long short-term memory recurrent neural network language models for speech recognition. In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5990–5994. IEEE. 86

Hsieh, S., Schubert, S., Hoon, C., Mioshi, E., and Hodges, J. R. (2013). Validation of the addenbrooke's cognitive examination iii in frontotemporal dementia and Alzheimer's disease. *Dementia and geriatric cognitive disorders*, 36(3-4):242–250. 22

Huckvale, M., Leff, J., and Williams, G. (2013). Avatar therapy: an audio-visual dialogue system for treating auditory hallucinations. In *Proc INTERSPEECH*, pages 392–396. ISCA. 141

Huijbregts, M. (2008). *Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled*. PhD thesis, University of Twente, The Netherlands. 105, 106, 110

Huijbregts, M., Va Leeuwen, D., and Wooters, C. (2012). Speaker diarization error analysis using oracle components. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):393–403. 106

Inoue, T., Tanaka, T., Nakagawa, S., Nakato, Y., Kameyama, R., Boku, S., Toda, H., Kurita, T., and Koyama, T. (2012). Utility and limitations of phq-9 in a clinic specializing in psychiatric care. *BMC psychiatry*, 12(1):73. 23

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., arbara Peskin, Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The ICSI Meeting Corpus. In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 1, pages I–364–I–367. 83, 107

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., and Ogar, J. (2014). ided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proc Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 27–37. 36, 123

Jiang, H. (2010). Discriminative training of HMMs for automatic speech recognition: A survey. *Computer Speech and Language*, 24(4):589–608. 72

Jones, D. (2015). A family living with Alzheimers disease: The communicative challenges. *Dementia*, 14(5):555–573. 4, 24

Jones, D., Drew, P., Elsey, C., Blackburn, D., Wakefield, S., Harkness, K., and Reuber, M. (2015). Conversational assessment in memory clinic encounters: interactional profiling for differentiating dementia from functional memory disorders. *Aging & Mental Health*, 7863:1–10. 4, 27, 42, 56

Jurafsky, D. and Martin, J. H. (2008). *Speech and language processing*. Prentice Hall, 2nd edition. 24, 91

Kanda, N., Harada, S., Lu, X., and Kawai, H. (2016). Investigation of semi-supervised acoustic model training based on the committee of heterogeneous neural networks. In *Proc INTERSPEECH*, pages 1325–1329. ISCA. 81

Karunanayaka, P., Martinez, B., Eslinger, P. J., and Yang, Q. X. (2017). Olfactory processing is highly cognitively demanding: Sensitive functional marker for cognitive deficits and dementia in ad. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 13(7):P1118–P1119. 39

Kaushik, M., Trinkle, M., and Hashemi-Sakhtsari, A. (2010). Automatic detection and

removal of disfluencies from spontaneous speech. In *Proceedings of the Australasian International Conference on Speech Science and Technology (SST)*. 70

Kemp, C., Griffiths, T. L., Stromsten, S., and Tenenbaum, J. B. (2004). Semi-supervised learning with trees. In *Advances in neural information processing systems*, pages 257–264. 80

Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P. (2007). Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1435–1447. 107

Khan, S. u. D., Becker, K., and Zimman, L. (2015). The acoustics of perceived creaky voice in american english. *The Journal of the Acoustical Society of America*, 138(3):1809–1809. 122

Khodabakhsh, A., Yesil, F., Guner, E., and Demiroglu, C. (2015). Evaluation of linguistic and prosodic features for detection of Alzheimers disease in turkish conversational speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):9. 17

Kindell, J., Sage, K., Keady, J., and Wilkinson, R. (2013). Adapting to conversation with semantic dementia: Using enactment as a compensatory strategy in everyday social interaction. *International Journal of Language and Communication Disorders*, 48(5):497–507. 4, 24, 26

Klimova, B., Maresova, P., Valis, M., Hort, J., and Kuca, K. (2015). Alzheimers disease and language impairments: social intervention and medical treatment. *Clinical interventions in aging*, 10:1401. 16, 17, 18

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(1):112–124. 37

Kramer, B. A. (1982). Depressive pseudodementia. *Comprehensive psychiatry*, 23(6):538–544. 19

Kroenke, K. and Spitzer, R. L. (2002). The phq-9: a new depression diagnostic and severity measure. *Psychiatric annals*, 32(9):509–515. 23

Kumar, S. S. and RangaBabu, T. (2015). Emotion and gender recognition of speech signals using svm. *Emotion*, 4(3). 71

Lafaille-Magnan, M.-E., Madjar, C., Hoge, R., and Breitner, J. C. (2015). Olfactory identification correlates with cerebral blood flow in cognitively normal adults at risk of Alzheimers dementia. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 11(7):P160–P161. 39

Laffitte, P., Sodoyer, D., Tatkeu, C., and Girin, L. (2016). Deep neural networks for automatic detection of screams and shouted speech in subway trains. In *Proc ICASS)*, pages 6460–6464. IEEE. 71

Larner, A. J. (2014). Impact of the National Dementia Strategy in a neurology-led memory clinic: 5-year data. *Clin Med*, 14:216. 55

Lau, J. H. and Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*. 126

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proc International Conference on Machine Learning*, pages 1188–1196. 126

Le, V., Mella, O., and Fohr, D. (2007). Speaker diarization using normalized cross likelihood ratio. In *Proc INTERSPEECH*, pages 873–876. ISCA. 102, 103, 104

Lease, M., Johnson, M., and Charniak, E. (2006). Recognizing disfluencies in conversational speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1566–1573. 70

Leff, J., Williams, G., Huckvale, M., Arbuthnot, M., and Leff, A. P. (2014). Avatar therapy for persecutory auditory hallucinations: what is it and how does it work? *Psychosis*, 6(2):166–176. 141

Lehr, M., Prud'hommeaux, E., Shafran, I., and Roark, B. (2012). Fully automated neuropsychological assessment for detecting mild cognitive impairment. In *Proc 13th Conference of the International Speech Communication Association*. 36

Lerner, G. H. (2004). *Conversation Analysis: studies from the first generation*. Amsterdam John Benjamins Pub. xix, xxvi, 4, 25, 223

Levinson, S. E., Rabiner, L. R., and Sondhi, M. M. (1983). An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *Bell System Technical Journal*, 62(4):1035–1074. 67

Li, S., Akita, Y., and Kawahara, T. (2015). Discriminative data selection for lightly supervised training of acoustic model using closed caption texts. In *Proc 6th Conference of the International Speech Communication Association*. 80, 81

Li, S., Akita, Y., and Kawahara, T. (2016). Semi-supervised acoustic model training by discriminative data selection from multiple asr systems' hypotheses. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(9):1520–1530. 81

Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*. xix, 75, 76

Lopez de Ipina, K., Alonso, J.-B., Travieso, C. M., Sole-Casals, J., Egiraun, H., Faundez-Zanuy, M., Ezeiza, A., Barroso, N., Ecay-Torres, M., Martinez-Lage, P., and Martinez de Lizardui, U. (2013). On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. *Sensors*, 13:6730–45. 35

Lopez de Ipina, K., Sole-Casals, J., Eguiraun, H., Alonsod, J. B., Travieso, C. M., Ezeiza, A., Barrosoa, N., Ecay Torres, M., Martinez Lage, P., and Beitia, B. (2015). Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach. *Computer Speech and Language*, 30:43–60. 35

Lopez-Otero, P., Docio-Fernandez, L., Abad, A., and Garcia-Mateo, C. (2017). Depression detection using automatic transcriptions of de-identified speech. In *Proc INTERSPEECH*, pages 3157–3161. ISCA. 126

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proc 49th Association for Computational Linguistics: Human language technologies*, volume 1, pages 142–150. Association for Computational Linguistics. 126

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60. 52

Martin, A. and Przybocki, M. (2004). Nist speaker recognition evaluation ldc2004s04. *Linguistic Data Consortium, Philadelphia*. 108

Martin, A., Przybocki, M., and Campbell, J. P. (2005). The nist speaker recognition evaluation program. In *Biometric Systems*, pages 241–262. Springer. 108

Martin, A. F. and Greenberg, C. S. (2009). Nist 2008 speaker recognition evaluation: Performance across telephone and room microphone channels. In *Proc 10th Conference of the International Speech Communication Association*. 105, 108

Martin, A. F. and Greenberg, C. S. (2010a). The 2009 nist language recognition evaluation. In *Odyssey*, volume 30. 106

Martin, A. F. and Greenberg, C. S. (2010b). The nist 2010 speaker recognition evaluation. In *Proc 11th Conference of the International Speech Communication Association*. 106

Mathuranath, P., Nestor, P., Berrios, G., Rakowicz, W., and Hodges, J. (2000). A brief cognitive test battery to differentiate Alzheimer's disease and frontotemporal dementia. *Neurology*, 55(11):1613–1620. 22

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157. 49

Mehta, M., Adem, A., and Sabbagh, M. (2012). New acetylcholinesterase inhibitors for Alzheimer's disease. *International Journal of Alzheimers disease*, 2012. 17

Meignier, S. and Merlin, T. (2010). LIUM SpkDiarization: an open source toolkit for diarization. *CMU SPUD Workshop*. 105, 110

Menon, R. and Larner, A. (2011). Use of cognitive screening instruments in primary care: the impact of national dementia directives (NICE/SCIE, National Dementia Strategy). *Fam Pract*, 28(3):272–276. 55

Miao, Y., Gowayyed, M., and Metze, F. (2015). Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. In *Proc Automatic Speech Recognition and Understanding (ASRU)*, pages 167–174. IEEE. 80

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. 125

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Proc 11th Conference of the International Speech Communication Association*. 78

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119. 125

Milner, R. and Hain, T. (2016). Dnn-based speaker clustering for speaker diarisation. In *Proc INTERSPEECH*, pages 2185–2189. ISCA. 109

Milner, R., Saz, O., Deena, S., Doulaty, M., Ng, R. W., and Hain, T. (2015). The 2015 sheffield system for longitudinal diarisation of broadcast media. In *Proc Automatic Speech Recognition and Understanding (ASRU)*, pages 632–638. IEEE. 109

Mioshi, E., Dawson, K., Mitchell, J., Arnold, R., and Hodges, J. R. (2006). The addenbrooke's cognitive examination revised (ace-r): a brief cognitive test battery for dementia screening. *International journal of geriatric psychiatry*, 21(11):1078–1085. 22

Mirheidari, B., Blackburn, D., Harkness, K., Venneri, A., Reuber, M., Walker, T., and Christensen, H. (2017a). An avatar-based system for identifying individuals likely to develop dementia. In *Proc INTERSPEEH*. ISCA. 7, 10

Mirheidari, B., Blackburn, D., Harkness, K., Walker, T., Venneri, A., Reuber, M., and Christensen, H. (2017b). Toward the automation of diagnostic conversation analysis in patients with memory complaints. *Journal of Alzheimer's Disease*, pages 1–15. 6, 7, 10, 40

Mirheidari, B., Blackburn, D., Reuber, M., Walker, T., and Christensen, H. (2016). Diagnosing people with dementia using automatic conversation analysis. In *Proc INTERSPEECH*, pages 1220–1224. ISCA. 6, 10

Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., and Christensen, H. (2019). Dementia detection using automatic analysis of conversations. *Computer Speech & Language*, 53:65–79. 11

Mirheidari, B., Blackburn, D., Walker, T., Venneri, A., Reuber, M., and Christensen, H. (2018a). Detecting signs of dementia using word vector representations. In *Proc INTERSPEECH*, pages 1893–1897. ISCA. 7, 10, 128

Mirheidari, B., Daniel, B., O'Malley, R., Walker, T., Venneri, A., Reuber, M., and Christensen, H. (2018b). Computational cognitive assessment: investigating the use of an intelligent virtual agent for the detection of early signs of dementia. *Submitted to ICASSP 2019*. 7, 11

Miro, A. (2006). *Robust speaker diarization for meetings*. PhD thesis, Universitat Politecnica de Catalunya, Bercelona, Spain. 106

Miro, X. A., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370. xx, 101, 102, 103, 104, 106

Moattar, M. H. and Homayounpour, M. M. (2012). A review on speaker diarization systems and approaches. *Speech Communication*, 54(10):1065–1103. 101, 102, 104

Mohri, M. and Pereira, F. (2002). Weighted Finite-State Transducers in Speech Recognition Weighted Finite-State Transducers in Speech Recognition. *Computer Speech & Language*, 16(1):69–88. 69

Moore, R. J. (2015). Automated Transcription and Conversation Analysis. *Research on Language and Social Interaction*, 48(3):253–270. 35, 70

Morandell, M. M., Hochgatterer, A., Fagel, S., and Wassertheurer, S. (2008). Avatars in assistive homes for the elderly. In *Proc Symposium of the Austrian HCI and Usability Engineering Group*, pages 391–402. Springer. 141

Mun, S., Park, S., Lee, Y., and Ko, H. (2016). Deep neural network bottleneck feature for acoustic scene classification. *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*. 66

Nakamura, M., Iwano, K., and Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech and Language*, 22(2):171–184. 70

Nakatani, S., Saiki, S., and Nakamura, M. (2018). Integrating 3d facial model with person-centered care support system for people with dementia. In *Proc International Conference on Intelligent Human Systems Integration*, pages 216–222. Springer. 141

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., and Chertkow, H. (2005). The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699. 21

Nguyen, Q. B., Gehring, J., Kilgour, K., and Waibel, A. (2013). Optimizing deep bottleneck feature extraction. In *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF International Conference on*, pages 152–156. IEEE. 66

NHS (2018). What are the treatments for dementia? https://www.nhs.uk/conditions/dementia/treatment [Online; accessed 2 June 2018]. 17

Novotney, S. and Schwartz, R. (2009). Analysis of low-resource acoustic model self-training. In *Proc 10th Conference of the International Speech Communication Association*. 80

Oba, H., Sato, S., Kazui, H., Nitta, Y., Nashitani, T., and Kamiyama, A. (2018). Conversational assessment of cognitive dysfunction among residents living in long-term care facilities. *International psychogeriatrics*, 30(1):87–94. 4, 27

Okuno, H. G., Ogata, T., and Komatani, K. (2007). Computational auditory scene analysis and its application to robot audition: Five years experience. In *Proc 2nd International Conference on Informatics Research for Development of Knowledge Society Infrastructure*, pages 69–76. IEEE. 71

Otterson, S. and Ostendorf, M. (2007). Efficient use of overlap information in speaker diarization. *Automatic Speech Recognition & Understanding*, pages 683–686. 106

Pakhomov, S. V., Eberly, L., and Knopman, D. (2016). Characterizing cognitive performance in a large longitudinal study of aging with computerized semantic indices of verbal fluency. *Neuropsychologia*, 89:42–56. 142

Pakhomov, S. V. and Hemmy, L. S. (2014). A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the nun study. *Cortex*, 55:97–106. 142

Park, D., Kim, S., Lee, J., Choo, J., Diakopoulos, N., and Elmqvist, N. (2018). Conceptvector: text visual analytics via interactive lexicon building using word embedding. *IEEE trans. on visualization and computer graphics*, 24(1):361–370. 126

Parsons, S., Rego, D. M., Shawe-Taylor, J., Firth, N. C., Primativo, S., Crutch, S. J., Shakespeare, T. J., Slattery, C. F., Macpherson, K., Carton, A. M., et al. (2017). Modelling eye-tracking data to discriminate between alzheimer's patients and healthy controls. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 13(7):P597–P598. 39

Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proc 6th Conference of the International Speech Communication Association*. xix, 78, 79

Pedregosa, F. and Varoquaux, G. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830. 48, 50

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proc EMNLP*, pages 1532–1543. 125

Perkins, L., Whitworth, A., and Lesser, R. (1998). Conversing in dementia: A conversation analytic approach. *Journal of Neurolinguistics*, 11:33–53. 4, 26, 47, 56

Petridis, S. and Pantic, M. (2016). Deep complementary bottleneck features for visual speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 2304–2308. IEEE. 66

Potkins, D., Myint, P., Bannister, C., Tadros, G., Chithramohan, R., Swann, A., O'Brien, J., Fossey, J., George, E., Ballard, C., et al. (2003). Language impairment in dementia: impact on symptoms and care needs in residential homes. *International Journal of Geriatric Psychiatry*, 18(11):1002–1006. 17

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *Proc IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE. 87

Povey, D., Kingsbury, B., Mangu, L., Saon, G., Soltau, H., and Zweig, G. (2005). fmpe: Discriminatively trained features for speech recognition. In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 1, pages I–961. IEEE. 66

Prince, S. J. and Elder, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In *Proc 11th International Conference on Computer Vision*, pages 1–8. IEEE. 104

Przybocki, M. A., Martin, A. F., and Le, A. N. (2006). Nist speaker recognition evaluation chronicles-part 2. In *Proc IEEE Speaker and Language Recognition Workshop*, pages 1–6. IEEE. 108

Quinn, J. F. (2014). *Dementia*. John Wiley & Sons Ltd. 15, 16, 20, 21

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proc IEEE*, volume 77, pages 257–286. IEEE. 66, 68

Rajnoha, J. and Pollák, P. (2011). Asr systems in noisy environment: Analysis and solutions for increasing noise robustness. *Radioengineering*, 20(1):74–83. 66

Raven, J. C. (1995). *Coloured Progressive Matrices Sets A, Ab, B. Manual Sections 1 & 2*. Oxford Psychologists Press. 42

Rawat, A. and Mishra, P. K. (2015). Emotion recognition through speech using neural network. *International Journal*, 5(5). 71

Reisberg, B., Doody, R., Stöffler, A., Schmitt, F., Ferris, S., and Möbius, H. J. (2003). Memantine in moderate-to-severe Alzheimer's disease. *New England Journal of Medicine*, 348(14):1333–1341. 17

Renals, S., Hain, T., and Bourlard, H. (2010). Recognition and understanding of meetings. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1–9. 83, 107

Rey, A. (1964). *Lexamen clinique en psychologie*. Presses universitaires de France., 2nd edition. 42

Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17(1-2):91–108. 104

Rizzo, A. A., Lucas, G. M., Gratch, J., Stratou, G., Morency, L.-P., Chavez, K., Shilling, R., and Scherer, S. (2016). Automatic behavior analysis during a clinical interview with a virtual human. In *MMVR*, pages 316–322. 141

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., and Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE transactions on audio, speech, and language processing*, 19(7):2081–2090. 35

Robinson, T., Fransen, J., Pye, D., Foote, J., and Renals, S. (1995). Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition. In *Proc International Conference on Acoustics Speech and Signal Processing*, volume 1, pages 81–84. IEEE. 83

Royal College of Psychiatrists, R. (2016). National audit of memory clinics 2014. http://www.rcpsych.ac.uk/memoryclinicsaudit. Accessed on March 26, 2016. 55

Rus-Calafell, M., Gutiérrez-Maldonado, J., and Ribas-Sabaté, J. (2014). A virtual reality-integrated program for improving social skills in patients with schizophrenia: a pilot study. *Journal of behavior therapy and experimental psychiatry*, 45(1):81–89. 141

Sacks, H., Schegloff, E., and Jefferson, G. (1974). A simplest systematics for the organization of turn taking for conversation. *Language*, 50:696–735. 4, 24

Sadjadi, S. O., Kheyrkhah, T., Tong, A., Greenberg, C., Reynolds, E. S., Mason, L., and Hernandez-Cordero, J. (2017). The 2016 nist speaker recognition evaluation. In *Proc INTERSPEECH*, pages 1353–1357. ISCA. 106

Sak, H., Senior, A., and Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*. xix, 76, 77

Salamon, J. and Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283. 66

Salamon, J., Bello, J. P., Farnsworth, A., and Kelling, S. (2017). Fusing shallow and deep learning for bioacoustic bird species classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 141–145. IEEE. 66

Salton, G. (1983). *Introduction to modern information retrieval*. New York, London, McGraw-Hill. 44

Sani, A., Lestari, D. P., and Purwarianti, A. (2015). Filled pause detection in indonesian spontaneous speech. In *Proc International Conference of the Pacific Association for Computational Linguistics*, pages 54–64. Springer. 70

Saon, G. and Chien, J.-T. (2012). Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances. *IEEE Signal Processing Magazine*, 29(6):18–33. 64, 68

Saon, G., Kuo, H.-K. J., Rennie, S., and Picheny, M. (2015). The ibm 2016 english conversational telephone speech recognition system. In *Proc INTERSPEECH*, pages 3–7. ISCA. 85

Satt, A., Sorin, A., Toledo-Ronen, O., Barkan, O., Kompatsiaris, I., Kokonozi, A., and Tsolaki, M. (2013). Evaluation of speech-based protocol for detection of early-stage dementia. In *Proc INTERSPEECH*, pages 1692–1696. ISCA. 37

Schmidtke, K., Pohlmann, S., and Metternich, B. (2008). The syndrome of functional memory disorder: definition, etiology, and natural course. *Am J Geriatr Psychiatry*, 16:981–8. 19

Seide, F., Li, G., and Yu, D. (2011). Conversational speech transcription using context-dependent deep neural networks. In *Proc INTERSPEECH*, pages 437–440. ISCA. 68, 85

Sell, G. and Garcia-Romero, D. (2014). Speaker diarization with plda i-vector scoring and unsupervised calibration. In *Proc Spoken Language Technology Workshop*, pages 413–417. IEEE. 107

Sell, G. and Garcia-Romero, D. (2015). Diarization resegmentation in the factor analysis subspace. In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 4794–4798. IEEE. 108

Sell, G., Snyder, D., McCree, A., Garcia-Romero, D., Villalba, J., Maciejewski, M., Manohar, V., Dehak, N., Povey, D., Watanabe, S., et al. (2018). Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural dihard challenge. In *Proc INTERSPEECH*, pages 2808–2812. 108, 111

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611. 51

Shen, Q., Wang, Z., and Sun, Y. (2017). Sentiment analysis of movie reviews based on cnn-blstm. In *Proc International Conference on Intelligence Science*, pages 164–171. Springer. 126

Shriberg, E. (2005). Spontaneous speech: How people really talk and why engineers should care. In *Proc INTERSPEECH*, pages 1781–1784. ISCA. 35, 70

Sidnell, J. and Stivers, T. (2012). *The Handbook of Conversation Analysis*. Wiley-Blackwell. 4, 24, 25

Siegler, M., Jain, U., Raj, B., and Stern, R. M. (1997). Automatic segmentation, classification and clustering of broadcast news audio. *Proceedings of the Speech Recognition Workshop*, pages 97–99. 104, 105

Sinclair, M. and King, S. (2013). Where are the challenges in speaker diarization? In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 7741–7745. IEEE. 109

Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., and Khudanpur, S. (2016). Deep neural network-based speaker embeddings for end-to-end speaker verification. In *Proc Spoken Language Technology Workshop*, pages 165–170. IEEE. 108

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21. 125

Spitzer, R. L., Kroenke, K., Williams, J. B., and Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097. 23

Stringer, G., Couth, S., Brown, L., Montaldi, D., Gledson, A., Mellor, J., Sutcliffe, A., Sawyer, P., Keane, J., Bull, C., et al. (2018). Can you detect early dementia from an email? a proof of principle study of daily computer use to detect cognitive and functional decline. *International journal of geriatric psychiatry*. 39

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *J Exp Psychol*, 18(6):643. 42

Sundermeyer, M., Schlüter, R., and Ney, H. (2012). Lstm neural networks for language modeling. In *Proc 13th Conference of the International Speech Communication Association*. 76, 78

Swietojanski, P. and Renais, S. (2016). Sat-lhuc: Speaker adaptive training for learning hidden unit contributions. In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5010–5014. IEEE. 86

Tanaka, H., Adachi, H., Ukita, N., Ikeda, M., Kazui, H., Kudo, T., and Nakamura, S. (2017). Detecting dementia through interactive computer avatars. *IEEE journal of translational engineering in health and medicine*, 5:1–11. 141

Tang-Wai, D. F. and Graham, N. L. (2008). Assessment of Language Function in Dementia. *Geriatrics and Aging*, 11(2):103–110. xxiii, 2, 18

Tao, J., Chen, L., and Lee, C. M. (2016). Dnn online with ivectors acoustic modeling and doc2vec distributed representations for improving automated speech scoring. In *Proc INTERSPEECH*, pages 3117–3121. ISCA. 126

Taylor, A., Marcus, M., and Santorini, B. (2003). The penn treebank: an overview. In *Treebanks*, pages 5–22. Springer. 124

Thomas, C., Keselj, V., Cercone, Rockwood, K., and Asp, E. (2005). Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In *Proc International Conference on Mechatronics & Automation*, pages 1569–1574. IEEE. 37

Tilk, O. and Alumäe, T. (2015). Lstm for punctuation restoration in speech transcripts. In *Proc 6th Conference of the International Speech Communication Association*. 70

Toth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatloczki, G., Biro, E., Zsura, F., Pakaski, M., and Kalman, J. (2015). Automatic detection of mild cognitive impairment from spontaneous speech using asr. In *Proc INTERSPEECH*. ISCA. 35

Toth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatloczki, G., Banreti, Z., Pakaski, M., and Kalman, J. (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research*, 15(2):130–138. 36

Tran, M. K. P., Robert, P., and Bremond, F. (2016). A virtual agent for enhancing performance and engagement of older people with dementia in serious games. In *Proc Workshop Artificial Compagnon-Affect-Interaction*. 141

Tranter, S. E. and Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1557–1565. 101

Verkhodanova, V. and Shapranov, V. (2015). Multi-factor method for detection of filled pauses and lengthenings in russian spontaneous speech. In *Proc International Conference on Speech and Computer*, pages 285–292. Springer. 70

Veselỳ, K., Ghoshal, A., Burget, L., and Povey, D. (2013). Sequence-discriminative training of deep neural networks. In *Proc INTERSPEECH*, pages 2345–2349. ISCA. 74

Vijayasenan, D. and Valente, F. (2012). Diartk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings. In *Proc INTERSPEECH*. ISCA. 105

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269. 69

Vu, N. T., Kraus, F., and Schultz, T. (2011). Rapid building of an asr system for under-resourced languages based on multilingual unsupervised training. In *Twelfth Annual Conference of the International Speech Communication Association*. 81

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. J. (1990). Phoneme recognition using time-delay neural networks. In *Readings in speech recognition*, pages 393–404. Elsevier. 78

Wald, M. (2013). An exploration of the potential of automatic speech recognition to assist and enable receptive communication in higher education. *Approaches to Developing Accessible Learning Experiences: Conceptualising Best Practice*, page 9. 70

Walker, T., Christensen, H., Mirheidari, B., Swainston, T., Rutten, C., Mayer, I., Blackburn, D., and Reuber, M. (2018). Developing an intelligent virtual agent to stratify people with cognitive complaints: A comparison of human–patient and intelligent virtual agent–patient interaction. *Dementia*, page 1471301218795238. 10

Wang, L., Zhang, C., Woodland, P., Gales, M., Karanasou, P., Lanchantin, P., Liu, X., and Qian, Y. (2016). Improved dnn-based segmentation for multi-genre broadcast audio. In *Proc International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5700–5704. IEEE. 109

Wechsler, D. (1945). Wechsler memory scale. *Psychological Corporation*. 23

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale*. The Psychological Corporation., 3rd edition. 42

Wechsler, D. (2014). Wechsler adult intelligence scale–fourth edition (wais–iv). *San Antonio, Texas: Psychological Corporation*. 22, 23

Weiner, J., Angrick, M., Umesh, S., and Schultz, T. (2018). Investigating the effect of audio duration on dementia detection using acoustic features. In *Proc INTERSPEECH*, pages 2324–2328. ISCA. 6, 38, 41, 175

Weiner, J., Engelbart, M., and Schultz, T. (2017). Manual and automatic transcriptions in dementia detection from speech. In *Proc INTERSPEECH*, pages 3117–3121. ISCA. 38

Weiner, J., Herff, C., and Schultz, T. (2016). Speech-based detection of Alzheimerś disease in conversational german. In *Proc INTERSPEECH*, pages 1938–1942. ISCA. 37

Williams, R. J. and Peng, J. (1990). An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural computation*, 2(4):490–501. 75

Wu, S., Zhang, D., Zhou, M., and Zhao, T. (2015). Efficient disfluency detection with transition-based parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 495–503. 70

Wu, X. and Matsumoto, Y. (2014). A hierarchical word sequence language model. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*. 91

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. (2016). The microsoft 2016 conversational speech recognition system. *arXiv preprint arXiv:1609.03528*. 85

Yancheva, M., Fraser, K., and Rudzicz, F. (2015). Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. *6th Workshop on Speech and Language Processing for Assistive Technologies*. 38

Yancheva, M. and Rudzicz, F. (2016). Vector-space topic models for detecting Alzheimer's disease. In *Proc 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2337–2346. 39

Yella, S. H. and Bourlard, H. (2014). Overlapping Speech Detection Using Long-Term Conversational Features for Speaker Diarization in Meeting Room Conversations. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(12):1688 – 1700. 107, 111

Yella, S. H. and Valente, F. (2012). Speaker diarization of overlapping speech based on silence distribution in meeting recordings. In *Proc INTERSPEECH*, pages 1–4. ISCA. 106

Yenter, A. and Verma, A. (2017). Deep cnn-lstm with combined kernels from multiple branches for imdb review sentiment analysis. In *Proc 8th Ubiquitous Computing, Electronics and Mobile Communication Conference*, pages 540–546. IEEE. 126

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. (2006). *The HTK Book Steve*. Cambridge University Engineering Department. xix, 67

Yu, D. and Seltzer, M. L. (2011). Improved bottleneck features using pretrained deep neural networks. In *Twelfth annual conference of the international speech communication association*. 66

Yu, Z., Yanqing, S., Jianping, Z., and Yonghong, Y. (2009). Speech emotion recognition system using both spectral and prosodic features. *Information Engineering and Computer Science, ICIECS 2009*, pages 1–4. 71

Zajıc, Z., Hrúz, M., and Müller, L. (2017). Speaker diarization using convolutional neural network for statistics accumulation refinement. In *Proc INTERSPEEH*. 110

Zavaliagkos, G. and Colthurst, T. (1998). Utilizing untranscribed training data to improve performance. In *DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne*. Citeseer. 80

Zhang, Y., Wilcockson, T., Kim, K. I., Crawford, T., Gellersen, H., and Sawyer, P. (2016). Monitoring dementia with automatic eye movements analysis. In *Proc Intelligent Decision Technologies*, pages 299–309. Springer. 39

Zhou, L., Fraser, K. C., and Rudzicz, F. (2016). Speech recognition in Alzheimerś disease and in its assessment. In *Proc INTERSPEECH*, pages 1948–1952. ISCA. 38

Appendix A

Conversation Analysis Symbols

Table A.1: *Some common symbols of Conversation Analysis (CA) (Lerner [2004]).*

Symbol	Name	Meaning
[	Left bracket	point of overlap onset, e.g. A:how t[all are you, Al B: [How tall'r you Al
]	Right bracket	end point of two overlapped utterances
=	Equal sign	no break or gap, one at the end of one line an another at the beginning of a next, e.g A: thirty five pounds= B:=AAUUGH
(0.0)	Numbers on parentheses	gap time by tenth of second, e.g. (0.2) for 0.2 second silence
(.)	Dot in parentheses	a brief interval, normally less than a tenth of second within or between utterances
_	Underscoring	stress via pitch and/or amplitude, e.g. A: Well D<u>ea</u>n has: uh:,h <u>totally</u> coop'rated with the U.S. Att<u>o</u>rney.
:	Colons	prolongation of the immediately prior sound, e.g [W o: : : : :] w
:_	Combination of colon and underscore	intonation contours, when underscore followed by colon it indicates up-to-down contour and vise a verse down-to-up contour, e.g. w<u>o</u>:rd comparing with wo:<u>r</u>d
↑↓	Arrows	Shifts into high or low pitch, e.g. ↑↑Th<u>an</u>k ↓you.
.,?	Punctuation marks	for usual intonation
WORD	Uppercase	loud sounds
°word°	Degree sign	bracketing around utterance indicate sounds are ofter that other parts
*	Asterisk	percussive non-speech sounds or creaky voice

Table A.2: *CA symbols continue..*

Symbol	Name	Meaning
t^*, d^*	An asterisk following a consonant	hardener, e.g. *thet** which is tantalised and hard version of that
t,d	Bold consonant	hardener e.g. i**t**
>word	A pre-positioned left carat	a hurried start for self-repair
word<	A post-positioned left carat	while the word is fully completed, it seems to stop suddenly
–	Dash	Cut off, e.g. I get- I get sick behind it.
><	Right/left carats bracketing	speeding up utterance or a part of utterance
<>	Left/right carats bracketing	speeding down utterance or a part of utterance
.hhh	A dot-prefixed row of 'h's	in breath, without dot, 'h's indicate an out breath
(h)	Parenthesised 'h'	plosiveness
()	Empty parentheses	transcriber was unable to get what said
(( ))	Doubled parentheses	transcriber's descriptions

Appendix B

General guidelines of the Hallamshire study

As the main aim of these guidelines is to elicit the descriptions of patients and accompanying others and to allow them to develop their own way of describing their subjective memory problems, doctors are asked to choose an opening which does not make any direct reference to memory problems, e.g. by starting with an enquiry such as "what can I do for you today?". This question creates an open initial phase, in which patients can set out their own agenda. During the next phase of the encounter, patients are prompted to describe events characterised by particularly memorable cognitive problems. This approach encourages patients to reconstruct subjective experiences (like the first time they realised something was wrong, or any examples of recent episodes of memory failure which were particularly significant or embarrassing for them). It is anticipated that the first 15-20 minutes of the clinical encounter will be devoted to this open questioning. During this part of the interview doctors are discouraged from interrupting patients or accompanying persons, to ask additional questions (other than for clarification), or to introduce new topics in the discussion. In a third and final part of the consultation the doctor can ask any relevant questions which have not been addressed.

Sample questions:

- Why have you come to clinic today and what are your expectations?
- Tell me about a problem with your memory that you found particularly embarrassing?
- Tell me about the most recent time that your memory failed?

- Is there anything else about your memory that you need to tell me?