Feasibility and utility of automated processing of speech-based cognitive assessments for early Alzheimer's Disease



Alankar Aryal Atreya

Lincoln College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hillary 2023

Dedicated to,

Vishnu Raj Atreya (1942 - 2020)

You continue to inspire.

Acknowledgement

Many people have helped me in their ways to complete my DPhil thesis. Firstly, I would like to thank my supervisors, Dr Chris Hinds, Professor Sarah Parish, and Associate Professor Marion Mafham, for their time and encouragement with scientific advice and for supporting me through twists, turns and pandemics during my DPhil study. I would also like to thank Dr Claire Lancaster and Professor Cornelia van Duijn for their valuable input to my research.

Thank you to Eli Lilly & Company and the Nuffield Department of Population Health for supporting my research. I should also thank the PREVENT and GameChanger study participants and the Mezurio smartphone development team, especially Simon Bond, who was always available for my questions on the Mezurio app's technical and developmental details.

This DPhil journey had many difficult moments: my father passed away due to cancer and a disturbance in my research due to the COVID-19 pandemic. I want to thank my family in Nepal, Cambodia, and the Netherlands, especially my mom, sisters, brothers, nieces, nephews, and in-laws, for their unwavering love, support and encouragement. Thank you to the wonderful friends in the Big Data Institute, including Devesh Batra, Diego Aguiler-Ramirez, Nomi Carlebach, David Shaw, Shing Chan, Sharib Ali, and Sunil Pokhrel, for amazing conversations on science. Thank you to many old and new friends, too many to name individually, but especially Andrija, Kamal, Prabin, Rakshya, Ayush, Mita, Bhawana, Niranjan, Pawan, Sonu, Agya, Shrochis, Gavin, Varsha, Pratik, Nivedita, Abhishek, Serene, Leela, Kailash, Bipin, Aradhana, and everyone from the Oxford Nepali Community. Thank you to all my teachers and mentors who have taught, motivated and supported me to get me here. Thanks especially to Associate Professor Lukas Esterle and Dr Anil Alexander, who guided me into this field of scientific research.

I would also like to thank my daughter Kavya, who joined us during this journey, providing me with happiness and pleasure.

Finally, to my lovely wife, Pramila, I would not even imagine doing it without you - or at least I would not have been motivated to. Your patience, understanding, and support helped me maintain my focus to complete my research. I love you!

Abstract

Feasibility and utility of automated processing of speech-based cognitive assessments for early Alzheimer's Disease

Alankar Aryal Atreya, Lincoln College, University of Oxford
Submitted for the degree of Doctor of Philosophy, Hilary Term 2023

Subtle changes in speech and language can be one of the earlier signs of Alzheimer's disease (AD) that is evident even decades before the appearance of observable symptoms leading to the AD diagnosis. Smartphone applications enable the frequent, extensive collection of raw speech data from a large population remotely and target specific behaviours that correlate with the disease. There is a need for automated and scalable processing of these large data to extract meaningful features for further analysis. Also, the newly developed measure's feasibility and utility must be addressed before it can be widely adopted.

A literature review of studies undertaking audio and language processing for cognitive assessment identified 131 studies and the range of techniques used. The study sizes were very small (75% of studies with < 100 participants; the largest sample size of 521), and the data were usually manually processed. Some of the evidence gaps identified were the lack of a standardised feature set with both audio and language features and the limited use of computational methods for scalability and automation.

Data available from the healthy PREVENT (n = 25) and GameChanger (n = 11239) study participants contained audio recordings of story narration (learning) from pictorial cues (18 stories in PREVENT and 7 stories in GameChanger Year-1 and an additional 5 stories in a subset of participants in Year-2) and their retellings from memory immediately and 24 hours later using the Story Time task available within the Mezurio smartphone app. In this thesis, a scalable and cohesive automated feature extraction pipeline was developed utilising automatic speech recognition (ASR) systems, natural language processing (NLP), audio signal processing (ASP), and artificial intelligence/machine learning (AI/ML) to extract audio and language features from speech recordings. A systematic comparison of transcription accuracy across six different ASR systems showed that Microsoft Azure transcription service provided transcripts with a lower word error rate (WER = 0.16) compared to other ASR systems, including KALDI

(WER = 0.33), which has been mostly used in AD research. Therefore, the ASR system provided by Microsoft Azure was included in the pipeline to convert speech to text. The quality of the information provided by the participants was automatically scored as Automated Information Units (AIU) considering NLP methods, including semantic neighbourhood and part-of-speech tags. A protocol to provide a human-derived measure of information unit recall [Manual Information Unit (MIU) score] was developed and applied in PREVENT to compare the quality of AIU scores. A linear mixed model (adjusted for age, sex, and education, with participants and stories as random effects) was used to estimate the change in about 50 pipeline speech-based features cross-sectionally. Associations between Story Time pipeline features and other cognitive measures (in-clinic test, Gallery Game, and Subjective Cognitive Decline scores) available within PREVENT and GameChanger were evaluated. The association of pipeline features with well-established socio-demographic variables associated with dementia risk, including age, gender, education, family history of dementia, and deprivation, were examined. In addition, stepwise regression was used to identify the informative speech-based features associated with measures of cognition.

In the pilot of 25 participants in PREVENT, the correlation between AIU and MIU scores was very high (r = 0.98). Of about 18000 participants in GameChanger, 11239 attempted the Story Time task. Older people (60-95) in the study were more committed to the task, and on average, they fully completed more stories (4.6 stories with a full set of the subtasks learning, immediate recall, and delayed recall) than the younger group (16-39: 3.2 stories; 40-59: 3.6 stories) out of 7 stories in Year-1. The test-retest reliability (Cronbach alpha) of learning, immediate, delayed, and average AIU recall scores were between 0.70-0.79, showing moderate to good reliability. Among those with a full set of sub-tasks, the AIU score was significantly lower at 24-hr delayed recall (Mean ± SD = 0.42 ± 0.14) than immediate recall (Mean ± SD = 0.50 ± 0.12) ($p < 10^{-300}$). The AIU recall score was one of the strong pipeline predictors of the Gallery Game recall score and subjective cognitive decline measure. In the sensitivity analysis, when only the audio features were considered that would be directly applicable to studies in other languages, the model explained only 4% of the variation of Gallery Game recall accuracy scores compared to 7%, using all the pipeline features, highlighting the importance of language features, but still obtaining half of the value by using audio features alone. Association of AIU recall scores with age and education were in the expected direction. Several pipeline features, including lexical, syntactic, speech fluency, and voice quality, were also statistically significantly associated with age, education, and deprivation.

In this thesis, a novel automated feature extraction pipeline was developed using various emerging technologies to extract speech-based features of potential relevance to dementia using speech data collected via a smartphone application. This pipeline has the potential to facilitate much larger studies and trials through its scalability and wide deployment. AIU recall score showed expected associations with other cognitive scores and demographics, demonstrating its utility as a clinical assessment method for examining the performance of participants in a story-retelling task. However, the association of the score with stages of dementia could not be assessed. The Mezurio app is currently being used in other European studies (some of which include the potential for comparison with dementia status), and the participants provide speech responses in their native language. ASR in non-English languages is improving. Therefore, this pipeline could be readily adapted to produce these potentially useful language features in those studies allowing further validation of the methods, but the audio feature is directly applicable.

Table of contents

LIST OF TABLES ..... **XIII**

LIST OF FIGURES ..... **XVIII**

LIST OF ABBREVIATIONS **23**

DECLARATION **26**

IMPACT OF THE COVID-19 PANDEMIC **27**

CHAPTER 1. INTRODUCTION **28**

1.1 EARLY DETECTION OF ALZHEIMER'S DISEASE 28

1.1.1 The burden of Alzheimer's disease *28*

1.1.2 Existing methods of AD detection *30*

1.1.3 The purpose of early detection *31*

1.1.4 The potential of digital tools for early AD detection *32*

1.2 SPEECH AND LANGUAGE IN ALZHEIMER'S DISEASE 34

1.2.1 Progression of speech and language changes in AD *35*

1.2.2 Speech elicitation methods in AD research *37*

1.2.3 Audio and language features from narrative speech *39*

1.2.4 Relevance of audio and language on other forms of dementia *40*

1.3 MEZURIO STORY TIME TASK 41

1.3.1 Background of Mezurio Story Time task *41*

1.3.2 Task design *42*

1.3.3 Task stimuli *43*

1.4 POTENTIAL ROLES OF SPEECH-BASED ASSESSMENT METHODS 44

1.5 AIMS AND OBJECTIVES OF THE THESIS 45

1.6 THESIS OUTLINE 46

CHAPTER 2. AUDIO AND LANGUAGE FEATURES IN AD **48**

2.1 INTRODUCTION 48

2.2 AUDIO FEATURES 49

2.2.1 Speech fluency features *49*

2.2.2 Spectral and voice quality features *50*

2.3 LANGUAGE FEATURES 52

2.3.1 Lexical and syntactic features *52*

2.3.2 Semantic features *54*

2.3.3 Lexical density *55*

2.3.4 Psycholinguistic features *56*

2.4 SUMMARY 57

CHAPTER 3. LITERATURE REVIEW ON AUDIO AND LANGUAGE PROCESSING FOR COGNITIVE ASSESSMENT 58

3.1 LITERATURE REVIEW 58

3.1.1 Objectives *58*

3.1.2 Search strategy *58*

3.1.3 Study selection and data extraction *61*

3.1.4 Study characteristics *64*

3.1.5 Use of computational methods and tools *80*

3.1.6 Informative speech and language features *83*

3.1.7 Disease-specific associations with audio and language features *84*

3.2 LIMITATIONS IN THE EXISTING LITERATURE 89

3.2.1 Lacking socio-demographic diversity *89*

3.2.2 Labour-intensive manual coding and feature extraction *89*

3.2.3 Speech task for collecting comprehensive language samples *90*

3.2.4 Limited ability to aggregate results across studies *90*

3.3 DISCUSSION 91

3.4 SUMMARY 93

CHAPTER 4. DESIGN AND COHORT CHARACTERISTICS OF THE PREVENT AND GAMECHANGER STUDIES 95

4.1 PREVENT 95

4.1.1 Study design *95*

4.1.2 Recruitment methods *96*

4.1.3 Baseline assessment *96*

4.1.4 Smartphone-based assessments *97*

4.1.5 Repeated assessments *97*

4.1.6 Baseline characteristics of PREVENT study participants *99*

4.1.7 Development of Manual Information Unit (MIU) scoring *100*

4.2 GAMECHANGER 104

4.2.1 Study design *104*

4.2.2 Recruitment methods *104*

4.2.3 Demographics and other questionnaires *105*

4.2.4 Smartphone-based assessments *108*

4.2.5 Repeated assessment *109*

4.2.6 Baseline characteristics of GameChanger participants *111*

4.3 SUMMARY 114

CHAPTER 5. DEVELOPING A FEATURE EXTRACTION PIPELINE TO DERIVE AUDIO AND LANGUAGE FEATURES FROM RAW SPEECH DATA 115

5.1 INTRODUCTION 115

5.2 A SYSTEMATIC COMPARISON OF ASR SERVICES AND EVALUATION OF THEIR TRANSCRIPTION RESULTS ON A SUBSET OF THE PREVENT DATASET 119

5.2.1 Automatic Speech Recognisers *120*

5.2.2 ASR optimisation techniques *122*

5.2.3 Methods *125*

5.2.4 Comparison results *128*

5.2.5 Benefits and limitations of different ASR systems and the utility of Word Error Rate (WER) as a transcription accuracy measure *132*

5.3 EVALUATION OF THE PERFORMANCE OF RETROFITTED WORD EMBEDDING MODELS USING A SUBSET OF THE PREVENT DATASET 133

5.4 FINE-TUNING THE BERT MODEL FOR SURROGATE PHRASE DETECTION AND EVALUATING ITS PERFORMANCE ON A SUBSET OF THE PREVENT DATASET 133

5.4.1 Methods *134*

5.4.2 Results *138*

5.5 FEATURE ENGINEERING: DERIVING FEATURES USING A COMBINATION OF COMPUTATIONAL METHODS 141

5.5.1 Semantic similarity-based Automated Information Unit (AIU) memory score *141*

5.5.2 Lexical, syntactic, and audio features from transcripts *143*

5.5.3 Audio features *144*

5.6 END-TO-END FEATURE EXTRACTION PIPELINE 146

5.6.1 Data ingestion *146*

5.6.2 Data pre-processing *146*

5.6.3 Feature extractor *147*

5.6.4 Postprocessing and reformatting, and storage of features *147*

5.7 IMPROVING PERFORMANCE AND SCALABILITY WITH PARALLEL AND DISTRIBUTED COMPUTING 148

5.8 SUMMARY 149

CHAPTER 6. PILOTING THE METHODS ON THE PREVENT DATASET 151

6.1 INTRODUCTION 151

6.2 METHODS 151

6.2.1 Study population *151*

6.2.2 Pipeline features derived from the Story Time task *152*

6.2.3 In-clinic neuropsychological tests *152*

6.2.4 Selection of confounders *152*

6.2.5 Statistical analyses *153*

6.3 RESULTS 158

6.3.1 Participant's engagement with the Story Time task *159*

6.3.2 Distributions of MIU scores on recall tasks *162*

6.3.3 Influence of the story stimuli on MIU recall scores *164*

6.3.4 Influence of 24-hr delay between recalls on MIU scores *165*

6.3.5 Association between MIU and AIU scores *167*

6.3.6 Effect of 24-hr delay between recalls on AIU scores *172*

6.3.7 Exploring associations between AIU recall scores, audio, and language features *173*

6.3.8 Construct validity of Story Time task *176*

6.4 DISCUSSION 179

6.4.1 Overall findings *180*

6.4.2 Strengths and limitations *182*

6.5 SUMMARY 183

CHAPTER 7. EVALUATING METHODS IN GAMECHANGER: A POPULATION-BASED STUDY 184

7.1 INTRODUCTION 184

7.2 METHODS 185

7.2.1 GameChanger *185*

7.2.2 Digital features *185*

7.2.3 Selection of confounders *186*

7.2.4 Statistical analyses *187*

7.3 RESULTS 191

7.3.1 Participant's compliance with the Mezurio schedule *191*

7.3.2 Influence of region on AIU scores *196*

7.3.3 Test-retest reliability of AIU scores *197*

7.3.4 Influence of 24-hr delay and time of the day on AIU recall scores *198*

7.3.5 Exploring associations between AIU recall scores, language, and audio features *201*

7.3.6 Associations of pipeline features with other Mezurio scores *206*

7.3.7 The association of participant's sociodemographic factors with the Story Time pipeline features *221*

7.4 DISCUSSION 239

7.4.1 Overall findings *239*

7.4.2 Strengths and limitations *246*

7.5 SUMMARY 247

CHAPTER 8. DISCUSSION 248

8.1 INTRODUCTION 248

8.2 SCALABILITY OF COMPUTATIONAL METHODS IN A LARGE DATASET 249

8.3 EVIDENCE OF FEASIBILITY 250

8.4 FACE VALIDITY 252

8.4.1 Association of AIU recall scores with in-clinic tests and other cognitive tests within the Mezurio app *252*

8.4.2 Association of AIU recall scores with the pipeline features *252*

8.4.3 Associations of pipeline features with the sociodemographic variables and other Mezurio tests *253*

8.4.4 Applicability of the feature extraction pipeline *254*

8.4.5 Preliminary analysis of audio features in the RADAR-AD cohort *255*

8.5 STRENGTHS AND LIMITATIONS OF THE STUDY 259

8.6 IMPLICATIONS FOR FUTURE RESEARCH 260

8.6.1 Validation of Story Time outcomes in cohorts with established biomarkers and cognitive measures *261*

8.6.2 Assessing the feasibility of the Story Time data extraction pipelines and outcome measures in other languages *262*

8.6.3 Assessing the value of Story Time reading variation *262*

8.7 SUMMARY 263

APPENDIX A. SUPPLEMENTARY TABLES AND FIGURES 264

A.1 SUPPLEMENTARY TABLES 264

A.2 SUPPLEMENTARY FIGURES 282

APPENDIX B. SUPPLEMENTARY DATA FILE 295

APPENDIX C. RADAR-AD COHORT 296

REFERENCES 299

List of Tables

Table 1.1 Advantages and disadvantages of cognitive assessments of AD using digital tools. 33

Table 1.2 Progression of speech and language changes in different stages of AD. 36

Table 1.3 Speech elicitation methods used for speech-based analysis in AD. 38

Table 3.1 Search terms used in different databases to identify studies reporting speech-based cognitive assessment on Alzheimer's Disease. 59

Table 3.2 Characteristics of studies analysing speech-based features for different stages of Alzheimer's Disease. 67

Table 3.3 Audio and language features analysed in different stages of AD. 80

Table 3.4 Informative audio and language features used in the AD research organised by type of the features as reported in the reviewed articles. 84

Table 4.1 Summary of sample characteristics on the PREVENT dataset and performance (Mean ± SE) on in-clinic tests. 99

Table 4.2 Example of a Story Time comic with the list of expected information units. 103

Table 4.3 Subjective Cognitive Decline Questionnaire deployed within Mezurio in Game Changer study. 106

Table 4.4 List of assessments deployed within the Mezurio smartphone app for the GameChanger study. 108

Table 4.5 Sociodemographic characteristics of those taking part in GameChanger and those who attempted the Story Time task. 112

Table 4.6 Sociodemographic characteristics of those taking part in year 1 and returning to year 2 of the GameChanger study. 113

Table 5.1 Optimisation methods provided by ASR systems. 123

Table 5.2 Summary of the cost associated per audio minute using various transcription providers. 129

Table 5.3 Summary of the data processing procedures and data security measures used by the ASR system providers.....130

Table 5.4 Top 5 words that are semantically close to the word "boy" before and after the domain adaptation. 133

Table 5.5 Examples of surrogate phrases that indicate difficulty in retelling stories. 134

Table 5.6 Surrogate phrases in the training, validation, and test set after the split. 135

Table 5.7 Description of BIOES tags used in labelling the data. 137

Table 5.8 Surrogate phrase corpus preparation using BIOES tagging scheme. 137

Table 5.9 Comparison of positive predictive value, sensitivity and F1 scores result on BERT models using surrogate phrase corpus developed for this study. 139

Table 5.10 Comparison of positive predictive value, sensitivity and F1 scores on real story recall test datasets. 139

Table 5.11 Audio and language features that are produced from the pipeline. 148

Table 6.1 Definitions of key measures for the analyses. 155

Table 6.2 Summary of sample characteristics on the PREVENT dataset and performance (Mean ± SE) on in-clinic tests. 159

Table 6.3 The proportions of participants completing each Story Time learning, immediate, and delayed recall distributed across the PREVENT assessment periods (baseline, 6-months, 12-months). 160

Table 6.4 Issues with Automated Information Unit extraction, shown as - transcript output (expected information unit) in 50 different story transcripts. 169

Table 6.5 Joint association of lexical and syntactic measures with immediate and delayed AIU recall scores. 173

Table 6.6 Joint association of psycholinguistic measures with immediate and delayed AIU recall scores.....174

Table 6.7 Joint association of speech fluency features with immediate and delayed AIU recall scores.....174

Table 6.8 Joint association of spectral and voice quality features with immediate and delayed AIU recall scores.....175

Table 7.1 Definitions of key outcomes from Mezurio, Story Time, Gallery Game, and Subjective Cognitive Decline (daily and six months).....186

Table 7.2 The mean compliance, task uptake, and the number of participants on each GameChanger Mezurio task across Year-1 and Year-2 assessment periods.....192

Table 7.3 Characteristics of participants on their commitment to Mezurio schedule based on the audio recordings available for the Story Time learning sub-tasks during Year-1 and Year-2 assessment.....193

Table 7.4 Influence of English regions on the AIU scores (at immediate recall sub-task)...197

Table 7.5 Cronbach coefficient of reliability (α) estimate on AIU scores considering first 2, 4, 6 and 7 stories in the GameChanger study, Year-1. 198

Table 7.6 Joint association of lexical and syntactic measures with immediate and delayed AIU recall scores.....202

Table 7.7 Joint association of psycholinguistic measures with immediate and delayed AIU recall scores.....203

Table 7.8 Joint association of speech fluency features with immediate and delayed AIU recall scores.....203

Table 7.9 Joint association of spectral and voice quality features with immediate and delayed AIU recall scores.....204

Table 7.10 Association of the rest of the pipeline features at recalls with AIU Recall score using stepwise regression analysis. 205

Table 7.11 Sensitivity analysis on the association of audio features at recalls with AIU recall scores using stepwise regression analysis. 206

Table 7.12 Association of pipeline features (recalls) with Gallery Game recall accuracy score. 211

Table 7.13 Association of pipeline features at recalls with Gallery Game Recall Accuracy score using stepwise regression analysis. 213

Table 7.14 Sensitivity analysis on the association of pipeline features (audio features) at recalls with Gallery Game recall accuracy score using stepwise regression analysis. 213

Table 7.15 Association of pipeline features (from recalls) with SCD questionnaire scores. 217

Table 7.16 Association of pipeline features at recalls with SCD scores using stepwise regression analysis. 220

Table 7.17 Sensitivity analysis on the association of audio features at recalls with SCD scores using stepwise regression analysis. 221

Table 7.18 The difference in Story Time pipeline features per year of age cross-sectionally. 223

Table 7.19 The difference in Story Time pipeline features per lower education cross-sectionally. 228

Table 7.20 The difference in Story Time pipeline features associated with having any family history of dementia as the predictor variable cross-sectionally. 232

Table 7.21 The difference in Story Time pipeline features per most deprived group as the predictor variable cross-sectionally. 236

Table 8.1 The difference in audio pipeline features on RADAR-AD clinical groups cross-sectionally. 257

Table A.1 PRISMA checklist for the systematic review of studies on audio and language-based cognitive assessment for Alzheimer's Disease (AD).....264

Table A.2 Word Error Rate (WER) for each of the Cloud-based ASR and open-source ASR system's outputs when compared to manual transcriptions. 267

Table A.3 Some examples of surrogate phrases contained in the real story retelling data and were detected by the fine-tuned BERT model.....269

Table A.4 Definition and description of audio and language features extracted from text. .270

Table A.5 Description and definition of audio features. 272

Table A.6 The proportions of participants completing each Story Time learning, immediate, and delayed recall distributed across the GameChanger assessment periods (Year 1 and Year 2). 277

Table A.7 Mean (SD) of each pipeline feature at each of the three subtasks among participants and stories they completed.....278

Table A.8 Baseline characteristics of individuals with and without a family history of dementia.....280

Table A.9 Baseline characteristics of individuals in various socioeconomic deprivation deciles. 281

Table C.1 Demographic characteristic of RADAR-AD participants per clinical group providing Story Time data. 298

List of Figures

Figure 1.1 The continuum of Alzheimer's disease. 29

Figure 1.2 An overview of audio and language features obtained by processing speech recordings in Alzheimer's disease research.....39

Figure 1.3 Representation of a Story Time learning activity; each activity contains nine isolated comic strip frames, which participants are asked to narrate aloud frame-by-frame...43

Figure 2.1 Representation of a test speech recording from the Story Time app and overview of audio features extracted from these signals. 48

Figure 2.2 Taxonomy of audio and language features extracted from speech recordings.....49

Figure 3.1 Flowchart describing the selection of articles in speech-based cognitive assessment on various stages of Alzheimer's Disease. 63

Figure 3.2 The number of articles published in ten years intervals between 1980 to 2020 after the final selection. 64

Figure 4.1 Participants flow diagram from the PREVENT study, data availability, and the number of assessments.....98

Figure 4.2 Representation of (a) learning, (b) recognition, and (c) recall (swipe direction: left, up, and right) in Gallery Game task within the Mezurio app. 109

Figure 4.3 Flow diagram of participation in the GameChanger study and data availability. 110

Figure 5.1 Flowchart of feature extraction pipeline developed for this doctoral research. ..116

Figure 5.2 Plot shows the average Word Error Rate (WER) obtained using different ASR systems and manual transcription outputs. 131

Figure 5.3 Word cloud of surrogate phrases used in Story Time (A) and wallet story (B) retellings.....140

Figure 5.4 Audio feature extraction procedure undergoing various steps such as audio pre-processing, calculating audio intensity, obtaining voiced and unvoiced blocks and locating local maxima within the voiced blocks. 145

Figure 5.5 Comparison of parallel and distributed computing performance using Dask on a machine with 16 cores and 32 GiB memory. 149

Figure 6.1 Proportion of participants completing Story Time subtasks stratified by assessment periods. 161

Figure 6.2 Scatterplot showing the association between compliance and the age of the participants. 162

Figure 6.3 Density plot depicting the distribution of MIU scores extracted from both Story Time immediate and delayed recall tasks where the y-axis shows the probability density estimates. 163

Figure 6.4 Plot showing the distribution of Manual Information Unit (MIU) scores on immediate and delayed recalls across Story Time. 164

Figure 6.5 Plots showing the influence of 24-hr delay between immediate and delayed recall tasks 165

Figure 6.6 Plot showing the beta estimates obtained by regressing MIU scores on age, gender, education, story IUs and recalls using a linear mixed effect model with participants and stories as a random effect. 167

Figure 6.7 Scatter plot of the Manual Information Unit (MIU) scores and Automated Information Units (AIU) scores estimated from PREVENT Story Time data using the KALDI ASR system for transcript generation. 168

Figure 6.8 Scatter plot of the Manual Information Unit (MIU) scores and Automated Information Units (AIU) scores estimated from PREVENT Story Time data using the transcripts obtained from the Microsoft Azure ASR system. 170

Figure 6.9 Bland-Altman agreement analysis plot on multiple observations per participant. 171

Figure 6.10 Plot showing the beta estimates obtained by regressing AIU scores on age, gender, education, story IUs and recalls using a linear mixed-effect model with participants and stories as random effects. 172

Figure 6.11 Scatterplot shows associations between AIU recall scores and in-clinic tests (A) COGNITO and (B) ACE-R from the PREVENT dataset. 177

Figure 6.12 Scatterplot shows associations between delta AIU and in-clinic tests (A) COGNITO and (B) ACE-R from the PREVENT dataset. 178

Figure 6.13 Scatterplot showing associations between participants' age and average delta AIU. Scores are shown in different colours to represent participants. 179

Figure 7.1 Participant's completion of stories by story index stratified by sub-tasks for each story in the GameChanger study. 195

Figure 7.2 The heatmap plot showing AIU immediate recall scores (mean) for twelve Story Time comics by English regions. 196

Figure 7.3 Beta estimates obtained by regressing AIU scores at recall on age, gender, education, story IUs and recall type using a linear mixed model with participants and stories as a random effect. 199

Figure 7.4 The effect of time of the day with AIU scores at immediate and delayed recalls using linear mixed models with participants and stories as random effects and time of the day as a fixed effect adjusted for age, gender, education, and story IU. 200

Figure 7.5 Relationship between AIU recall scores from the Story Time task and Mezurio Gallery Game recall accuracy scores and averaged per participant. 207

Figure 7.6 Relationship between scores on the Mezurio Gallery Game recall accuracy scores and delta AIU scores from the Story Time task averaged per participant. 208

Figure 7.7 Plot showing Mean and 95% CI on average AIU recall scores for Gallery Game recall accuracy score quartiles. 209

Figure 7.8 Plot showing Mean and 95% CI on average delta AIU scores for Gallery Game recall accuracy score quartiles. 209

Figure 7.9 Plot showing mean and 95% CI on average AIU recall scores for Subjective Cognitive Decline score quartiles on daily and yearly questionnaires. 214

Figure 7.10 Plot showing mean and 95% CI on average delta AIU scores for Subjective Cognitive Decline score quartiles on daily and yearly questionnaires. 215

Figure A.1 Word similarity threshold result based on precision. 282

Figure A.2 Association of in-clinic test scores and age. 283

Figure A.3 Association of in-clinic test scores and years of education. 284

Figure A.4 Density plot showing the frequency distribution of various audio features of each PREVENT participant. 285

Figure A.5 Distribution of average AIU recall scores in each age group. 286

Figure A.6 Distribution of average delta AIU scores in each age group. 287

Figure A.7 Distribution of average AIU recall scores in each education group. 288

Figure A.8 Distribution of average delta AIU scores in each education group. 289

Figure A.9 The task schedule of the GameChanger study showing various cognitive tasks and self-report questionnaires as they appear in the 30 days schedule. 290

Figure A.10 Distribution of the number of participants fully completing (learning, immediate recall, and delayed recall) different number of stories in GameChanger. 291

Figure A.11 Correlation of the pipeline features extracted from the Story Time recall task. 292

Figure A.12 Density plot of various psycholinguistic features in the GameChanger story responses at recalls. 293

Figure A.13 Bar plot showing the number of participants in various age groups stratified by education in the GameChanger study. 294

List of abbreviations

ABCD	Arizona Battery for Communication Disorders of Dementia
ACE-R	Addenbrooke's Cognitive Examination - Revised
AD	Alzheimer's Disease
ADL	Activities of Daily Living
AI	Artificial Intelligence
AIU	Automatic Information Unit
aMCI	amnestic Mild Cognitive Impairments
aMCI/MD	amnestic MCI/Multiple Domain
aMCI/SD	amnestic MCI/Single Domain
ANOVA	Analysis of Variance
AOA	Age of Acquisition (subjective rating of words based on the age at which individuals typically acquire them for the first time)
APD	Average Pause Duration
API	Application Programming Interface
APOE	Apolipoprotein E
AR	Articulation Rate
ASAD-Cog	Alzheimer's Disease Assessment Scale-Cognitive Subscale
ASD	Average Syllable Duration
ASP	Audio Signal Processing
ASR	Automatic Speech Recognition
AWD	Average Word Duration
BBC	British Broadcasting Corporation
BERT	Bidirectional Encoders Representation from Transformers
BNC	British National Corpus
CAMCOG	Cambridge Cognitive Examination
CAMDEX	Cambridge Mental Disorders of the Elderly Examination
CDR	Clinical Dementia Rating
CERAD	The Consortium to Establish a Registry for Alzheimer's Disease.
CG	Control Group
CI	Confidence Intervals
CBD	Cortico-Basal Degeneration
COGNITO	Computerized Assessment of Information Processing
ConCERT-D	Consent to Consent for Research and Trials - Dementia
COV	Coefficient of Variation
CSF	Cerebrospinal Fluid
DAT	Dementia of Alzheimer's type
EDoN	Early Detection of Neurodegenerative diseases
EO	Early Onset
FTD	Frontotemporal Dementia
GDS	Global Deterioration Scale

HAMD	Hamilton Depression Scale
HC	Healthy controls
HD	Huntington Disease
HNR	Harmonics to Noise Ratio
IBM	International Business Machine
ICM	Information Concentration-Memory test
IMD	Index of Multiple Deprivation
IQR	Inter Quartile Range
IU	Information Units
JPEG	Joint Photographic Experts Group
LO	Late Onset
LSA	Latent Semantic Analysis
MADRS	Montgomery and Asberg Depression Rating Scale
MANOVA	Multivariate Analysis of Variance
MATTR	Moving Average Type Token Ratio
MCI	Mild Cognitive Impairments
MFCC	Mel-Frequency Cepstral Coefficients
MIU	Manual Information Unit
ML	Machine Learning
MLU	Mean Length of Utterance
MMSE	Mini-Mental State Examination
MoCA	Montreal Cognitive Assessment
MRI	Magnetic Resonance Imaging
MSE	Mean Square Error
naMCI	non amnestic Mild Cognitive Impairments
NART	National Adult Reading Test
NER	Named Entity Recognition
NLP	Natural Language Processing
NR	Not Reported
OA	Older Adults
OOV	Out-Of-Vocabulary
OR	Odds Ratio
pAD	probable AD
PCA	Principal Component Analysis
PCM	Pulse Code Modulation
PD	Parkinson's Disease
PET	Positron Emission Tomography
PPA	Primary Progressive Aphasia
QPA	Quantitative Production Analysis
RADAR-AD	Remote Assessment of Disease and Relapse - Alzheimer's Disease
RMS	Root Mean Square
RMT	Remote Monitoring Technologies
SCD	Subjective Cognitive Decline

SD	Standard Deviation
SDK	Software Development Kit
SR	Speaking Rate
SVM	Support Vector Machine
TTR	Type Token Ratio
VaD	Vascular Dementia
WAIS	Wechsler Adult Intelligence Scale
WAIS-R	Wechsler Adult Intelligence Scale Revised
WER	Word Error Rate [ratio of errors in a transcript (substitution, insertion, and deletion) to the total spoken words]
WMS	Wechsler Memory Scale
YA	Younger Adults

Declaration

I, Alankar Aryal Atreya, declare that this thesis entitled "Feasibility and utility of automated processing of speech-based cognitive assessments for early Alzheimer's Disease" is my own composition, and the research contained herein is my own original work under the supervision of Dr Chris Hinds, Professor Sarah Parish, and Associate Professor Marion Mafham. I conducted the literature review and planned the analysis. I produced all the tables and figures except where otherwise indicated. I performed all statistical analyses and wrote up all the aspects of the study presented in this thesis.

I confirm that I was not involved in designing the Mezurio Smartphone application or the Story Time task and in the data collection process. My supervisor Dr Chris Hinds provided the raw audio data presented in this thesis. No portion of this work has been submitted to support an application for any other degree.

Impact of the COVID-19 pandemic

According to the university guidance, this section provides the impact of the COVID-19 pandemic on work for this thesis. The thesis began in October 2018, and the effect of the COVID-19 pandemic in the UK started in March 2020. Following national and university guidance, I worked from home during that time.

Some of the impacts that I faced during COVID were:

i. I did not have access to the data delaying progress on core project (a large sized dataset was only made available in January 2022)
ii. working from home was less efficient than office-based access
iii. contact with my main supervisor and other key colleagues was substantially reduced over an extended period
iv. I was ill with COVID and tested positive 2 times. During that time my child was also not allowed to go to nursery (due to government policy), which reduced my work efficiency.

Chapter 1. Introduction

1.1 Early detection of Alzheimer's disease

1.1.1 The burden of Alzheimer's disease

The global burden of Alzheimer's disease (AD) is substantial as it is a leading cause of dementia, disability, and dependency among older adults. According to the World Health Organisation (WHO), an estimated 55 million people are living with dementia worldwide, and this number is expected to increase to 75 million by 2030 [1].

AD is a progressive, degenerative brain disorder characterised by the loss of cognitive function, including memory, language, planning and organisational skills (executive function), and problem-solving ability. People living with AD face increasing cognitive difficulties that interfere with their daily activities of living. The progression of AD can be divided into several stages, which reflect the increasing severity of the disease and the decline in cognitive and functional abilities. The stages of Alzheimer's disease [2] are:

a) Pre-clinical stage: This is the earliest stage of AD, in which changes occur in the brain [3], [4], but there are no detectable symptoms. Although the individual may notice changes, these are not detectable on neuropsychological tests.
b) Mild Cognitive Impairment (MCI): During this stage, the individual may experience more noticeable cognitive impairments, including memory loss, difficulty with language, and problems with spatial orientations. They may also have difficulty in executive functions such as planning and organising skills and may experience changes in mood and behaviour. Individuals with MCI are still able to independently complete most activities of daily living.

c) Mild and Severe AD: People with mild AD may have significant memory loss and difficulty communicating. They may have a decline in their ability to perform daily activities and may need assistance with basic self-care. Individuals with severe AD have severe memory loss and difficulty communicating. They may also lose the ability to walk and may be completely dependent on others for care [5].

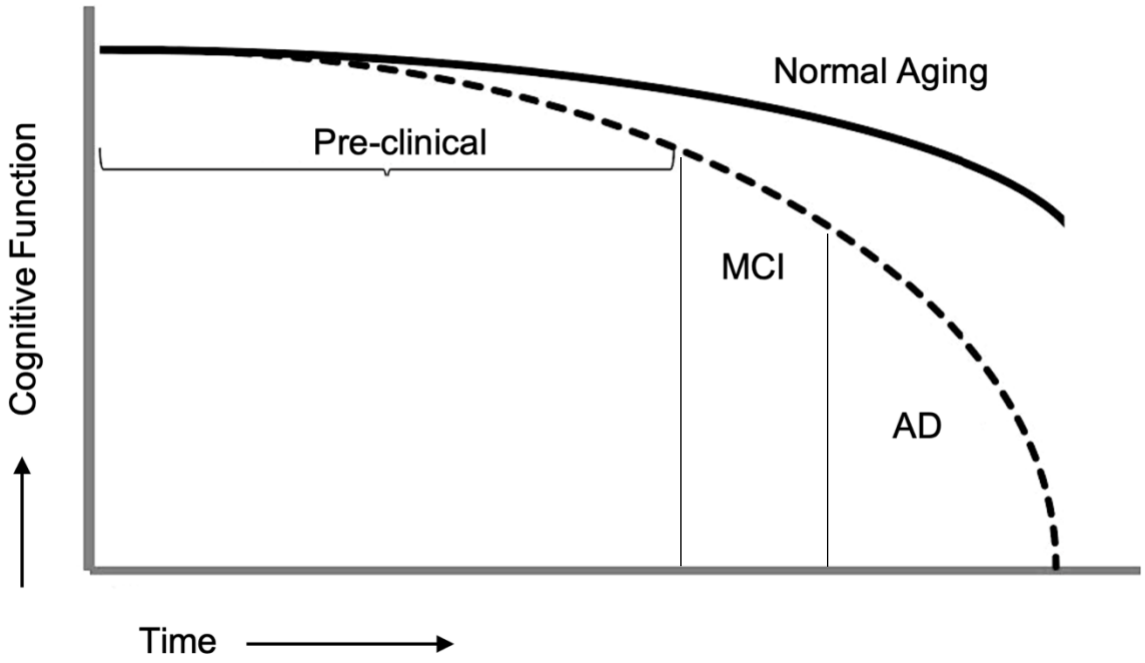


Figure 1.1 The continuum of Alzheimer's disease. The figure is adapted from [2].

The rate of AD progression is illustrated in **Figure 1.1**, highlighting that cognitive function declines rapidly from MCI and AD. However, some people may experience a rapid decline, while others may progress more slowly and these stages are only general guidelines.

There has been limited progress in developing and testing new AD treatments. Until recently, four drugs have been available: donepezil, galantamine, rivastigmine, and memantine [6]. These were developed to help the symptoms and have shown limited benefits [7]. The United

States Food and Drug Administration (FDA) recently provisionally approved Biogen's aducanumab, an antibody therapy for treating AD [8]. The approval was based on the long-term goal of the ability of the drug to remove amyloid plaques in the brain. This treatment still needs more clinical trials to prove its effectiveness and for full approval. Donanemab and Lecanemab, monoclonal antibodies, are the most recent addition to the therapy spectrum, which have been FDA approved under the Accelerated Approval Pathway [9], [10]. They have been tested in persons with early Alzheimer's disease, showing that these drugs reduced amyloid markers and resulted in less decline of cognition and function. However, these drugs were associated with adverse events. As such, no cure is currently available for this disease, and the expected life expectancy has been found to be less than ten years after the diagnosis [11].

The economic burden of this disease is also significant. Globally, the direct and indirect cost associated with caring for people with AD is estimated to be billions of dollars annually [2]. These costs include medical care, long-term care, and lost productivity due to caregivers taking time off work. Caring for a person with AD can be physically and emotionally demanding for family caregivers, who are reported to experience a high level of stress, anxiety, and depression [11], [12]. Therefore, there is an increasing need to identify and test potential treatments early in the disease to prevent irreversible loss of cognitive function.

1.1.2 Existing methods of AD detection

There are several methods for detecting AD, which can be broadly divided into:

a) Clinical assessment: These assessments involve thoroughly evaluating the individual's medical history, symptoms, and cognitive abilities. These may also include physical and neurological examinations and cognitive tests. The most commonly used cognitive

tests are the Mini-Mental State Examination (MMSE) [13], Cambridge Cognition Examination (CAMCOG) [14], Addenbrookes Cognitive Examination-Revised (ACE-R) [15], and Montreal Cognitive Assessment (MoCA) [16]. These tests can help to identify the presence of AD-related symptoms such as memory problems, confusion, disorientation, problems with speech and language, and difficulty with daily tasks. Other tests that are used to identify the disease severity are the Global Deterioration Scale (GDS) [17], the Clinical Dementia Rating (CDR) scale [18], and Alzheimer's Disease Assessment Scale-Cognitive Subscale (ASAD-Cog) [19].

b) Biomarker: Biomarkers are substances in the body that can be measured to indicate the presence or risk of a particular condition. In the case of AD, biomarkers can be found in blood, cerebrospinal fluid (CSF), or imaging tests such as positron emission tomography (PET) or magnetic resonance imaging (MRI) scans [20]. These biomarkers can help confirm the diagnosis of AD and track the progression of the disease.

c) Genetic testing: Certain genetic mutations have been linked to an increased risk of developing AD [21]. Therefore, the genetic testing method is used to identify whether an individual with a family history of the disease has these mutations.

It is important to note that no single test can definitively diagnose AD. Instead, a diagnosis is typically made based on a combination of these methods.

1.1.3 The purpose of early detection

The concept of 'pre-clinical AD' has evolved to refer to those cognitively normal individuals without any appearance of clinical symptoms but the presence of established biomarkers (Aß amyloid and CSF tau) and neurodegeneration [2], [3]. This stage of AD has become a major focus of research because it is thought that interventions during this stage may have an

enhanced potential for therapeutic success. Detection of AD in the pre-clinical stage can be beneficial for the following reasons:

a) Allowing for earlier interventions: Detecting AD in its early stage allows for earlier interventions, including lifestyle changes and medication, which may slow the progression of the disease.
b) Opportunities for clinical trials: Early detection may allow individuals to participate in clinical trials for new treatments or therapies that may not be available to those with more advanced stages of the disease because most therapies may be better at slowing the progression or improving AD symptoms rather than reversing existing disease. The literature has highlighted that clinical trials need to target and recruit participants at the earliest pre-clinical stages of the neurodegenerative disease for developing effective therapeutics to prevent or slow the disease progression [22], [23].
c) Planning for the future: Early detection can also help with making long-term care plans and addressing any financial and legal issues while they are still able to do so. Knowing the possibility of having AD early can help families and caregivers prepare for the changes that may occur as the disease progresses [24]. This can allow them to make necessary arrangements and provide the support needed for the individual with AD.

1.1.4 The potential of digital tools for early AD detection

The existing AD detection methods (section 1.1.2) have limitations, including being invasive, requiring face-to-face contact, and being insensitive at the pre-clinical stage. Obtaining data from neuroimaging technologies and CSF is costly and burdensome [20]. Biomarker status can be estimated through a blood test; however, it requires face-to-face contact. Also, biomarker status does not help to track the disease progression. Existing neuropsychological tests can

track progression at a late stage; however, they are insensitive at the pre-clinical stage [25], as they were designed to detect impairments where the symptoms have already manifested [26].

Smartphones enable frequent data collection from large populations remotely without in-clinic visits and a lower participant burden to capture specific behaviours that correlate with the disease. Also, increased digital literacy among older adults has removed the barrier to using digital tools for health assessment [27]. Therefore, digital tools provide cost-effective, non-invasive, and scalable measures for preclinical screening of AD, allowing an acute longitudinal change in cognitive function to be accurately assessed.

As shown in **Table 1.1**, digital tools promise benefits for AD research with some disadvantages. Novel digital outcomes require validation against traditional biomarkers of the disease. Also, there is a data privacy concern regarding the ethical and sustainable use of personal data [28].

Table 1.1 Advantages and disadvantages of cognitive assessments of AD using digital tools.

Advantages	**Disadvantages**
- Assessments are low cost. - Frequent and remote assessments with a lower burden to participants. - Avoid human biases in evaluating scores. - Can be deployed on a larger population providing better statistical power for studies and the possibility of including a more diverse population. - Scalability with automated assessments.	- Automated assessments can be affected by socio-demographic biases in data. - Most of the automated assessments lack validation studies. - Assessments do not automatically lead to treatment or intervention. - Data privacy concern, if the digital tool collects personal data (e.g., voice, demographics data such as age and location)

Monitoring intra-individual patterns of cognitive change could provide a sensitive measure of cognitive function capable of detecting declining function early in the disease and the response to treatment. Frequent measurements will give more stable assessments of the cognitive state, which may depend on various confounders such as stress levels, sleep deprivation and intoxications, among others [29], [30]. This could help to improve current shortcomings in risk management, early detection of AD and care if integrated within a healthcare context. Technological advancements have made it possible to deploy such a technology in a large population allowing frequent collection of a large amount of both clinical and non-clinical data.

Digital tools provide cheap, remote, and scalable assessment; however, the larger data obtained from these tools can only be effectively utilised if automated. Therefore, there is a need for both remote assessment technologies and automated processing of the data.

1.2 Speech and language in Alzheimer's disease

Speech is an act of human communication associated with multiple cognitive and psychological processes, including retrieval from memory and production and perception of syntax by reliance on the working memory [31], [32]. Spoken language consists of rich, naturally occurring elements, including disfluency (filled pauses, hesitations) and use of surrogate words/phrases (e.g., "I mean", "you know", "right", "well", "so", "ok") compared to reading or written language. Also, these features provide a critical “window” to the level of cognitive functioning from semantic storage and retrieval, working memory, and speech production to thought processing [33]–[35].

This rich speech phenomenon has attracted researchers to use spoken language for analysing cognitive changes. Subtle changes within speech characteristics, including language use, are

reported in individuals more than a decade before clinical diagnosis of Alzheimer's disease and evolve throughout time [36]–[40]. Therefore, there is a growing interest in analysing raw speech recordings to find associations between subtle speech changes and early cognitive decline compared to invasive and costly techniques presently used to detect neurodegenerative pathology, including MRI, PET brain scans, and CSF samples. Research outcomes on speech-based (audio and language) analysis show great promise for the utility of speech as a marker of subtle cognitive impairment, enabling more timely detection of AD, an important avenue for therapeutic development [41].

Cognitive assessments using speech recordings could be comfortably taken remotely, such as at a participant's residence, reducing in-clinic visits and other clinic-related constraints. Also, audio signal processing (ASP), natural language processing (NLP), automatic speech recognition (ASR), and machine learning/artificial intelligence (ML/AI) technologies have rapidly improved and unfolded many emerging solutions. This recent development presents a promising opportunity to capture a large chunk of raw speech data and develop computational techniques for scalable analysis [42], [43].

1.2.1 Progression of speech and language changes in AD

Studies have shown that speech and language impairment is most notably gradual and parallel to the disease progression [31], [44]. Grouping disease stages into MCI, mild-to-moderate AD, and severe AD, **Table 1.2** shows various speech and language impairments in these populations.

Table 1.2 Progression of speech and language changes in different stages of AD.

Key domains	MCI	Mild-to-moderate AD	Severe AD
Semantic content	Use of empty words	Reduced information units with empty words	Almost unintelligible
Lexical/Syntactic content	Repetitions of words	Difficulties with more repetitions and inappropriate pronoun use	Failure to communicate properly
Speech fluency	Mostly fluent speech	Increasingly disfluent with hesitations and pauses	Disfluent speech
Logical flow	Incoherent response	Increasingly incoherent	Unintelligible

MCI - Mild Cognitive Impairment

In early-stage MCI, lexical access (word-finding) and semantic knowledge become difficult, which manifests in a reduction in semantic content, causing vague, empty or surrogate words/phrases ("something", "that thing", "I suppose", "I cannot remember", "I do not know") indicating confusion, forgetfulness, doubts or concerns [45]–[47]. Adults with MCI may be talking more without providing content while maintaining grammar to compensate for cognitive decline [48], [49]. Also, their speech contains repetitions and a decrease in the use of metaphors [50]. Although their fluency remains intact, they showed difficulty in retrieving the names of people and places and substituting them with words such as things, this, that, stuff, and so on [42], [51]–[57].

In the speech produced by the mild-to-moderate AD group, the filled pauses such as 'um' and 'er' start to be very prominent [58]. They start using more words to explain a simple concept where fewer words would suffice, making their speech very disfluent and empty. Speech from this group of patients demonstrates incoherent narration, semantic impairment, word-finding difficulty highlighted by slower speech rate and more prolonged hesitation, increasingly high repetitions and limited executive functions associated with poor monitoring of the speech narration [50], [59]–[61].

Severe AD patient's speech is mostly disfluent with increased stuttering, hoarseness, and breathiness. Their speech becomes unintelligible on semantic, syntactic, and logical flow aspects [62], [63].

1.2.2 Speech elicitation methods in AD research

This section describes the speech elicitation methods used in AD research. Various methods have been used to collect speech from the participants (**Table 1.3**). Picture description tasks are most commonly used for speech elicitations methods. Participants are asked to elaborate in detail on what they have seen in the picture. The main objective of the picture description task is to measure the quality of information an individual can provide from visual stimulus. The Boston Cookie Theft [64] and the picnic scene from the Western Aphasia Battery [65] are the two most common pictures used in the picture description task.

Another commonly used in-clinic or over-the-phone test is the semantic or verbal fluency task [66], [67], a time-constraint task where the participants are asked to produce a list of nouns from a particular category. Both fluency and picture description tasks were initially designed for aphasia, given that this disease affects speech and language abilities.

Memory-based recall tests are included in cognitive batteries such as the Wechsler Adult Intelligence Scale (WAIS) [68], Wechsler Memory Scale (WMS) [69] and Arizona Battery for Communication Disorders of Dementia (ABCD) [70], where participants are asked to recall short passages of text from their memory. Also, the frog story by Mercer Mayer [71] and the well-known fairy tale Cinderella story are used as story recall tasks. Participants are asked to recall the story immediately and delayed (e.g., after 30-60 minutes).

Table 1.3 Speech elicitation methods used for speech-based analysis in AD.

Speech elicitation methods	Description of the method
Picture description task	Look at the picture and describe it as detailed as possible.
Semantic/verbal fluency	Produce a list of nouns from a certain category.
Story narrations	Participants are either asked to read text-based narration or picture-based narration and recall from memory.
Discourse production task	Recording the interview between the patient and the physician and analysing the recorded sample.
Read aloud texts	Participants are asked to read aloud a literature text.

Discourse production or conversational speech tasks are used to examine the participant's ability to communicate with others. These include specific questions from describing steps to do an activity, e.g., How to prepare a tea? to open-ended biographical questions. One of the popular discourse production tasks is where the participants are asked to generate information for the question - 'You are taking a trip to New York; please describe all the activities you are planning to carry out?' [72]. Recently, a task involving map description has been developed to generate conversational speech data [73].

Another standard method for generating speech data is the use of read-aloud tests. In these tests, participants are asked to read text narratives. The popular text extracts used in these reading tasks are the grandfather [74] and rainbow passage [75]. These tasks are less semantically demanding, and speech production is less naturalistic compared to picture description tasks. However, some speech signals in reading speech have shown variations between disease stages of AD [76].

1.2.3 Audio and language features from narrative speech

The narrative speech provides both audio and language features. While audio features are extracted after processing speech recordings, language features are extracted from textual data obtained after transcribing those recordings (**Figure 1.2**). This choice of data representation (voice or text) determines the nature of the features to be extracted.

1.2.3.1 Audio features

Audio features that are studied in AD research are mostly temporal or fluency features such as speech rate and pause rate and features extracted from transcripts such as filled pauses and repetitions.

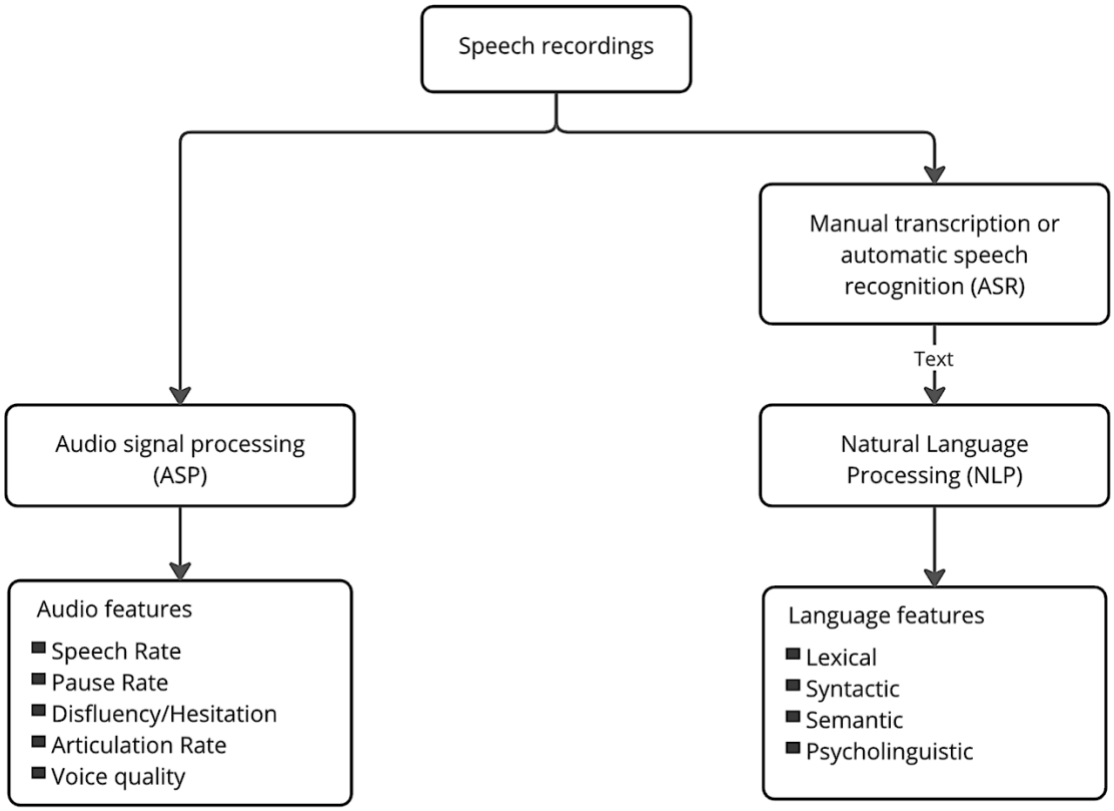


Figure 1.2 An overview of audio and language features obtained by processing speech recordings in Alzheimer's disease research. Audio signal and natural language processing technologies are used in extracting audio and language features for clinical decision-making.

1.2.3.2 Language features

The most commonly extracted language feature from the picture description task is information units (IU), also known as semantic units, idea units, and thematic elements. This feature is measured by comparing the descriptions provided by the individual to a pre-prepared list of information units that accurately convey relevant information on eliciting stimulus, for instance: objects, actions and scenes from the pictures [42], [52], [77]. Other language features studied in AD include lexical (word count and lexical richness) and syntactic features (use of different part-of-speech tags - nouns, verbs, adjectives, adverbs, and length of utterances) [53], [78].

More detailed information on both audio and language features is described in Chapter 2, and the literature related to the use of these features in AD research is reviewed in Chapter 3.

1.2.4 Relevance of audio and language on other forms of dementia

Similarly, there are other types of dementia, such as vascular dementia, frontotemporal dementia (FTD), Lewy body dementia (LBD), Primary Progressive Aphasia (PPA), and others that can lead to a decline in speaking and language comprehension abilities [79]. For example, language difficulties such as semantic impairment or understanding complex sentences can be prominent features of vascular dementia [80], and PPA and FTD patients may experience problems with word finding and fluent speech [81]. LBD patients may experience reduced speech rate, long pauses between utterances, and grammatical difficulty [82].

1.3 Mezurio Story Time task

Story Time is a voice-based story narration task deployed within the Mezurio[1] smartphone app, freely available on iOS and Android devices, which contains several novel cognitive tests and self-report questionnaires. Mezurio smartphone application containing various cognitive tasks, including Story Time task, has been deployed in various studies, including PREVENT [83], GameChanger [84], Deep and Frequent Phenotyping [85], RADAR-AD [86], and Early Detection of Neurodegenerative diseases [87]. In this doctoral research, speech data from PREVENT and GameChanger study cohorts recorded using the Story Time task, will be studied.

This task provides a test of speech, language, and memory; however, it does not provide any feedback about the performance to the app users. In this section, an overview of the background of Story Time, followed by a detailed summary of the task design, and finally, the selection of the stimuli used in the task are discussed.

1.3.1 Background of Mezurio Story Time task

The smartphone application is designed with the motivation of short but frequent assessment of cognitive function, aiming to accelerate early AD detection [88]. The device microphone collects raw speech, providing more naturalistic spoken language chunks than standardised in-clinic assessment and enabling the analysis of temporal and fluency-related speech features. This daily assessment task provides rich speech content compared to picture description tasks

[1] Mezurio (https://mezur.io/) is a smartphone-based clinical research tool developed by Digital Phenotyping Group, Big Data Institute, the University of Oxford, led by Dr Chris Hinds, containing several cognitive tasks aimed at accelerating early AD detection.

widely used for speech-based cognitive assessment. In addition, participants can complete the task in their own homes, providing the opportunity for repeated assessments (multiple narration tasks and immediate and 24-hr delayed recall within each task). The tasks themselves are a relatively low burden as short micro-interactions, enabling this tool to be deployed in a wide range of populations.

1.3.2 Task design

To start the task, participants must permit access to their device microphone. The voice check feature demonstrates to the user whether their microphone is working and how loud they are speaking with the help of the audio wave visualisation. The noise level that can be expected can vary depending on the environment in which the test is conducted. In a home setting, the background noise level can be influenced by factors such as household activities, ambient sounds, or the presence of other individuals. The Story Time task contains a background noise check feature that indicates the background noise level to ensure the participant is in a quiet area where the background noise level is really low before taking the test. There are three subtasks to each Story Time task: learning, immediate recall and 24-hour delayed recall.

The user interface representation of the Story Time learning subtask within the Mezurio smartphone app is shown in **Figure 1.3**. In a learning subtask, participants are presented with comic strips and are asked to record their speech samples narrating a story describing what they see in those comics. The first scene shown to the participants contains the text, "Once upon a time, there was a boy named Corky ..." to start and frame the story narration. Then, the participants can move through the comic frames at their own pace by clicking the "Next" button, and they have to say aloud what is going on in that picture. Participants will also be

indicated if they have talked enough for each frame by displaying a progress bar that turns orange when the suggested number of words is reached for that frame.



Figure 1.3 Representation of a Story Time learning activity; each activity contains nine isolated comic strip frames, which participants are asked to narrate aloud frame-by-frame.

The immediate recall follows the learning subtask, where participants are asked to recall aloud the story they just learned without the picture stimuli. Similarly, in the delayed recall subtask presented on the next day (after 24 hours), participants are again asked to recall the same story they learned the day before aloud from their memory. Overall, the task allows simultaneous analysis of audio and language features while also potentially accessing episodic memory.

1.3.3 Task stimuli

The stimuli for Story Time include eighteen comic stories deployed as a series of adventures involving the main character Corky. The task designer[2] acquired these comics from a freely

[2] Dr Chris Hinds designed the Story Time task and provided information on the selection of stimuli.

available online comic platform[3]. There is an equal number of nine frames in each comic story. The designer considered two metrics for selecting comics that were a) the number of words included in each comic and b) the JPEG compression size. The idea behind using JPEG compression size was that non-complex pictures could be compressed to a greater degree than complex pictures given the exact picture dimensions, resulting in larger file sizes in the more complex pictures [89], [90]. The above metrics helped to ensure consistency between verbal context and visual complexity of cartoons presented during the baseline assessment, relevant to those shown at subsequent follow-ups.

1.4 Potential roles of speech-based assessment methods

As mentioned in section 1.1.2, existing clinic-based cognitive tests are insensitive to early AD detection. Biomarker status estimated through neuro-imaging technologies and CSF is invasive and costly. Section 1.1.4 discussed that digital assessment methods are particularly advantageous being non-invasive, low cost and can be easily administered remotely or in a person's environment, such as at home which is convenient for the patient. Therefore, these assessments can increase accessibility and reduce the burden of in-clinic visits.

Speech has been used to investigate cognitive changes as it enables remote and efficient memory assessment (recall as a modality). Speech may add additional information from audio and language features (such as speech rate, pause rate, and semantic content) over and above a memory-based test providing valuable longitudinal data on cognitive function and helping researchers understand the impact of treatments over time.

[3] https://comicbookplus.com/?dlid=27126

Speech-based assessments are typically designed to complement existing cognitive tests rather than replace them. They could be used in conduction with other digital measures. These assessment methods could add value to the diagnostic process and be used to screen populations to identify people with early-stage AD for clinical trials. During clinical trials, speech-based assessments might be used as an additional or supplementary outcome measure to assess the efficacy of treatments. Changes in speech patterns or language content could be monitored over time to track the progression of the disease or the impact of an intervention.

1.5 Aims and objectives of the thesis

Firstly, this thesis identifies and assimilates previous studies on speech-based cognitive assessment in AD. Subsequently, it aims to develop scalable computational methods and establish their feasibility and utility on the raw audio data collected from the smartphone app. Finally, it aims to investigate associations between speech-based test scores and other neuropsychological scores and whether there is a relationship between speech-based outcomes and variables associated with cognitive decline, including age, education, subjective cognitive complaints, deprivation, and family history of dementia.

The main objectives of this thesis are to:

1. Develop feature extraction pipelines to process raw audio data to produce meaningful speech metrics relating to memory and features within the speech narration. Compare automated and manual scorings of language-based features to evaluate the accuracy of automation.
2. Examine the general acceptability and feasibility of remote, digital speech-based cognitive assessment. This includes calculating compliance on task scheduling and

completion and studying whether these metrics are affected by socio-demographic variables across a large population.

3. Estimate associations of the speech-based cognitive test score against standardised or well-established neuropsychological test scores to establish validity.
4. Explore the utility of speech-based cognitive assessment by studying the distribution of speech-based measures in different socio-demographic groups and investigating if there is a relationship between Story Time outcomes and well-established socio-demographic risk factors for AD.

While this thesis focuses on audio and language assessments in AD, as discussed in section 1.2.4, the results from this thesis will have implications for other forms of dementia.

1.6 Thesis outline

The organisation of this thesis is as follows.

Chapter 2 briefly describes audio and language features used in AD research that are extracted from speech recordings.

Chapter 3 covers a systematic literature review of speech-based cognitive assessments on AD. This chapter first describes the search strategy and exclusion criteria in the electronic records for literature review. Then the studies on cognitive processing using audio signals and natural language processing are discussed in detail. The last section in Chapter 3 focuses on the limitations of the current studies.

Chapter 4 describes the design and cohort characteristics of the two datasets used in this thesis: the PREVENT and GameChanger studies.

Chapter 5 presents the feature extraction pipeline developed to derive speech-based features from raw speech data using automatic speech recognition, natural language processing, and audio signal processing techniques.

Chapters 6 and **Chapter 7** evaluate the utility and feasibility of newly developed methods on two datasets with speech recordings. In Chapter 6, the computational methods are piloted in the small-sized PREVENT dataset. In Chapter 7, the evaluation of the method is extended to GameChanger, a population-based study where the feasibility and utility of the speech-based cognitive assessment are analysed. The associations of Story Time features with well-established socio-economic variables known to be associated with dementia risk and scores from two other tasks (Gallery Game, Subjective Cognitive Decline questionnaire) within the Mezurio smartphone application are evaluated.

Finally, **Chapter 8** gives a summary of the thesis along with suggesting directions for future research.

Chapter 2. Audio and language features in AD

2.1 Introduction

Chapter 1 briefly discussed the audio and language features studied in AD research. This chapter aims to provide more detailed information on how these features are extracted from speech recordings.

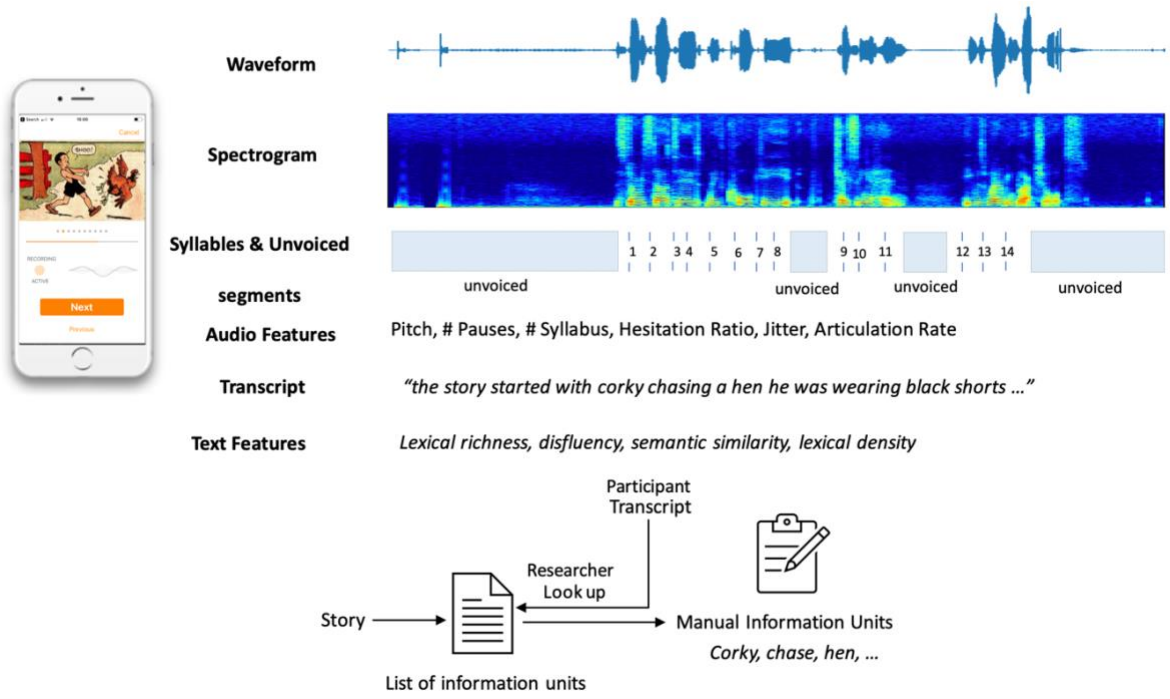


Figure 2.1 Representation of a test speech recording from the Story Time app and overview of audio features extracted from these signals. Transcripts from the speech can be further processed to extract language features.

An example of a speech recording obtained by completing one of the stories within the Story Time task, its representation in waveform and spectrogram, the estimation of unvoiced segments and syllables, and the extraction of information units (IU) manually from the transcribed text is shown in **Figure 2.1**. This shows the potential of speech recordings in providing various audio and language features (**Figure 2.2**) for evaluating expressive audio

and language performance. Throughout the thesis, audio features will represent all acoustic, paralinguistic, prosodic features as they are used interchangeably in the literature.

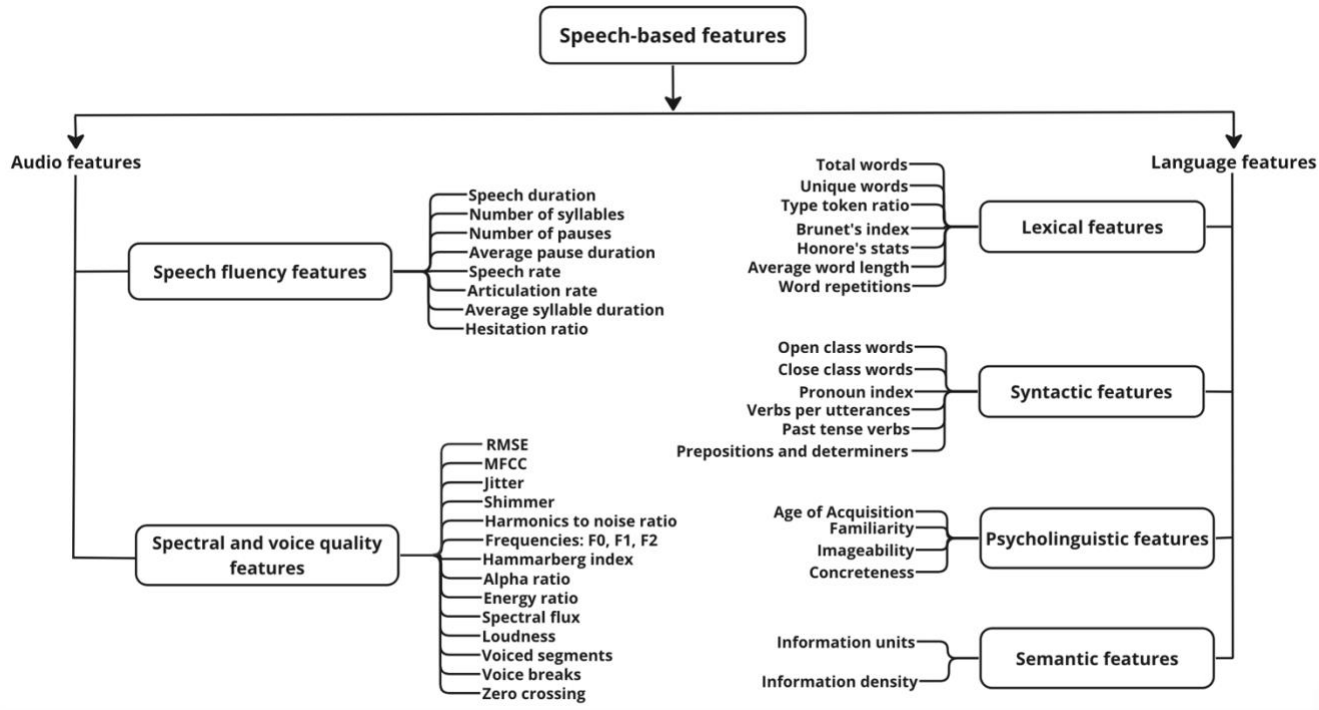


Figure 2.2 Taxonomy of audio and language features extracted from speech recordings.

2.2 Audio features

The influence of speech characteristics of the patients is known to provide information on their cognitive ability [51]. The deterioration of speech fluency with distinct, more prolonged hesitation, a lower speech rate, and word-finding difficulty has been found to change significantly between healthy controls and participants in different AD stages [47], [91], [92]. These changes are also further investigated using spectral [55], and voice quality [93] features and the method to calculate those will be discussed in the following sub-section.

2.2.1 Speech fluency features

The commonly extracted speech fluency features include recording duration, duration of voiced segments, silence duration, and many others [94]–[96]. Other features such as speech

and articulation rates are calculated in syllables, phonemes, or word levels. In this thesis, syllable level speech rate and articulation rate will be calculated. The formula for calculating speech and articulation rates is as follows:

$$Speech\ rate\ = \frac{\#\ syllables}{total\ recording\ duration} \qquad Eq.\ 2.1$$

$$Articulation\ rate\ = \frac{\#\ syllables}{total\ speech\ duration} \qquad Eq.\ 2.2$$

Similarly, average syllable duration is calculated as the ratio between the total speech duration and the number of syllables, as shown in Eq. 2.3.

$$Average\ syllable\ duration\ = \frac{total\ speech\ duration}{\#\ syllables} \qquad Eq.\ 2.3$$

Another common speech fluency feature is the pause duration. In the existing literature, the duration of silence to define a pause varies across studies [54]. However, most authors reported silence greater than 200 milliseconds as pauses in speech [67], [94], [97].

The hesitation ratio is calculated as the total duration of pause to the total audio duration:

$$Hesitation\ ratio\ = \frac{total\ pause\ duration}{total\ recording\ duration} \qquad Eq.\ 2.4$$

where the value of 0 means there are no pauses, and 1 means there is no speech in the audio recording.

2.2.2 Spectral and voice quality features

The features related to frequency, such as pitch (F0), formants (F1, F2, F3), jitter, shimmer, loudness, and harmonics-to-noise ratio, are extracted from speech. F0, also known as fundamental frequency, measures the rate at which the vocal fold vibrates when speech is

produced [98]. Formants provide information on the resonant frequencies of the vocal tract [99]. The loudness of speech is calculated based on the sound intensity [100].

Jitter is calculated as a variation of the fundamental frequency (F0) [101]:

$$Jitter = \frac{1}{N-1}\sum_{i}^{N-1} |T_i - T_{i-1}| \qquad Eq.2.5$$

where N is the number of fundamental frequency (F0) periods, and Ti represent the period length of fundamental frequency (F0).

Shimmer provides the measure of the variability of amplitude from peak to peak [102] and is calculated as:

$$Shimmer = \frac{1}{N-1}\sum_{i}^{N-1} |20\ log\ (\frac{A_{i+1}}{A_i})\ | \qquad Eq.2.6$$

where A_i represents the amplitude of the signal, and N is the number of fundamental frequency (F0) periods.

Harmonics to noise ratio (HNR) calculates the logarithm of the ratio of energy signals that are associated with harmonics (periodic - integer multiple of the fundamental frequency) and noise (non-periodic) components of the speech [102], and this measure is calculated as:

$$HNR = 10\ log_{10}\frac{Harmonic\ component\ of\ speech}{Noise\ component\ of\ speech} \qquad Eq.2.7$$

Other features extracted from speech recordings include alpha ratio, Hammarberg index, spectral slope, and Mel-Frequency Cepstral Coefficients (MFCC) features. The alpha ratio is calculated as a ratio of the summed energy from 50-1kHz and 1-5kHz [103]. Hammarberg index is calculated as the difference between the maximum energy in the energy band lower

than 2kHz and the 2 to 5kHz band [104]. The spectral slope is calculated as a linear regression slope of the logarithmic power spectrum within the two given bands [55]. Mel-frequency cepstral coefficients (MFCC) estimate the spectral envelope of the speech [105]. These features are used in AD research to investigate cognitive impairment as they are known to provide information on speech irregularities, including quiet or muffled and breathy or hoarse speech [42], [105], [106]. There are several other audio features, including voice breaks [102], zero-crossing rate [107], and peak slope [108].

The reliability and standard of these spectral and voice quality features depend on various factors, such as the length of the time window used in feature extraction (openSMILE library uses a sliding window of 120 milliseconds [109]) and the method used for spectral decomposition (e.g., Fast Fourier Transform (FFT) or Wavelet Transform) [110]. Due to the above factors, the values of these features may differ across the software used for feature extraction.

2.3 Language Features

Text samples are known to provide information on linguistic abilities that are related to cognition [37], [40]. The following sub-sections show how features related to cognition are extracted from text, assuming that an accurate transcript from the audio recording has been obtained.

2.3.1 Lexical and syntactic features

Lexical features include counting the number of total words and unique words in the speech response.

The lexical or vocabulary richness is measured using the type-token ratio (Eq. 2.8), Brunet's Index (Eq. 2.9) and Honore's statistics (Eq. 2.10). As the type-token ratio depends on the text length, the moving average type-token ratio is recommended in the text analysis [42].

$$\text{Type Token Ratio (TTR)} = \frac{U}{N} \qquad Eq.2.8$$

$$\text{Brunet Index (BI)} = N^{U^{-0.165}} \qquad Eq.2.9$$

$$\text{Honore's Stats (HS)} = \frac{100\ log\ N}{1 - \frac{N_1}{U}} \qquad Eq.2.10$$

where,

N = Total number of words

U = Total number of unique words (after removing repeats)

N_1 = Total number of words that occur only once in the entire recording (hapax legomena)

Honore's stats measure lexical richness by counting words that are used only once in speech utterances or text, known as the hapax legomena [23]. The higher the value of Honore's stats richer is the vocabulary. The values obtained from Brunet's Index generally vary between 10 and 20; the lower the value richer is the speech.

Other features include the use of word-class, verbs, including past tense verbs, proportions of prepositions and determiners to the total words in the speech sample. Open-class words and close-class words are primarily lexical measures. The proportion of word-class can also have syntactic implications with adjectives modifying nouns and adverbs modifying verbs and adjectives [111], [112], they are grouped into syntactic features in this thesis, with the evidence from the study by Mueller et al. (2018), where these features were grouped into syntactic factors using the latent factor analysis [49]. Studies report measuring syntactic complexity by

calculating the mean length of utterance (MLU), i.e., the average number of words per utterance (a sentence or collection of words terminated by a pause) and its variants as phrase length, mean clause length (also called c-unit length) and number of words per clause (c-unit) [113], [114].

Disfluency measures how fluent the spoken language is, reflecting lexical decisions and discourse planning [67]. Literature shows that there are various measures calculated under disfluency. Most of the studies count the number of filled pauses (including 'uh', and 'um') [108], word repetitions [115], and occurrence of false starts (e.g.:, "Flow... I mean Flower") [116], from speech.

2.3.2 Semantic features

IUs are accurate and relevant information about the content of the picture. Traditionally, the semantic content of the connected speech samples was quantified using a pre-prepared list of IUs (as shown in **Figure 2.1**) [28], [29], where the occurrence of the IUs are counted based on the pre-defined list prepared using the method described by Croisile et al. (1996) [47].

With NLP advancement, semantic content is measured by embedding the tokens (words) into the high-dimensional vector space [117]. Semantic similarity is then calculated with the distance between the word vectors. The earliest method developed on word embedding was Latent Sentiment Analysis (LSA). This method represents the text sample as vectors in a high-dimensional semantic space [118], [119]. Then the similarity between the two vectors is calculated using the formula below:

$$Cos(v_1, v_2) = \frac{v_1 . v_2}{||v_1|| ||v_2||} \qquad Eq.\ 2.11$$

where v_1, v_2 are two-word vectors, the higher the cosine similarity, they are semantically similar.

In recent years, methods such as Word2Vec [120] and GloVe [121] were developed by training the model in a large corpus of texts providing reliable word vectors compared to the ones obtained from the LSA method [122]. One commonly used example demonstrating how good these methods are at capturing the semantic relationship between words is the vectors used to represent king, queen, man, and woman, showing the analogy of king as queen as man is to woman.

$$king - queen = man - woman$$

Recently, a massively trained language model named BERT was developed by Google that captures the meaning and context of words in a sentence and provides high-quality context-aware word embeddings [123]. Although these models were developed for language translation and question answering, there is an increasing interest in using those models in AD research to classify the AD speech samples from healthy controls [124], [125].

2.3.3 Lexical density

Lexical density is another measure used to estimate speech or written text's informativeness [126]. This term is referred to in the literature as idea density, information density, content density or propositional density. To extract this measure first, the sample is divided into tokens (words) and labelled using part-of-speech (POS) tags to calculate this measure. The information units, such as nouns, verbs, adjectives and adverbs, are extracted from the POS tag labels and divided by the number of total words in the sample (Eq. 2.12).

$$Lexical\ density\ = \frac{\#\ Information\ Units}{N} \qquad Eq.\ 2.12$$

The proportions of nouns, verbs, adjectives, and adverbs are also estimated from the POS tags.

2.3.4 Psycholinguistic features

An increased reliance/dependency on highly familiar words has been found to indicate semantic impairment [127], [128]. Subjective ratings of words based on the age of acquisition (AOA), familiarity, imaginability and concreteness are available from Bristol and Gilhooly-Logie norms [129]. Following are the details on these features:

AOA: AOA refers to the age at which a word was learnt and has been proposed as a significant contributor to language and memory processes. People with AD are found to produce words acquired earlier in life, as well as items that are more frequent, more typical, and shorter in length compared to healthy controls [128]. Among AOA ratings, the word *'momma'* (the first word a child says typically) is rated 1.58, and *'eisteddfod'* (a word derived from Welsch which means session) is rated 25.

Familiarity: Familiarity is interpreted as a measure of the frequency of exposure to a word. Individuals with genetic markers of AD (i.e., apolipoprotein E) have been found to produce words of higher frequency than individuals without those genetic markers [127]. For example, the word *'vertex'* is rated low in familiarity, while *'zebra'* is high in familiarity.

Imageability: Imageability is a semantic variable that measures how easy it is for a word to arouse mental images. It has been used to evaluate the effect of meaning on memory and word

recognition. For example, the word *'hope'* is rated low in imageability, while *'house'* in high imageability.

Concreteness: Concreteness indicates the degree to which a concept refers to a perceptible entity. For example, 'couch' is a concrete word, while 'ideal' is an abstract word.

2.4 Summary

This chapter provided an overview of audio and language features and their extraction methods from speech and text samples. The next chapter, Chapter 3, begins by reviewing literature reporting audio and language analysis on spontaneous speech used as the data collection in various stages of AD.

Chapter 3. Literature review on audio and language processing for cognitive assessment

3.1 Literature review

3.1.1 Objectives

The general introduction of AD, with progressive impairment of cognitive and functional abilities, burdening patients and carers and having a severe economic and societal impact, was described in Chapter 1. In Chapter 2, the audio and language features studied in AD research were briefly described. This chapter presents a literature review seeking to summarise the current evidence regarding cognitive assessment using speech recordings and applying speech and natural language processing to study AD. First, the literature review process is described, followed by the details on the speech-based assessments, use of computational methods, and informative audio and language features in the AD research. The main aims of this review are:

a) to summarise the existing findings.
b) to provide insight into data collection methods, audio and language features, and the technology used.
c) to highlight gaps in current research

3.1.2 Search strategy

This literature review was conducted using the PRISMA guideline (see **Table A.1** for the checklist). This search was performed on Medline (Ovid 1946 to 2020 Week 40), Embase (1974 to 2020 Week 40), PsycINFO (1806 to 2020 Week 40), and Web of Science (Inception to 2020 Week 40. All the articles using the audio and language processing methods for AD

research specifying "mild cognitive impairment", "subjective cognitive impairment", and "pre-clinical AD" or other related terminologies commonly used in this research area were searched. No limits were applied for language or publication date for the literature search. The full search terms are shown in **Table 3.1**.

Table 3.1 Search terms used in different databases to identify studies reporting speech-based cognitive assessment on Alzheimer's Disease.

Medline (Ovid, 1946 to 2020 Week 40)	
1	Alzheimer Disease/ or Alzheimer*.mp.
2	subjective cognitive impairment.mp. or subjective cognitive decline.mp.
3	mild cognitive impairment.mp. or MCI.mp.
4	pre*clinical AD.mp. or pre*clinical Alzheimer* disease.mp. or prodromal AD.mp. or prodromal Alzheimer* disease.mp.
5	1 or 2 or 3 or 4
6	spontaneous speech.mp. or connected speech.mp. or pathological speech.mp. or connected language.mp. or spontaneous language.mp. or spoken language.mp. or oral discourse.mp. or discourse.mp.
7	speech degradation.mp. or speech deterioration.mp. or speech disorder*.mp. or Speech Disorder/ or speech impairment*.mp. or Speech-Language Pathology/ or Speech*language pathology.mp.
8	Linguistics/ or linguistics.mp. or linguistic features*.mp. or semantic feature*.mp. or linguistic complexity.mp.
9	audio feature*.mp. or temporal feature*.mp. or acoustic feature*.mp. or prosodic feature*.mp. or prosody.mp. or Phonetics/ or phonetics.mp. or phonetic feature*.mp.
10	speech analysis.mp. or discourse analysis.mp. or audio signal processing.mp. or audio processing.mp. or acoustic measure*.mp. or acoustic analysis.mp. or automatic speech recognition.mp. or ASR.mp. or Automated Speech Recognition.mp. or Quantitative Production Analysis.mp. or Narrative Language Analysis.mp. or speech segmentation.mp.
11	Narration/ or narration.mp. or narrative*.mp. or story*telling.mp. or narrative discourse.mp. or picture description.mp.
12	lexical semantics.mp. or lexical production.mp. or lexical coherence.mp.
13	Natural Language Processing/ or Natural Language processing.mp. or NLP.mp. or text processing.mp.
14	filled pause*.mp. or hesitation ratio.mp. or silent pause*.mp. or d#sfluenc*.mp. or empty speech.mp. or speech pause*.mp. or filler word*.mp.
15	Idea density.mp. or information unit*.mp. or idea unit*.mp. or semantic unit*.mp.
16	Semantic content.mp. or syntactic content.mp. or semantic coherence.mp.
17	Part of speech tag*.mp. or POS tag*.mp.
18	multi*modal data.mp.
19	Vector space model*.mp. or word embedding*.mp. or word vector*.mp. or word2vec.mp. or Latent Semantic Analysis.mp. or LSA.mp.
20	speech rate.mp. or jitter.mp. or shimmer.mp. or articulation rate.mp. or pause duration.mp. or harmonics to noise ratio.mp. or HNR.mp. or average syllable duration.mp. or Voice Quality/ or voice quality.mp. or type token ratio.mp. or number of pauses.mp.
21	6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20
22	5 and 21
Embase (1974 to 2020 Week 40)	
1	Alzheimer disease/ or Alzheimer*.mp.
2	subjective cognitive impairment.mp. or subjective cognitive decline.mp.
3	mild cognitive impairment/ or mild cognitive impairment.mp. or MCI.mp.
4	pre*clinical AD.mp. or pre*clinical Alzheimer* disease.mp. or prodromal AD.mp. or prodromal Alzheimer* disease.mp.
5	1 or 2 or 3 or 4
6	spontaneous speech.mp. or connected speech.mp. or pathological speech.mp. or connected language.mp. or spontaneous language.mp. or spoken language.mp. or oral discourse.mp. or discourse.mp.
7	speech degradation.mp. or speech deterioration.mp. or speech disorder*.mp. or speech disorder/ or speech impairment*.mp. or Speech*language pathology.mp.

8	linguistics/ or linguistics.mp. or linguistic feature*.mp. or semantic feature*.mp. or linguistic complexity.mp.
9	audio feature*.mp. or temporal feature*.mp. or acoustic feature*.mp. or prosodic feature*.mp. or prosody.mp. or phonetics/ or phonetics.mp. or phonetic feature*.mp.
10	speech analysis/ or Speech analysis.mp. or discourse analysis/ or discourse analysis.mp. or audio signal processing.mp. or audio processing.mp. or acoustic measures.mp. or acoustic analysis/ or acoustic analysis.mp. or automatic speech recognition/ or automatic speech recognition.mp. or ASR.mp. or automated speech recognition.mp. or Quantitative Production Analysis.mp. or Narrative Language Analysis.mp. or speech segmentation.mp.
11	narration.mp. or narrative*.mp. or storytelling/ or story*telling.mp. or narrative discourse.mp. or picture description.mp.
12	Lexical semantics.mp. or lexical production.mp. or lexical coherence.mp.
13	natural language processing/ or Natural Language processing.mp. or NLP.mp. or text processing.mp.
14	filled pause*.mp. or hesitation ratio.mp. or silent pause*.mp. or speech pause*.mp. or d#sfluenc*.mp. or empty speech.mp. or filler word*.mp.
15	Idea density.mp. or information unit*.mp. or idea unit*.mp. or semantic unit*.mp.
16	Semantic content.mp. or syntactic content.mp. or semantic coherence.mp.
17	Part of speech tag*.mp. or POS tag*.mp.
18	multi*modal data.mp.
19	vector space model*.mp. or word embedding*.mp. or word vector*.mp. or word2vec.mp. or Latent Semantic Analysis.mp. or LSA.mp.
20	speech rate/ or speech rate.mp. or jitter.mp. or shimmer.mp. or articulation rate.mp. or pause duration.mp. or harmonics to noise ratio.mp. or HNR.mp. or average syllable duration.mp. or voice quality.mp. or type token ratio.mp. or number of pauses.mp.
21	6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20
22	5 and 21
PsycINFO (1806 to 2020 Week 40)	
1	Alzheimer's Disease/ or Alzheimer*.mp.
2	subjective cognitive impairment.mp. or subjective cognitive decline.mp.
3	mild cognitive impairment.mp. or MCI.mp.
4	pre*clinical AD.mp. or pre*clinical Alzheimer* disease.mp. or prodromal AD.mp. or prodromal Alzheimer* disease.mp.
5	1 or 2 or 3 or 4
6	spontaneous speech.mp. or connected speech.mp. or pathological speech.mp. or connected language.mp. or spontaneous language.mp. or spoken language.mp. or oral discourse.mp. or discourse.mp.
7	speech degradation.mp. or speech deterioration.mp. or speech disorder*.mp. or Speech Disorders/ or speech impairment*.mp. or Speech Language Pathology/ or Speech*language pathology.mp.
8	Linguistics/ or linguistics.mp. or Cognitive Linguistics/ or linguistic feature*.mp. or semantic feature*.mp. or linguistic complexity.mp.
9	audio feature*.mp. or temporal feature*.mp. or acoustic feature*.mp. or prosodic feature*.mp. or Prosody/ or prosody.mp. or Phonetics/ or phonetics.mp. or phonetic feature*.mp.
10	speech analysis.mp. or Discourse Analysis/ or discourse analysis.mp. or audio signal processing.mp. or audio processing.mp. or acoustic measures.mp. or acoustic analysis.mp. or Automated Speech Recognition/ or Automated Speech Recognition.mp. or automatic speech recognition.mp. or ASR.mp. or Quantitative Production Analysis.mp. or Narrative Language Analysis.mp. or speech segmentation.mp.
11	Narratives/ or narrative*.mp. or narration.mp. or Storytelling/ or story*telling.mp. or narrative discourse.mp. or picture description.mp.
12	Lexical semantics.mp. or lexical production.mp. or lexical coherence.mp.
13	Natural Language Processing/ or Natural Language processing.mp. or NLP.mp. or text processing.mp.
14	filled pause*.mp. or hesitation ratio.mp. or silent pause*.mp. or Speech Pauses/ or speech pause*.mp. or d#sfluenc*.mp. or empty speech.mp. or filler word*.mp.
15	Idea density.mp. or information unit*.mp. or idea unit*.mp. or semantic unit*.mp.
16	semantic content.mp. or syntactic content.mp. or semantic coherence.mp.
17	part of speech tag*.mp. or POS tag*.mp.
18	multi*modal data.mp.
19	vector space model*.mp. or word embedding*.mp. or word vector*.mp. or word2vec.mp. or Latent Semantic Analysis.mp. or LSA.mp.

20	Speech Rate/ or speech rate.mp. or jitter.mp. or shimmer.mp. or articulation rate.mp. or pause duration.mp. or harmonics to noise ratio.mp. or HNR.mp. or average syllable duration.mp. or voice quality.mp. or type token ratio.mp. or number of pauses.mp.
21	6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20
22	5 and 21
Web of Science (Inception to 2020 Week 40)	
#1	TS=("Alzheimer Disease" OR Alzheimer*)
#2	TS=("subjective cognitive impairment" OR "subjective cognitive decline")
#3	TS=("mild cognitive impairment" OR MCI)
#4	TS=("pre*clinical AD" OR "pre*clinical Alzheimer* disease" OR "prodromal AD" OR "prodromal Alzheimer* disease")
#5	#4 OR #3 OR #2 OR #1
#6	TS=("spontaneous speech" OR "connected speech" OR "pathological speech" OR "connected language" OR "spontaneous language" OR "spoken language" OR "oral discourse" OR discourse)
#7	TS = ("speech degradation" OR "speech deterioration" OR "speech disorder*" OR "speech impairment*" OR "Speech*language pathology")
#8	TS=(linguistics OR "linguistic feature" OR "semantic feature" OR "linguistic complexity")
#9	TS=("audio feature*" OR "temporal feature*" OR "acoustic feature*" OR "prosodic feature*" OR prosody OR phonetics OR "phonetic feature*")
#10	TS=("speech analysis" OR "discourse analysis" OR "audio signal processing" OR "audio processing" OR "acoustic measures" OR "acoustic analysis" OR "Automated Speech Recognition" OR "automatic speech recognition" OR "ASR" OR "Quantitative Production Analysis" OR "Narrative Language Analysis" OR "speech segmentation")
#11	TS=("speech analysis" OR "discourse analysis" OR "audio signal processing" OR "audio processing" OR "acoustic measures" OR "acoustic analysis" OR "Automated Speech Recognition" OR "automatic speech recognition" OR "ASR" OR "Quantitative Production Analysis" OR "Narrative Language Analysis" OR "speech segmentation")
#12	TS=("lexical semantics" OR "lexical production" OR "lexical coherence")
#13	TS=("Natural Language Processing" OR NLP OR "text processing")
#14	TS=("filled pause*" OR "hesitation ratio" OR "silent pause*" OR "speech pause*" OR d$sfluenc* OR "empty speech" OR "filler word*")
#15	TS=("Idea density" OR "information unit*" OR "idea unit*" OR "semantic unit*")
#16	TS=("semantic content" OR "syntactic content" OR "semantic coherence")
#17	TS=("part of speech tag*" OR "POS tag*")
#18	TS=("multi*modal data")
#19	TS=("vector space model*" OR "word embedding*" OR "word vector*" OR word2vec OR "Latent Semantic Analysis" OR LSA)
#20	TS=("speech rate" OR jitter OR shimmer OR "articulation rate" OR "pause duration" OR "harmonics to noise ratio" OR HNR OR "average syllable duration" OR "voice quality" OR "type token ratio" OR "number of pauses")
#21	#20 OR #19 OR #18 OR #17 OR #16 OR #15 OR #14 OR #13 OR #12 OR #11 OR #10 OR #9 OR #8 OR #7 OR #6
#22	#21 AND #5

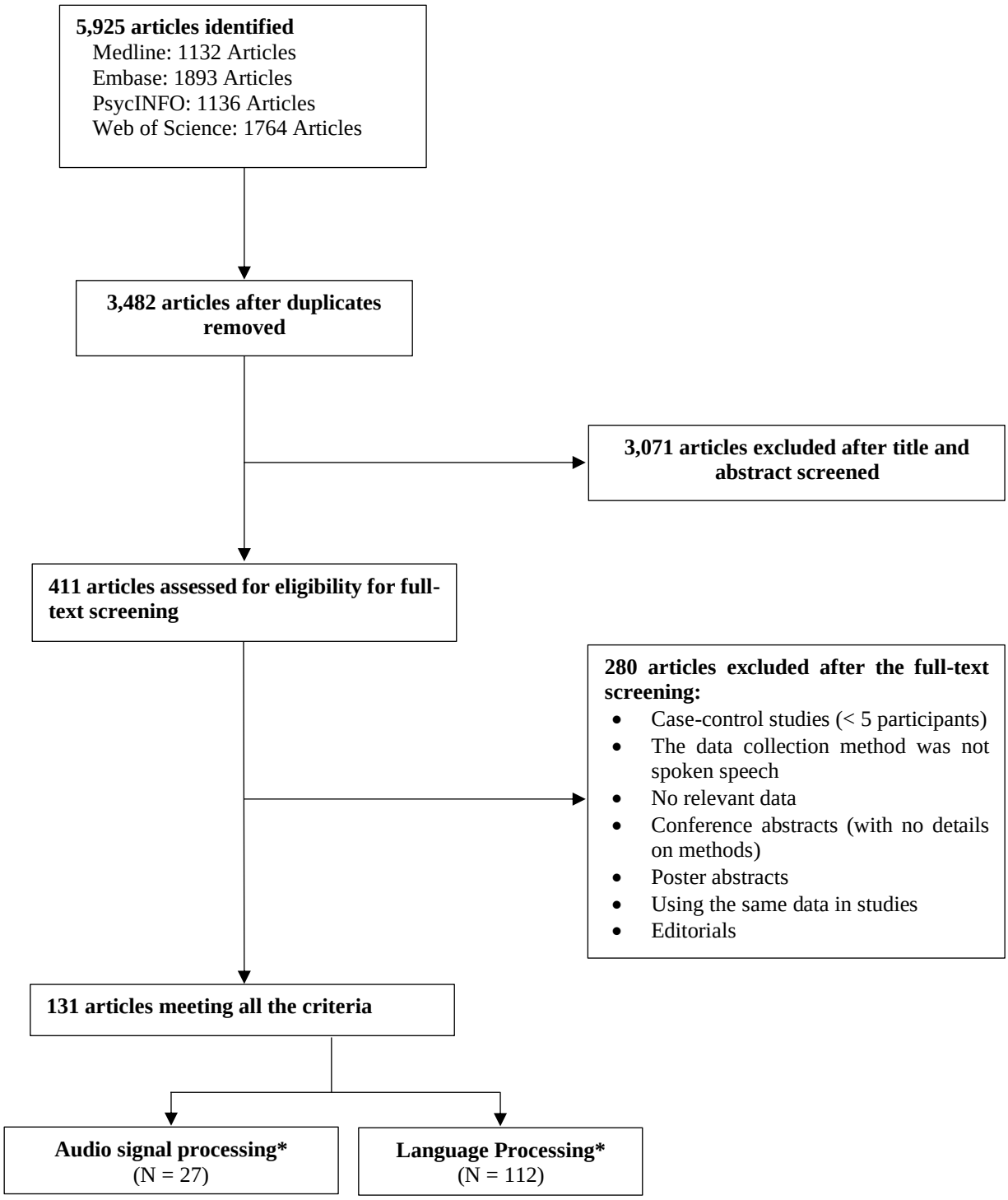
Note: .mp in the search term represents "multipurpose" and is used for searching the term in multiple fields such as title, abstract, keywords, and unique identifier.

3.1.3 Study selection and data extraction

The inclusion criteria for this literature review were that the studies used speech-based cognitive assessment and reported manual or automatic data evaluation. As mentioned in section 3.1.2, the articles were not restricted to language in the search, but in the selection step, articles written in English were only selected.

Studies that did not collect speech data were excluded. Also, the studies reporting other dementia, including Primary Progressive Aphasia (PPA), PD (Parkinson's Disease), FD (Frontotemporal Dementia, and HD (Huttington Disease), were excluded, with the exception that they also included patients with AD. The studies that did not have full text or insufficient information and case studies with a sample size of fewer than five participants were excluded from the selection.

After compiling the articles from multiple databases, duplicate entries were removed using the deduplication tool available within the EndNote application. Finally, titles, abstracts, and full text of articles (for potentially relevant studies) were screened. The following data were extracted from the final list of articles: information on the author, study year, study location, speech data collection methods, languages in which speech tasks were completed, participant selection, sample size, gender, and mean age of the sample population, including the type of in-clinic cognitive assessment used in determining different stages of AD, and statistical methods used for analysis. While reviewing the articles, it is assumed that the manual transcription technique was used if the method to acquire transcripts from the audio recordings was not reported. **Figure 3.1** provides the details on the selection of articles.



* 8 studies investigated both audio and language features

Figure 3.1 Flowchart describing the selection of articles in speech-based cognitive assessment on various stages of Alzheimer's Disease.

3.1.4 Study characteristics

An overview of the characteristics of identified articles is presented in **Table 3.2**. The study year ranged from 1982 to 2020, with a marked increase in the number of articles on this topic over time, with 79 articles published since 2011 (**Figure 3.2**). The sample size ranged from 5 to 521 participants. All the studies comprised male and female participants except three studies, where one study [130] contained only male participants and two studies [131], [132], only female participants. In most studies, the percentage of female participation was higher than males.

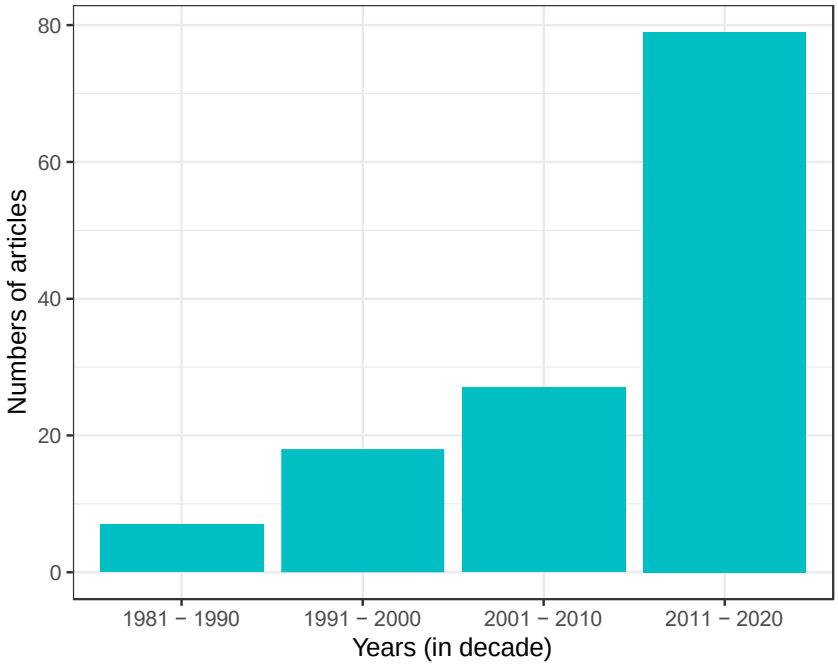


Figure 3.2 The number of articles published in ten years intervals between 1980 to 2020 after the final selection.

Detailed characteristics of the identified articles are presented below:

Speech data collection

Reviewed articles collected speech data using various speech elicitation tasks, mostly picture stimuli, including the Cookie Theft Picture from Boston Diagnostic Aphasia Examination (BADE) as a picture description task [64].

Other picture-based stimuli included:

- The Norman Rockwell picnic scene [133]
- Circus procession [134]
- Picture depicting a bank robbery [135]
- A frog story from a children's book [136]
- Various domestic/everyday scenes [77], [137]–[139].

One of the articles used the dog story containing seven pictures in sequence for the speech data collection method [140]. Participants were shown a silent film in another study and asked to recall it immediately [141]. Also, a few studies asked participants to tell well-known "Little red riding hood" or "Cinderella" fairy tale stories for speech production. Other studies used open-ended autobiographical questions including, "Tell me about the best day of your life" [54], "Tell me about the school you went to" [142], "Tell me about the previous day" [141], and "tell about your trip to New York City" [143]. Speech tasks such as describing the procedure of making a cup of tea [144], [145], changing a car tyre [130] and explaining the route on the map [146] were also used for eliciting speech. Conversational speech between participants and the in-clinic examiners was also used in speech analysis [59], [106], [111], [116], [132], [147]–[149]. A few studies reported the use of multiple speech elicitation tasks [130], [141], [150]–[153].

Languages used

The language used by participants in completing the task was mainly English. Other languages used for speech elicitation include Arabian, Basque, Brazilian Portuguese, Catalan, Chinese, Czech, Finnish, French, Deutsch, Hebrew, Italian, Japanese, Korean, Persian, Portuguese, Taiwanese, Turkish, Spanish, and Swedish (**Table 3.2**).

Disease status and diagnostic criteria

Most studies included patients with mild cognitive impairment (MCI) and probable, moderate, and severe AD. One of the studies reported Subjective Cognitive Decline (SCD) as a disease status in participants with high amyloid burden who reported experiencing memory, word finding and navigation difficulties [154]. Although few studies reported mixed patients with other types of dementia, the results from MCI and various stages of Alzheimer's are reported in this review. As discussed in section 1.1.2, for the diagnostic criteria, studies used the Global Deterioration Scale (GDS), Clinical Dementia Rating (CDR) scale, Geriatric Depression Scale (GDS), Mini-Mental State Examination (MMSE), Cambridge Cognition Examination (CAMCOG), Addenbrookes Cognitive Examination-Revised (ACE-R), Montreal Cognitive Assessment (MoCA), Alzheimer's Disease Assessment Scale-Cognitive Subscale (ASAD-Cog). A few studies used the Cambridge Mental Disorders of the Elderly Examination (CAMDEX), Generalized Anxiety Disorder Questionnaire and Patient Health Questionnaire in addition to the tests mentioned above. Some of the studies used more than one test as diagnostic criteria. The most commonly used diagnostic guidelines are those of NINCDS-ADRDA (**Table 3.2**).

Audio and language features

Table 3.3 reports the audio and language features studied on various disease severity groups. The description of features is available in Chapter 2.

Table 3.2 Characteristics of studies analysing speech-based features for different stages of Alzheimer's Disease. Values are displayed as Mean ± SD or Median (range) or (range) as reported in the study.

First Author	Study Year	Study location (Language)	Speech Data collection	No. Subjects	Gender (% Female)	Age (years)	Psychiatric Assessment	Features	Methodology
Ahmed [44]	2013	UK (English)	Picture description task	HC: 15; MCI: 15; Mild AD: NR	HC: 46 %; MCI; 40%; Mild AD: NR	HC: 76.0 ± 5.6; MCI 71.2 ± 8.5; MD 71.8 ± 7.9	CAMCOG, MMSE	Lexical, semantic, syntactic, and audio features	Independent t-test to compare demographic and cognitive measures between controls and patients. Mann-Whitney U-test to compare linguistic measures between controls and patients and Wilcoxon signed-rank test to compare linguistic measures between the clinical stages of AD. Repeated measure ANOVA to investigate linear trends in the quantitative production analysis
Ahmed [52]	2013	UK (English)	Picture description task	HC: 18; AD: 18	HC: 50%; AD: 33%	HC: 79.1±7.3; AD: 74.0± 8.8	CAMCOG, MMSE	Lexical and semantic features	Group comparison using independent samples t-test; association between semantic measures and performance using Pearson's correlation
Al-Hameed [98]	2019	UK (English)	Open ended biographical questions	FMD: 60, ND: 53	FMD: 25% ND: 28.3%	FMD: 57.8±2.0, ND: 63.7±2.3	ACE-R; MMSE; PHQ9; GAD7	Audio features	Classification between two groups using Linear SVM; Random Forest; Adaboost; MLP; Linear via SGD;
Alhanai [106]	2018	USA (English)	Audio recordings of neuropsychological examinations	HC: 71; Impaired: 21	51%	68±17	NR	Lexical, syntactic, and audio features	Binomial logistic regression model to classify between groups using different feature set
Almor [50]	1999	USA (English)	Semi-standardized interviews	HC: 9; AD: 11	NR	HC: 77 ± 4.4; AD: 80.7±6.8	MMSE, CT, MRI	Lexical and semantic features	Pearson's correlations with MMSE, semantic, and working memory impairment
Altmann [111]	2001	USA (English)	Speech data from conversation	HC: 15; pAD: 10	NR	HC: 76.4, pAD 80.4	MMSE	Lexical and syntactic features	ANOVA for group difference
Anna [155]	2008	France (NR)	Story narration	HC:16; AD:16	NR	HC: 78.4 ± 6.3; AD: 79.4 ±5.2,	MMSE; EGG; CT brain scan	Semantic features	Repeated measure ANOVA between groups; t-student tests between groups
Appell [58]	1982	Canada (English)	Story Recall	AD: 25; HC: 25	AD: 44%; HC: 44%	76 ± 7	NR	Lexical and semantic features	T-test for group difference
Aramaki [48]	2016	Japan (Japanese)	Written Narrative, Spoken Narrative	HC: 14; MCI: 7	HC: 50%; MCI: 42.85%	HC: 77.2 ± 2.1; MCI: 80.3 ± 4.2	HDS-R (similar to MMSE)	Lexical and syntactic features	Mann-Whitney U tests to compare the language abilities between the Healthy and Low HDS-R groups
Asgari [147]	2017	USA (English)	Speech data from conversation	HC:27; MCI:14	HC: 63%; MCI: 86%	HC: 78.9 ± 5.5; MCI: 83.4 ± 8	CDR	Lexical features	Nonlinear SVM and linear SVM Classifier between groups

First Author	Study Year	Study location (Language)	Speech Data collection	No. Subjects	Gender (% Female)	Age (years)	Psychiatric Assessment	Features	Methodology
Ash [61]	2007	USA (English)	Story narration	HC:10; AD:20	NR	NR	NR	Lexical and syntactic features	T-test for group difference
Ash [156]	2015	USA (English)	Picture Description Task	HC:19; naPPPA: 20; svPPA:14; lvPPA: 15; bvFTD: 15; AD: 18; PD: 26; PDD/DLB: 15	HC: 68% naPPA: 50% svPPA: 36% lvPPA: 47% bvFTD: 33% AD: 67% PD: 38% PDD/DLB: 33%	68.2 ± 9.2	MMSE	Semantic and audio features	Spearman's correlation between features and speech rate.
Baillon [150]	2004	UK (English)	Picture description task and story narration	VaD:68; AD:103	NR	NR	CT or MRI; CAMCOG; MMSE; MADRS, NART	Semantic features	Odds ratio; confidence interval; Mann-Whitney U for group difference
Beales [151]	2018	Australia (English)	Picture description task, story retelling, biographical narrative	svPPA: 3; lvPPA: 3; AD:3	33%	60 (55-86)	ACE-III	Lexical and syntactic features	Intra class correlation
Beltrami [157]	2016	Italy (Italian)	Picture description task	HC: 48; a-MCI:16, md-MCI:16, e-D: 16	NR	NR	MMSE	Audio features	ML classifier
Blanken [57]	1987	Germany (Deutsch)	Semi-standardized interviews	HC:5; SDAT:10, Wernicke's aphasia: 5	45%	HC:77 (69-80); SDAT: 83 (67-89), Wernicke's: 53 (48-71)	SDAT with CT and EEG scans,	Lexical features	Mann-U Whitney and Chi square test for group comparison
Brandão [158]	2009	Brazil/Spain (Spanish)	Biographical narrative	HC:18; AD: 18	NR	HC 79.5 ± 3.2; AD 79.8 ± 3.6	MMSE; Global Deterioration scale	Lexical features	Mann-Whitney test to compare Neuropsychological profile of the groups; Kruskall-Wallis test to compare incomplete and repeated propositions between groups; Pearson Correlations with cognitive variables
Bridges [159]	2013	USA (English)	Biographical narrative	HC: 5; EO: 5; LO: 6	NR	HC: 63.6 ± 13.8; EO 53.8 ± 5.5; LO: 85.5 ± 3.9	MMSE on AD participants	Lexical and semantic features	group difference Kruskall-Wallis test, Mann-Whitney's U, one-way ANOVA to check formulaic expression difference between groups
Broderick [160]	2018	USA (English)	Picture description task	HC: 96; AD: 142	NR	NR	NR	Audio features	PCA to dimensionality reduction; KNN & SVM for classification

First Author	Study Year	Study location (Language)	Speech Data collection	No. Subjects	Gender (% Female)	Age (years)	Psychiatric Assessment	Features	Methodology
Bschor [63]	2001	Germany (German)	Picture description task	HC: 40; MCI: 34; pDAT: 21; DAT: 20	HC: 55%; MCI: 50%; pDAT: 71.42%; DAT: 85%	65.9 ± 13	GDS and ICD-10	Lexical and semantic features	Chi-square and one-way analyses of variance (ANOVAs) for group difference; post hoc pair wise comparisons
Bucks [161]	2000	UK (English)	Semi-standardized interviews	HC: 16; pDAT: 8	HC: 56%; pDAT: 50%	HC 61.8 ± 7.7; pDAT: 67.4 ± 6.2;	MMSE	Lexical features	Group comparisons using Mann-Whitney U analysis; Classification using Linear Discriminant Analysis
Can [162]	2019	Turkey (Turkish)	Story-Picture Sequencing Test	HC:25; EOAD:25	51%	HC: 62.9; EOAD: 59.9	MMSE, Clinical dementia rating	Lexical, syntactic and semantic features	Chi-square test, median test
Can [62]	2019	Turkey (Turkish)	Picture description task	HC:26; LOAD: 39	NR	HC: 62.9; LOAD: 78.5	NR	Lexical, syntactic and semantic features	Group difference using Chi-square, U Mann-Whitney and student t-test
Chapman [163]	1998	USA (English)	Fables retell, Picture story generation	HC: 10; mFA 10; AD 10;	HC: 50%; mFA: 40%; AD: 50%	HC: 65 (47-75); mFA: 65 (47-74); AD: 65 (49-74);	MMSE; WAIS-R	Lexical and semantic features	Group comparisons on discourse tasks using repeated measure ANOVA
Chapman [164]	2002	USA (English)	Biographical narrative	HC: 25; MCI: 20; AD: 24	HC: 72%; MCI: 45%; AD: 75%	HC 76.1 ± 6.9; MCI 72.7 ± 6.1; AD 72.4 ± 9.6	MMSE	Lexical features	Jonckheere-Terpstra (JT) test to evaluate difference; Spearman rank order correlation to see relationship between gist level and detail; Kruskall-Wallis to see difference in age, gender, education level and MMSE
Chien [165]	2019	Taiwan (Taiwanese)	Picture description task	HC: 10; AD:10	HC: 60%; AD: 40%	HC: 67.2 ± 8.4; AD: 78.6 ± 6.5	NR	Lexical features	Feature sequence generator using Convolutional Recurrent Neural Network (CRNN) and bidirectional RNN for classification between groups
Choi [166]	2019	USA (English)	Picture description task	HC: 50; MCI: 50	HC: 62%; MCI: 60%	HC: 65.6 ± 6.80, MCI: 66.0 ± 8.38	MoCA, CDR, FAQ	Lexical and semantic features	Feed forward neural networks for classification using all the extracted features
Croisile [47]	1996	France (French)	Picture description task	HC: 24; AD: 22	70%	HC: 68.2 ± 7.8; AD: 70.7 ± 7.1	NR	Lexical and semantic features	ANOVA, student t-test
Croot [116]	2000	UK (English)	Conversational speech	AD: 6; DAT: 4	40%	65 (54-77)	MMSE; Neuroimaging showing Brain abnormality	Lexical and semantic features	Descriptive analysis

First Author	Study Year	Study location (Language)	Speech Data collection	No. Subjects	Gender (% Female)	Age (years)	Psychiatric Assessment	Features	Methodology
Cuetos [21]	2007	Spain (Spanish)	Picture description task	Carriers: 19; Non-carriers: 21	67.50%	Carriers: 43.2 ± 3.1; Non-carriers: 45.3 ± 3.7	MMSE; GDS	Lexical and semantic features	Performance of two groups were compared using multivariate analyses of variance (MANOVAs)
Davis [132]	2009	USA (English)	Conversational speech recordings	AD: 5	100%	85-90	NR	Lexical features	Quantitative analysis using Mean (SD) on extracted features
Dijkstra [167]	2004	USA (English)	Picture description task; biographical questions	HC: 30; Dementia: 30	NR	NR	MMSE	Lexical and semantic features	Group difference using independent t-test;
Duong [168]	2003	Canada (English)	Picture based narrative	HC: 27; AD: 5	NR	HC: 65-84; AD: 63-78	NR	Lexical and syntactic features	Descriptive analysis of discourse ratio and repetition of ideas
Drummond [169]	2019	Brazil (Brazilian Portuguese)	Narrative discourse	HC: 39, aMCI: 31, AD: 14	HC: 38%; aMCI: 48%; AD: NR	HC: 71.8 ± 4.3; aMCI: 72.2 ± 4.8, AD: 75.3 ± 6.8	MMSE, GDS, WMS-III, DSM-5, and MRI	Semantic features	Pearson's correlation of discourse to different AD spectrum (CSF-AD biomarkers)
Dunn [118]	2002	Australia (English)	Narrative recall	Intact: 58; Impaired: 14	75%	Intact 75 (64-92); Impaired: 78 (72-86)	MMSE, CAMCOG	Semantic features	Pearson's correlations between CAMCOG & MMSE with recall features
Ehrlich [137]	1997	USA (English)	Picture based narrative	HC: 16; DAT: 16	HC: 50%; DAT: 31.25%;	HC: 73.2 ± 5.3; DAT: 74.3 ± 7.4;	MMSE; ICM; WAB	Semantic features	ANNOVA for group difference
Faris [170]	2012	USA (English)	Open ended biographical questions	HC: 37; MCI: 21; Dementia: 23	NR	76.3 ± 6.86	DSM III R; CDR	Semantic features	Pearson's correlation coefficient to estimate relationship between ID and various demographic variables; t-test to compare ID across groups
Farivar [171]	2019	Iran (Persian)	Storytelling task	HC: 14; mild AD: 14	NR	NR	CDR	Lexical and semantic features	Independent t-test and Mann-Whitney U test to examine difference between groups
Faroqi-Shah [172]	2020	USA (English)	Picture description task	HC: 24; MCI: 20; AD: 20; PPA: 26	HC: 58%; MCI: 50%; AD: 50%; PPA: 46%	HC: 68.8 ± 7.2; MCI: 66.1 ± 9.2; AD: 68.0 ± 7.4; PPA: 68.6 ± 7.8	NR	Lexical and semantic features	One-way ANOVA for language measure for group; Kruskal-Wallis non- parametric test; pairwise comparisons using the Mann-Whitney U test
Fleming [152]	2009	USA (English)	Story retell and discourse production	HC: 9; MCI: 8	NR	HC: NR; MCI: 71.75 ± 7.48	BNT; LMS MMSE	Lexical and semantic features	Spearman correlation for participants initial and follow-up performance; Wilcoxon Signed Ranks Test to compare group performance on the initial administration and performance on the 6-month follow-up

First Author	Study Year	Study location (Language)	Speech Data collection	No. Subjects	Gender (% Female)	Age (years)	Psychiatric Assessment	Features	Methodology
Forbes [51]	2002	UK (English)	Picture description task	minimal AD: 11; mild AD: 11	NR	minimal AD: 73 ± 6.8; mild AD: 78.2 ± 4.5	MMSE	Audio features	Multiple regression analysis to examine demographic influences
Fraser [42]	2015	Canada (English)	Picture description task	HC: 233; AD: 240	HC: 65%; AD: 66%	HC: 65.2 ± 7.8; AD: 71.8 ± 8.5;	MMSE	Lexical and semantic features	Principal axis factor for factor analysis; logistic regression method for correlation with diagnosis
Fraser [173]	2019	Canada (Swedish)	Picture description task	HC: 29; MCI: 26	HC: 72%; MCI: 54%	HC: 67.8 ± 7.7, MCI 70.6 ± 5.8	MMSE and GDS	Lexical, semantic and audio features	Cascaded ML classifier (LR, SVM as baseline); Spearman's correlations between classifier predictions and neuropsychological test scores
Fraser [174]	2019	Canada (English and Swedish)	Picture description task	Gothenburg (HC: 36; MC: 31;), Dementia Bank (HC: 19; MCI: 19); Karolinksa (HC: 96); Dementia bank (HC: 78) Dementia Bank (HC: 97; pAD: 166)	Gothenburg (HC: 64%; MCI: 52%); DementiaBank (HC: 53%; MCI: 53%); Karolinska (HC: 54%); Dementia bank (HC: 62%) DementiaBank (HC: 60%; pAD: 67%)	Gothenburg (HC: 67.9 ± 7.2; MCI: 70.1 ± 5.6); Dementia Bank (HC: 66.4 ± 9.2; MCI: 66.7 ± 8.5); Karolinska (HC : 57.2 ± 19.9); Dementia Bank (HC: 63.9 ± 7.8); Dementia Bank (HC: 64.4 ± 8.1; pAD: 71.9 ± 8.3)	NR	Lexical and semantic features	Group comparisons with multi way ANOVA; using pre-trained word vectors (GloVe) to extract 300-dimensional word vectors; Classification using linear SVM classifier
Frobes-McKay [175]	2013	UK (English)	Picture description task	Cross-sectional study (HC: 30; AD: 31); Longitudinal study (AD: 15)	Cross-sectional study: (HC: 40%, AD: 38.70%); Longitudinal study: NR	Cross-sectional study (HC: 78.3 ± 4.6, AD: 76.0 ± 8.6); Longitudinal study (AD: 76.5 ± 7.9)	MMSE	Lexical features	One-way ANOVA for group difference, repeated ANOVA for longitudinal assessment
Gayraud [54]	2011	France (French)	Biographical narrative	HC: 20; pAD: 20	NR	pAD: 76.6 ± 9.2, HC: 76.85 ± 5.29	GDS and MMS	Semantic features	Group comparison using Mann-Whitney U-test
Gely-Nargeot [176]	2002	France (French)	Narrative recall	HC: 14; pAD: 14	HC: 43%; pAD: 43%	HC: NR; pAD: 67 (54-75)	MMSE; GDS	Lexical and semantic features	ANOVA for group difference
Giles [145]	1996	UK (English)	Picture description task	HC: 18; minimal: 16; mild: 15; moderate:17	73%	HC: 68.4 ± 7.3, minimal 72.8 ± 7.1, mild 66.6 ± 8.7, moderate 63.3 ± 9.1	MMSE	Syntactic and semantic features	One-way ANOVA for group difference, Newman-Keuls test for post-hoc pairwise difference between groups

First Author	Study Year	Study location (Language)	Speech Data collection	No. Subjects	Gender (% Female)	Age (years)	Psychiatric Assessment	Features	Methodology
Glosser [177]	1990	USA (English)	Biographical narrative	HC: 17; FA:9; CHI:9; AD:9	NR	HC: 55.06; FA: 60.22; CHI: 24.33; AD: 64.33	BDAE, MMSE	Audio features	Interviews were manually transcribed; verbal production were segmented; ANOVA for group difference in the length of discourse; MANOVA for group difference on the syntactic, lexical error and discourse cohesion
Gosztolya [141]	2019	Hungary (Hungarian)	Story recall and open-ended interview	HC: 25; MCI: 25; mAD:25	NR	HC: 70.7 ± 5.0; MCI:72.4 ± 3.6; mAD:73.9 ± 6.8	MMSE, CT, MRI, ADAS-Cog	Lexical and audio features	Extracting audio and language features using ASR, Classification between groups using SVM
Groves-Weight [178]	2004	USA (English)	Picture description task	HC: 14, mild AD: 14; moderate AD: 14	HC: 50%; mild AD: 50%; moderate AD: 36%	NR	MMSE	Lexical features	Four two-way ANNOVAs to compare verbal and written task on groups
Guinn [179]	2015	USA (English)	Conversational speech	56	NR	NR	NR	Lexical and syntactic features	Classification using decision trees and Naive Bayes classifier
Haider [180]	2020	UK (English)	Picture description task	HC: 82; AD: 82	56 %	50-80	MMSE	Audio features	Group classification using KNN, LDA, Random Forest and Support Vector Machine
Harris [181]	2008	USA (English)	Discourse Production Task	YA: 30; OA: 22; ND: 10; 10 MCI	NR	YA: 20.3 ± 2.3, OA: 66.6 ± 13.7; ND: 63.5 ± 9.7; MCI: 73.3 ± 8.8	MMSE	Lexical and semantic features	Difference between groups on standard tests and experimental discourse task with MANOVA and ANOVA, relationship between various standardised test scores through Pearson's correlation
Hays [144]	2004	New Zealand (English)	Picture description task and open-ended questions	HC: 13; AD: 13	HC: 85%; AD: 85%	HC: 77.7 ± 3.7; AD: 78.8 ± 8.1	MMSE	Semantic features	Mixed analyses of variance (ANOVAs); repeated measures ANOVA was used to test for group differences on measures of interest. Associations of pragmatic scales and other measures were explored with Pearson's R coefficients
Heller [182]	1992	Canada (English)	Narrative recall task	HC: 16; AD: 16	HC: 56%; AD: 56%;	HC: 67.4 ± 10.3; AD: 67.1 ± 10.0;	NINCDS and ADRDA	Semantic and audio features	ANOVA for group difference
Hernandez-Dominguez [183]	2018	Canada (English)	Picture description task	HC: 217; MCI: 43; AD: 257	HC: 65%; MCI: 37%; AD: 66%	NR	MMSE	Lexical and semantic features	Classification between groups using SVM, Random Forest using mixture of features
Hoffmann [184]	2010	Hungary (Hungarian)	Audio recordings of neuropsychological examinations	HC: 15; mild AD: 10; moderate AD: 10; severe AD: 10	NR	HC: 67.8 ± 8.5; mild AD: 71 ± 8.5; moderate AD: 69.4 ± 9.4; severe AD: 71.4 ± 3.1;	MMSE; CT or MRI	Lexical features	Kruskal-Wallis and Mann-Whitney for group difference and spearman correlation on feature set with MMSE
Illes [185]	1989	USA (English)	Biographical narrative	HC: 10; eAD: 5 mAD: 5, eHD: 5; mHD: 5; ePD: 5; mPD:5	NR	HC: NR; eAD: 71.2; mAD: 64.5; eHD: 44.0; mHD: 52.4; ePD: 56.4; mPD: 66.0	GDS	Lexical and semantic features	ANOVA and MANOVA for group differences

First Author	Study Year	Study location (Language)	Speech Data collection	No. Subjects	Gender (% Female)	Age (years)	Psychiatric Assessment	Features	Methodology
Kavé [186]	2018	Israel (Hebrew)	Picture description task	HC: 35; AD: 35	61%	HC: 76.7 ± 4.2; AD:77.9 ± 6.3	MMSE	Lexical and semantic features	Speech samples were transcribed, t-test to find group difference among features, Pearson's correlations to examine association between MMSE and language feature
Kemper [56]	1994	UK (English)	Picture description task	AD: 16, Spouse: 16	AD: 38%, Spouse: NR	AD: 67.8 ± 4.5; Spouse: 64.0 ± 9.2	MMSE	Syntactic and semantic features	Pearson's correlation between results of AD and spouse
Kemper [187]	2001	UK (English)	Open ended questions	HC: 30	NR	71.76 ± 2.98	MMSE	Syntactic and semantic features	Graphical plots to examine individual growth curves using spaghetti plots, mixed effect models
Khodabakhsh [188]	2014	Turkey (Turkish)	Conversational speech recordings	HC: 20; AD: 20	NR	NR	NR	Lexical features	SVM, LDA, decision tree
Khodabakhsh [189]	2015	Turkey (Turkish)	Conversational speech recordings	HC: 51; AD: 28	HC: 39%; AD: 36%	HC: 75.9 ± 6.4; AD: 75.0 ± 6.0	NR	Lexical and audio features	SVM, Naïve Bayes classifiers, and classification tree (Ctree)
Kim [190]	2019	South Korea (Korean)	Picture description and narrative discourse tasks	HC: 21; aMCI: 30, naMCI: 22,	73%	HC 71.9 ± 6.8; aMCI 73.8 ± 6.4; naMCI 70.1 ± 6.3,	MMSE, SLVT, CDR, RCFT	Lexical features	Logistic regression analysis
Konig [100]	2015	France (French)	Conversational speech recordings	HC: 15; MCI: 23; AD: 26	HC: 60%; MCI: 52%; AD: 50%	HC: 72 (60-79); MCI: 73 (67-79); AD: 80 (72-86)	MMSE, ADLs	Audio features	Mann-Whitney U test, Fisher's exact test, Mean EER and SE
Lai [191]	2012	Taiwan (Chinese)	Picture description and narrative discourse	HC: 30; AD: 30	NR	NR	CT and MRI; CDR	Lexical and semantic features	Chi square analysis and Fisher exact test to examine whether two groups differ significantly.
Lai [192]	2014	Taiwan (Chinese)	Conversational speech recordings	HC: 20; AD: 20	NR	HC 68.2 ± 3.8; AD: 72.8 ± 5.2	CT and MRI; CDR	Lexical features	ANOVA for difference in groups
Laine [193]	1998	Finland (Finnish)	Conversational speech recordings	HC: 19; AD: 11; VaD: 8	NR	HC: 67.4 ± 6.3; AD: 67.4 ± 8.2; vaD: 67.1 ± 6.5	MMSE; CDR;	Lexical and semantic features	Group comparisons with one-way ANOVA
Lehr [194]	2012	USA (English)	Story recall task	HC: 37; MCI: 35	NR	HC 87.3; MCI 87.2	CDR	Semantic features	Classification using SVM
Lira [125]	2014	Brazil (Brazilian Portuguese)	Picture description task	HC: 20; mild AD: 15; moderate AD: 11	NR	HC: 71.1 ± 5.2, mild AD: 68.3 ± 9.8, moderate AD: 75.7 ± 7.8	MMSE	Lexical and semantic features	Group difference using Tukey test; Linear regression analysis for the relationship between groups and discourse performance

First Author	Study Year	Study location (Language)	Speech Data collection	No. Subjects	Gender (% Female)	Age (years)	Psychiatric Assessment	Features	Methodology
Lira [140]	2019	Brazil (Brazilian Portuguese)	Story narrative	HC: 61; AD: 60	NR	HC: 69.7 ± 6.5; AD: 77.2 ± 6.4	MMSE; ADAS-Cog	Lexical and semantic features	Group difference using Student's t-test, Mann-Whitney Test; stepwise forward likelihood ratio criterion to identify the relevant combination of micro linguistic items for group differentiations
Liu [195]	2020	China (Chinese)	Open ended biographical questions	HC: 15; AD: 15	NR	HC: 82.1 ± 4.3; AD: 85.7 ± 4.9	NR	Lexical and semantic features	Language network based on the dependency grammar
Lock [196]	1997	UK (English)	Picture description task	NYA: 20; NOA: 20; AD: 20; ANO: 20	NYA:90%; NOA: 50%; AD: 85%; ANO: 40%	NYA: 17-38 NOA: 60-84; AD: 71-87 ANO: 61-84	MMSE, NOA - BDAE	Lexical and semantic features	One-way ANOVA to compare means between groups; Chi square to compare group pairs
López-de-Ipiña [197]	2013	Spain (English, French, Spanish, Catalan, Basque, Chinese, and Arabian)	Biographical narrative	HC: 20; AD: 20	HC: 50%; AD: 40%	HC: 20-98; AD: 68-96	NR	Audio features	Classification accuracy using Multi-layer perceptron (MLP), SVM, Decision Trees (DT), KNN, Naive Bayes
Lunsford [45]	2015	USA (English)	Narrative recall task	HC 28; Impaired: 15	NR	NR	NR	Lexical features	Wilcoxon Rank Sum and Fisher Exact test for group difference
Lyons [59]	1994	USA (English)	Interviews at psychometric assessment	HC: 108; very mild AD: 22; mild AD: 20	HC: 64%; very Mild AD: 55%; mild AD: 55%	HC: 76.2 ± 9.5, very mild, 75.3 ± 11.5, mild 74.3 ± 8.0	CDR	Lexical and syntactic features	One-way analysis of variance between groups
March [146]	2009	Australia (English)	Picture and route description tasks	HC: 26, pAD: 26	NR	NR	MMSE	Lexical and syntactic features	Group difference using Mann-Whitney U test, correlations between language use and neuropsychological variables on combined group data; regression analyses for examining the neuropsychological predictors of spatial deixis
Mueller [198]	2016	USA (English)	Picture description task	HC: 39; pMCI: 39	HC: 56%; MCI: 56%;	HC: 63.1 ± 6.3; pMCI: 63.1 ± 6.4	CES-D, Parental family history of AD, APOE e4 alle	Lexical, syntactic, and semantic features	Features were z-score; paired sample t-test or Chi square test for group difference; Pearson correlation coefficients among discourse verbal fluency, verbal memory
Mueller [49]	2018	USA (English)	Picture description task	HC: 200; early MCI: 64	HC: 70%; early MCI: 56%	HC: 61.1 ± 6.5; early MCI: 64.2 ± 5.9	MMSE; RAVT;	Lexical and semantic features	Pearson's correlations between language factors and standardized neuropsychological tests; group difference on language variables using ANCOVA
Mirheidari [199]	2017	UK (English)	Conversational speech recordings	FMD: 15; ND: 15	FMD: 60%; ND: 53%	FMD: 57.8 ± 2.02; ND: 63.7 ± 2.3	ACE-R, MRI brain imaging	Lexical features	ML Classifiers: SVM, Random Forest, Adaptive Boost (AdaBoost), Perceptron, and Stochastic Gradient Descent (SGD)

First Author	Study Year	Study location (Language)	Speech Data collection	No. Subjects	Gender (% Female)	Age (years)	Psychiatric Assessment	Features	Methodology
Mitzner [131]	2003	USA (English)	Biographical narrative	118	100%	83 ± 2.9	MMSE; ADLs	Lexical, syntactic, and semantic features	Correlations between age, education, MMSE, ADL, oral and written scores
Nasrolahzadeh [200]	2016	Iran (Persian)	Biographical narrative	HC: 30; AD: 30	HC: 50%; AD: 53%;	HC: 52-58; AD: 52-98	MMSE, CDR	Audio features	visual comparison of signal variance; cross-correlation vector- optimum filter coefficient and point of error surface of two groups
Nasrolahzadeh [110]	2020	Iran (Persian)	Biographical narrative	HC: 30; AD: 30	HC: 50%; AD: 53%	HC: 52-58; AD: 52-98	MMSE, CDR	Audio features	Classification using pattern recognition neural network (PRNN)
Nevler [201]	2019	USA (English)	picture description task	HC: 31; AD: 89 lvPPA: 26	NR	NR	NR	Audio features	Compared between carriers/ non-carriers; and sex using Pearson's correlation.
Nicholas [202]	1985	USA (English)	Picture description task	AD: 19 HC: 30; Wernicke's: 16; anomic Aphasia: 8	NR	HC: 62.9 ± 9.0; AD: 66.5 ± 8.7; Wernnicke's: 61.0 ± 9.6; Anomic Aphasia 57.6 ± 6.5;	DSM III, Boston Diagnostic Aphasia Examination (BDAE);	Lexical and semantic features	Multiple t-test to compare group pairs
Nitrini [203]	2008	Brazil (Brazilian Portuguese)	Narrative recall task	363	NR	49.2 ± 13.9	MMSE,	Semantic features	Mann U Whitney test for comparing test scores with demographics; spearman correlation between test scores and demographics
Orimay [204]	2017	Malaysia (English)	Picture description task	HC: 99; pAD: 99	NR	HC: 65.3 ± 8.4; pAD: 70.5 ± 8.9	NR	Lexical, syntactic features	Student t-test, logistic regression and ML
Orimay [205]	2018	USA (English)	Picture description task	Study 1: (HC: 99; pAD: 99;); Study 2: (HC: 19; MCI: 19)	NR	Study1: (HC: 46-81; pAD: 55-90;); Study2: (HC: 55-74; MCI 49-81)	MMSE	Lexical, syntactic, and semantic features	DNN language model classification using stochastic gradient descent optimization.
Ostrand [206]	2020	USA (English)	Picture description task; biographical narrative	Session 1: (MCI: 39); Session 2: (MCI: 30)	Session 1: (MCI: 69 %); session 2: (MCI: 63%)	NR	3 MS, Hopkins Verbal Learning Test	Lexical, syntactic, and semantic features	ANOVA for group difference
Perri [207]	2013	Italy (Italian)	Story recall task	HC: 34; AD: 20; fvFTD: 20; SIVD: 20; LBD: 20;	HC: NR; AD: 40%; fvFTD: 50%; SIVD: 20%; LBD: 40%	HC: 68.5 ± 7.1; AD: 66.1 ± 7.3, fvFTD: 64.0 ± 12.3, LBD: 72.1 ± 6.2; SIVD: 70.0 ± 10.8	CT, MRI, NINDS-AIREN criteria, MMSE for AD	Semantic features	Two-way ANCOVAs (age as covariate) between features/ within features; ROC to evaluate accuracy of discriminating group
Pistono [208]	2019	France (French)	Picture based narrative	HC: 17; AD: 17	HC: 53%; AD: 59%	HC: 69 (66-73); MCI: 72 (66-77)	MMSE, MRI, Amyloid assessment	Lexical and semantic features	Group comparison using Mann-Whitney U-test and Chi-square test
Pistono [67]	2019	France (French)	Picture based narrative	HC: 17; AD: 17	HC: 53%; AD: 59%	HC: 69 ± 5; AD: 72.1 ± 2	MMSE	Lexical and semantic features	Pearson’s correlations between pause frequency in the AD participants
Pompili [209]	2020	Portugal (English)	Picture description task	HC: 98; AD: 147	NR	HC: 46-80; AD: 53-88	MMSE	Lexical, syntactic, and semantic features	Random Forest Classifier
Powell [210]	1988	USA (English)	Picture description task	MID: 18; DAT: 14	MID: 11%; DAT: 50%	MID: 67.4 ± 2.5; DAT: 73.1 ± 2.6	MMSE	Semantic and audio features	Pearson correlation between MMSE and speech and language measurements

First Author	Study Year	Study location (Language)	Speech Data collection	No. Subjects	Gender (% Female)	Age (years)	Psychiatric Assessment	Features	Methodology
Prud'hommeaux [211]	2011	USA (English)	Narrative recall task	HC: 52; MCI: 72	NR	HC: 87 ± 4.6; MCI: 88.7 ± 6.0	CDR	Semantic features	Retellings were hand aligned; SVM used for classification between disease status
Prud'hommeaux [212]	2015	USA (English)	Narrative recall and picture description task	HC: 163; MCI: 72	NR	HC: 87.0 ± 4.6; MCI: 88.7 ± 6.0	CDR; MMSE	Semantic features	Word alignment for WLM Scoring using graph-based method; Latent Semantic Analysis (LSA) cosine distance, BLEU (Bilingual Evaluation understudy), unigram for text similarity, Group classification using LibSVM and evaluation of classifier via receiver operating characteristic (ROC) curves;
Rentoumi [213]	2014	UK (English)	Picture description task	pAD: 18; mAD 18	pAD: 33%; mAD: 67%	pAD: 74.3 ± 9.13; mAD: 75.1 ± 7.34	MMSE CAMCOG	Lexical and semantic features	Representation of narrative speech transcripts by vectors (ngrams); Naïve Bayes classifier (NB) trained in combining features
Ripich [114]	2000	USA (English)	Picture description task	early AD: 10; mid AD: 10; late AD: 10	early AD: 80% mid AD: 80% late AD: 80%	early AD: 70.8 ± 7.8; mid AD: 78.0 ± 4.5; late AD: 77.6 ± 7.9	MMSE	Lexical and semantic features	ANOVA and ANCOVA (Analysis of covariance)
Roark [214]	2011	USA (English)	Story recall task	HC: 37, MCI: 37	NR	HC: 88.8 ± 6.0, MCI: 89.8 ± 5.7	CDR, MMSE, CERAD, WLM	Lexical and semantic features	SVM for classification between groups
Romero [215]	1996	Germany (Deutsch)	Open ended biographical questions	63	73.02%	73 (56-87)	CAMDEX CDR	Syntactic, semantic, and audio features	Wilcoxon test (matched-pair design), Mann-Whitney U test, Spearman's rank correlation
Sajjadi [216]	2012	UK (English)	Picture description task and narrative discourse	HC: 30; SD: 16; mild AD: 20	HC: 50%; SD: 50%; mAD: 40%	HC 67.5 (51-80); SD: 67.0 (60-79); AD: 68.0 (60-79)	MMSE; ACE-R; WMS	Syntactic and semantic features	One-way ANOVA to compare the performance in different groups; Kruskal-Wallis test was used for multiple group comparisons; paired t-test and Wilcoxon test to compare different sample lengths
Sherratt [130]	2007	Australia (English)	Narrative recall and procedures description	32	0%	50-94	NR	Lexical and syntactic features	Spearman's rho correlation test
Shi [217]	2014	China (Chinese)	Story recall	HC: 249; MCI: 134, AD: 97	NR	NR	MMSE, CDR, HAMD, ADL	Semantic features	Sensitivity and specificity using Receiver Operating Characteristic (ROC); positive predicative values and negative predicative values; correlation between MMSE and story recall

First Author	Study Year	Study location (Language)	Speech Data collection	No. Subjects	Gender (% Female)	Age (years)	Psychiatric Assessment	Features	Methodology
Shimada [139]	1998	Japan (Japanese)	Picture description task	HC:17 AD: 23	HC: 59%; AD: 74%	HC: 55-87; AD: 60-88	CG: medical interview; physical and MMSE; laboratory test, ECG, CT, AD: CT; NINCDS-ADRDA	Semantic features	Independent t-test for difference of task variables; Pearson's correlation between Cerebral Blood Flow and information conveyed and efficiency of descriptions
Sidtis [149]	2015	USA (English)	Structured conversation	HC: 18; PD: 16; AD: 12	NR	HC: 69.6; PD: 66.7, AD: 77.9	MMSE	Lexical and semantic features	Student t-test for group difference
Simpson [218]	2019	Canada (English)	Picture description task	111	NR	55-90	MoCA	Lexical, syntactic, and semantic features	Mean score difference between baseline and 6 months change and Pearson's correlation with MoCA score
Sluis [219]	2020	Australia (English)	Picture description task	HC: 20; moderate AD: 20; mild AD: 20	HC: 65%; moderate: 70%; mild: 75%;	HC: 66.9 ± 8.7; moderate 72.2 ± 7.6; mild 70.1 ± 9.4	MMSE	Audio features	one-way ANOVA for group difference
Smith [220]	1989	Australia (English)	Picture description task	HC:18; AD: 18	NR	NR	NR	Lexical and semantic features	T-test for independent measures
Smolik [221]	2016	Czech (Czech)	Biographical narrative	HC: 20; aMCI: 20	50%	HC: 72.1 ± 7.4; MCI: 72.1 ± 7.9	MMSE	Semantic features	Group comparison using t-test and the relations between measures using correlation coefficients
Stevens [222]	1992	UK (English)	Picture description task	Dysphasic: 8, SDAT: 9	Dysphasic: 38%; SDAT: 89%	Dysphasic: 75 ± 6.2; SDAT: 83.2 ± 6.0	BDAE; mental test score (Hodkinson, 1972)	Lexical, syntactic, and semantic features	Multiple regression analysis
Stuck [223]	2018	USA (English)	Speech recordings	HC: 35; Dementia: 107	NR	NR	NR	Audio features	Random Forest Classifier
Sumali [107]	2020	Japan (Japanese)	Free talk	53	NR	57-84	MMSE, GDS	Audio features	SVM
Suribhatla [153]	2004	UK (English)	Picture description and story recall task	EO: 40 LO: 90	EO: 33%, LO: 36.%	EO: 60.0; LO: 77.0	CAMCOG; MADRS; NAR	Semantic features	Spearman's correlation between CAMCOG and age; odds ratio between groups
Taler [224]	2020	Canada (English)	Story recall	Study 1 Young: 21 Old: 20 Study 2 Old:18 MCI:18	NR	Study 1 Young 22.4 ± 2.4 Old 75.2 ± 7.9 Study 2 Old 69.3 ± 4.7 MCI 75.6 ± 6.6	MoCA; CVLT-II	Semantic features	Repeated measures ANOVA for each scoring method

First Author	Study Year	Study location (Language)	Speech Data collection	No. Subjects	Gender (% Female)	Age (years)	Psychiatric Assessment	Features	Methodology
Themistocleous [93]	2020	Sweden (Swedish)	Picture description task	HC: 30; MCI: 25	58%	59.0 ± 6.4	MMSE	Audio features	Regression analysis
Thomas [225]	2005	Canada (English)	Open ended biographical questions	95	NR	NR	MMSE	Syntactic and semantic features	Classification based on discretized MMSE scores
Thomas [55]	2020	USA (English)	Story recall task	HC: 69; Impaired: 61; Mild: 62; Moderate: 39; Severe: 25	HC: 54%; Impaired: 57%; Mild: 61; Moderate: 69%; Severe: 60%	HC: 77.7; Impaired: 80.0; Moderate 84.9; Severe: 85.0	DSM 4 criteria	Audio features	Chi squared test and Spearman correlation
Toledo [226]	2018	Brazil (Brazilian Portuguese)	Story narration	HC: 20; aMCI: 20; mAD: 20	NR	HC: 74.8 ± 11.3; aMCI: 73.3 ± 5.9, mAD: 78.2 ± 5.1	CDR, DSM-IV criteria, MMSE, VFT	Semantic features	Kruskal-Wallis test, Tukey multiple comparison test
Tomoeda [65]	1996	USA (English)	Picture description task	GDS1: 52; GDS3: 5; GDS4: 32; GDS5: 31	GDS1: 69%; GDS2: 80%; GDS4: 66%; GDS5: 77%	GDS1: 72.2 ± 7.0, GDS3: 73.3 ± 6.4, GDS4: 74.9 ± 9.0, GDS5: 78.2 ± 7.0	MMSE, GDS	Lexical and semantic features	Mean performance and moving average measures compared to MMSE score, Logistic regression
Toth [227]	2018	Hungary (Hungarian)	Story recall task	HC: 35; MCI: 48	HC: 66%; MCI: 67%	HC: 64.1 ± 7.1; MCI: 73.1 ± 7.9	MMSE, ADAS-COG	Audio features	Features are extracted from ASR; Classification between groups using ML (Naïve Bayes; Random Forest; SVM)
Treviso [228]	2018	Brazil (Brazilian Portuguese)	Picture description task	HC: 20; MCI: 20; AD: 20	HC: 80%; MCI: 70%; AD: 50%	HC: 74.8 ± 11.3; MCI: 73.3 ± 5.9; AD: 78.2 ± 5.1	NR	Lexical, syntactic, and semantic features	Group difference using Mann-Whitney nonparametric test for unpaired data
Tsantali [229]	2013	Greece (Greek)	Picture description task	HC: 38; aMCI: 28, mAD: 53	HC: 58%; aMCI: 32%; mAD: 62%	HC: 66.5 ± 5.3; aMCI: 69.2 ± 7.3, mAD: 66.9 ± 6.9	MMSE	Lexical and semantic features	Pearson's correlations between BDAE scores and MMSE scores, One-way ANOVA for performance difference between groups
Verfaillie [91]	2019	Netherlands (Dutch)	Picture description task	LAB: 44; HAB: 19	LAB: 39%, HAB: 45%	LAB: 61.5 ± 7.4; HAB: 68.2 ± 8.1	Either amyloid PET scans or CSF; MMSE	Lexical, syntactic, and semantic features	Analyses of variance (ANOVA) for continuous demographic and neuropsychological data; Mann-Whitney U test for (raw) linguistic data and Spearman Rho and t-test for linguistic parameters and sex
Vincze [230]	2016	Hungary (Hungarian)	Story recall	HC: 36; MCI: 48	HC: 64%; MCI: 67%;	HC: 69.3; MCI: 73.1	NR	Lexical, syntactic, and semantic features	Classification using SVM; Evaluation of SVM using the accuracy, precision, recall and F-measure metrics
Vuorinen [80]	2000	Finland (Finnish)	Picture description task	HC: 20; AD: 13; VaD: 10	HC: 50%; AD: 54%; VaD: 60%	HC: 67.4 ± 6.1; AD: 66.6 ± 7.7; VaS: 68.4 ± 7.0	MMSE, CDR	Semantic features	One-way ANNOVAs and group comparisons (Tukey Test)

First Author	Study Year	Study location (Language)	Speech Data collection	No. Subjects	Gender (% Female)	Age (years)	Psychiatric Assessment	Features	Methodology
Walker [142]	2020	UK (English)	Semi-standardized interviews	HC: 14; MCI: 8	HC: 57%; MCI: 63%	NR	ACE-R, GAD 7; PHQ-9	Audio features	Mann-Whitney Test for group difference;
Wicklund [231]	2006	USA (English)	Story retelling	HC: 48; AD: 33; FTD: 20	HC: 27%; AD: 33% FTD: 60%	HC: 72.1 ± 7.2; AD: 72.6 ± 9.6; FTD: 61.9 ± 8.4	MMSE, WMS, CERAD	Semantic features	T-test on Immediate vs Delayed
Wisler [105]	2020	New Zealand (English)	Semi-standardized interviews	521	64%	74.1 ± 6.2	MoCA	Audio features	LASSO Regression Analysis
Yancheva [43]	2016	Canada (English)	Picture description task	HC: 98; AD: 168	NR	NR	NR	Semantic features	Clustering information units using k-means; Random Forest classifier with 10-fold cross-validation
Yu [232]	2018	USA (English)	Story recall	214	66%	NR	NR	Audio features	Classification between patient groups using SVM

Abbreviations: AD: Alzheimer's Disease; ADL: Activities of Daily Living; aMCI/MD: amnestic MCI/Multiple Domain; aMCI/SD: amnestic MCI/ Single Domain; aMCI: amnestic Mild Cognitive Impairments; ANO: Anomic Aphasic; ANOVA: Analysis of Variance; ASAD-Cog: Alzheimer's Disease Assessment Scale-Cognitive Subscale ; CAMCOG: Cambridge Cognitive Examination; CAMDEX: Cambridge Mental Disorders of the Elderly Examination; CDR: Clinical Dementia Rating; CERAD: The Consortium to Establish a Registry for Alzheimer's Disease; CHI: Closed Head Injury; CSF: Cerebrospinal Fluid; DAT: Dementia of Alzheimer's type; DAT: Dementia of the Alzheimer Type; eAD: early AD; eHD: early Huntington Disease; EO: Early Onset; EOAD: Early Onset Alzheimer's Disease; ePD: early Parkinson's Disease; FAQ: Function Assessment Questionnaire; FMD: Functional Memory Disorder; FTD: Frontotemporal Dementia; GDS: Global Deterioration Scale; HAB: High Amyloid Burden; HAMD: Hamilton Depression Scale; HC: Healthy Controls; ICD-10: International Classification of Diseases-10; ICM: Information Concentration-Memory test; LAB: Low Amyloid Burden; LO: Late Onset; mAD: moderate AD; MADRS: Montgomery and Asberg Depression Rating Scale; MANOVA: Multivariate Analysis of Variance; MCI: Mild Cognitive Impairments; mHD: moderate Huntington Disease; MID: Multi-Infarct Dementia; MMSE: Mini Mental State Examination; MoCA: Montreal Cognitive Assessment; mPD: moderate Parkinson's Disease; naMCI: non-amnestic Mild Cognitive Impairments; NART: National Adult Reading Test; NART: National Adult Reading Test; ND: Neurodegenerative Disorders; NOA: Normal Old Adults; NR: Not Reported; NYA: Normal Young Adults; OA: Older Adults; pAD: probable AD; PCA: Principal Component Analysis; PHQ-9: Patient Health Questionnaire-9; SCD: Subjective Cognitive Decline; SVM: Support Vector Machine; VaD: Vascular Dementia; WAIS-R: Wechsler Adult Intelligence Scale Revised; YA: Younger Adults

Table 3.3 Audio and language features analysed in different stages of AD.

Speech and language domains	SCD	MCI	Mild AD	Moderate AD	Severe AD	References on each domain
Lexical, Semantic and Syntactic						
Word finding and word retrieval difficulties	+	+	+	+	+	[21], [39], [42]–[45], [47], [50]–[52], [54]–[58], [62], [63], [66], [72], [78], [80], [91], [92], [105], [111], [114], [116], [118], [125], [130], [131], [137], [139], [140], [144]–[146], [150]–[152], [155], [156], [162]–[164], [166]–[172], [175]–[177], [181]–[183], [186]–[188], [190]–[193], [195], [196], [198], [203], [204], [206]–[209], [211], [214]–[218], [220], [221], [225]–[231], [233]–[235]
Part-of speech tags (noun, verb, adjective, pronoun rates), open and closed classed words, sentence structure	-	+	+	+	+	
Information units, information density	+	+	+	+	+	
Filled pauses	-	+	+	+	+	[31], [45], [54], [65], [67], [98], [116], [130], [147], [158], [168], [172], [185], [188], [190], [212], [219], [227]
Repetition, repairs, and incomplete words	-	+	-	+	+	
Audio features						
Speech fluency (speech rate, articulation rate, silent pauses, speech tempo, hesitation ratio)	-	+	+	+	+	[42], [55], [93], [98], [100], [105]–[107], [110], [142], [145], [157], [160], [165], [173], [178], [180], [184], [189], [197], [199]–[201], [223], [227], [232]
Spectral and voice quality (MFCC, intonations, pitch, duration of speech and silence, RMSE, filter bank energy coefficient, shimmer, jitter, voice breaks, zero crossing, HNR)	-	+	+	+	+	

Note: + represents features that are studied in the patient groups; - represents feature that were not studied on those patient groups; AD: Alzheimer's disease; HNR: Harmonics to Noise Ratio; MCI: Mild Cognitive Impairment; MFCC: Mel Frequency Cepstral Coefficients; MSE: Mean Square Error; RMS: Root Mean Square; SCD: Subjective Cognitive Decline

3.1.5 Use of computational methods and tools

This section discusses the use of computational methods for feature extraction and puts it into perspective within the research context. Studies have reported using existing technologies or

developed computational methods, including ASP and NLP, to extract relevant audio and language features (see **Figure 1.2**).

3.1.5.1 Use of automatic speech recognition for speech-to-text

Audio recordings need to be processed to get the transcription to extract language (text-based) features. Transcriptions can be obtained by manually listening to the speech data and recording the information or using an Automatic Speech Recognition (ASR) system. Out of 131 articles, only seven reported using ASR for transcription. Articles reported the use of several ASR systems such as Google Cloud Speech to Text (English language) [209], IBM system (English language) [194], Temi (English language) [166], and Kaldi-ASR (English and Italian languages) [43], [157], out of which the Kaldi-ASR is developed for research purpose [236], and the other three are commercially available ASR systems. To extract information about the content words, word-level transcription was used. However, some articles reported the extraction of phoneme (in the Hungarian language) [141] and utterance-level (in the English language) [234] transcriptions.

Studies showed that the transcription accuracy of commercial ARS systems did not reach the accuracy level obtained from manual transcripts; because atypical speech (elderly speech) recognition using ASR showed high transcription errors [209], [237]. Some of the articles developed their own ASR system for transcribing Hungarian and English speech responses, focusing mainly on extracting the temporal aspect of speech (speech rate, pauses, hesitation ratio) and not on the quality of the information content [141], [199], [227].

3.1.5.2 Use of computational methods and tools for feature extraction

Information Units were a widely extracted feature from the picture description tasks. The reviewed articles showed that WordNet [238] - a lexical database containing human-curated

semantic relationships between words and word embeddings [120] - an automated word representation where similar words have similar representation were used to extract information units automatically. Fraser et al., 2016 developed a semi-automatic approach to extract information units described in [47] by generating a set of synonyms using WordNet and searching those words in the transcribed narratives. In the same year, a method of identifying information content units (ICUs) using word vector representations from picture descriptions was developed [43].

Prior to the development of word embedding, another vector representation technique called Latent Semantic Analysis (LSA) was used in the analysis of AD speech. In the LSA technique, the cosine angle was evaluated between the vector representations of words [239]. Studies mostly used open source Natural Language Toolkit (NLTK) [240] to apply other NLP techniques, such as tokenisation (splitting paragraphs and sentences into words) [55], lemmatisation (reducing a word to its root meaning) [115], [241], removal of stop-words (common words including articles, prepositions, and conjunctions) [42], [206], part of speech tagging to extract relevant language information for the speech recordings [48], [213].

Studies used PRAAT [242], OpenSMILE toolkit [243], and their own implementations of audio signal processing algorithms to extract audio features directly from the raw speech data [100], [105], [180], [244]. The features (they are defined in Chapter 2) extracted using PRAAT were mainly speech fluency features (speech rate, articulation rate, number of pauses) and voice quality features (pitch, jitter, harmonics to noise ratio). Spectral features (they are defined in Chapter 2), including MFCC, alpha ratio, Hammarburg index, spectral slope, harmonic difference, and loudness, were extracted using the ComParE [109] or GeMAPS [103] feature set using OpenSMILE toolkit or the author's own implementation of the audio signal processing algorithms including Speech Activity Detector in Python or MATLAB.

A few studies used speech-based features extracted from spontaneous speech (phonemes, syllables, words, energy and entropy-based features) to train the deep neural network model and classified healthy and AD patients with a classification accuracy as high as 98.3% [110], [165]. The deep neural network model based on verbal utterance learned several linguistic biomarkers in the form of higher order *n*-grams (sequence of n words that appear together [245]) to distinguish the AD patients from the healthy older adults with greater accuracy (AUC = 0.91) [204]. These results showed that audio and language features are informative in distinguishing AD patients from healthy participants, encouraging these techniques to be used in earlier disease stages.

3.1.6 Informative speech and language features

The most evident problem in the earlier phase of AD is the word-finding difficulty, which reduces speech rate, and adds the use of pauses and more hesitation in speech. Some other speech characteristics that are known to be affected are the lexical richness of the language with repetitions, the use of few nouns or replacing them with pronouns. From the reviewed articles, computationally extracted feature names corresponding to the feature types reported by more than one article are listed in **Table 3.4** (for definitions of features, see Chapter 2). Most of the articles did not provide details on the feature extraction process. Some articles extracted audio features using ASP methods, and others used the word timing information obtained from the transcripts. Disease-specific associations with speech and language features reported in the reviewed literature will be discussed in section 3.1.7.

Table 3.4 Informative audio and language features used in the AD research organised by type of the features as reported in the reviewed articles.

Audio features	
Feature type	**Feature names**
Speech fluency	Speech rate; Articulation rate; Pause rate; Total speech; Average of syllables; Pauses; Maximum pause
Hesitations	Filled pauses; Silent pauses; Voice breaks; Hesitation ratio
Frequency related	Pitch (F0, fundamental frequency); Jitter; Formants (1, 2, and 3); Formant frequency and structure
Energy/Amplitude	Shimmer; Loudness; Harmonics to Noise (HNR) ratio; Zero crossing
Spectral features	Mel Frequency Cepstral Coefficients (MFCC); Alpha ratio; Hammarberg index; Spectral slope; Harmonic difference
Language Features	
Feature Type	**Feature names**
Lexical richness	Type-token ratio; Moving average type-token ratio; Brunet's index; Honor's stats
Frequency of use	Words; Unique words; Nouns; Verbs; Nouns; Predicates; Coordinate and subordinate phrases; Reduced phrases; Incomplete phrases; Fillers; Repetitions; Word replacement
Speech fluency	Total speech; Speech rate; Speech time; Average of syllables; Pauses; Maximum pause; Hesitations/Speech duration
Part of Speech tags	Nouns; Pronouns; Verbs; Adjectives; Adverbs; Determinants
Semantic content	Information units; Information density

3.1.7 Disease-specific associations with audio and language features

The disease stages such as MCI and AD are well known to be associated with the poverty of speech, semantically incoherent speech and slow thought process. This section provides the details of disease-specific cognitive functioning measured with audio and language features from the reviewed literature.

3.1.7.1 Cognitive functioning measured with audio features

Among 131 studies, 27 used audio features indicative of cognitive functions consisting of temporal and spectral analysis. The most distinctive speech changes in the patient group were more prolonged hesitation, slower speech, and lower periodicity, which have shown discrimination between disease groups from healthy controls [100], [148].

Studies have shown that regardless of the age of onset, pause fillers were used significantly more in those with MCI and mild AD than controls [45], [147], [159], [214]. Silent pauses were present in MCI [185], [227], and the total pause duration (in seconds) was significantly higher according to the increasing severity of the disease (moderate AD: Mean (SD) = 47.3 (21.1), mild AD: Mean (SD) = 32.6 (8.5), healthy controls: Mean (SD) = 21.7 (12.0)) [219]. The pause frequency showed a positive correlation with memory performance and grey matter density of the anterior temporal lobe indicating pause frequency as the earliest marker of AD [67].

Speech/word fluency, measured as speech rate, was reported to differentiate subjects with mild AD and control groups [105], [178]. Giles et al. (1996) found that the number of syllables per information unit produced by moderate AD group was significantly lower compared to the control group [145].

Audio features including voice segments, silence segments, periodic segments, and aperiodic segments discriminated between AD, MCI and control groups [100], [106]. Spectral analysis of speech consists of computation of the MFCC that estimates frequency to match more closely what humans hear, identifying irregularities between patients and healthy individuals that can be tracked over time [42]. Although audio features alone have shown greater classification accuracies between control and patient groups [107], [189], adding linguistic features [199], [223] and demographic information [106] have shown to improve model performance. The

voice quality measures, such as jitter, shimmer and speech fluency, have shown a significant difference between MCI and healthy controls [93]. The mean square error (MSE) surface and contour plots investigated to determine the dynamics of the speech signal suggested lower-dimensional chaotic motion in Alzheimer's speech [200]. A study comparing relationships between audio features and APOE-E4 carriers found that AD carriers had a preserved intonation range. However, a higher dose effect of APOE-E4 was associated with a higher pitch range (F0) [201].

3.1.7.2 Cognitive functioning measured with language features

The most frequently cited measure of language degradation was the semantic content of picture descriptions analysed by extracting information units. Studies used multiple terms for measuring informativeness ("semantic units", "information units", "thematic units", "micro-propositions", and "content units"), making it difficult to compare results across studies. The word-finding difficulty appeared from the earliest stage of the disease itself, showing difficulty in retrieving the names of people and places and substituting them without semantic content (e.g. thing) [42], [51]–[57]. As a result, the language of AD patients can seem "empty", lacking coherence with slower speech and shorter utterances [57], [58], [61], [132], [149], [202], [210], [246]. Along with word-finding deficits, patients from the different stages of AD have also shown reduced semantic content, incoherent responses, fewer unique words, and verbosity (too much unnecessary information) [52], [58], [60], [63], [66], [72], [80], [139], [140], [145], [150], [155], [161], [163], [166], [168], [176], [181]–[183], [186], [192], [198], [214], [220], [230], [231], [233], [234]. Studies showed a reduction of information units with the disease severity [65], [114], [172]; for example, Ripich et al. (2000) found the main effect of time (using ANCOVA) on information units as $F(2, 27) = 3.97$, $p < .05$ and in late AD participants, there was a significant decrease in Means ± SD from baseline (4.6 ± 1.2) to 6-months (3.1 ± 1.7) and 12-months (2.7 ± 1.3) assessment period. Another study showed that the control group

used more information units and semantically and syntactically more complex utterances in speech elicitation tasks than AD patients [56].

Mueller et al. (2018) found that early MCI status participants showed faster semantic deterioration than the control group, with a significant interaction between time and cognitive status ($\beta = -0.04$, $p = 0.03$) [49]. Participants with subjective cognitive decline (SCD) measured with amyloid status demonstrated a high amyloid burden associated with fewer concrete nouns (OR 6.7 [95% CI 1.2 - 37.1]) and content words (OR 6.3 [95% CI 1.0 - 38.1]) [91].

Shi et al. (2014) analysed episodic memory in story recall tasks on 249 healthy controls, 134 MCIs, and 97 AD participants. They found that both MCI (Mean ± SD = 13.91 ± 8.79) and AD (Mean ± SD = 4.02 ± 6.06) groups provide significantly fewer information units in immediate recall than healthy controls (Mean ± SD = 30.03 ± 8.58). The number further decreased in delayed recall tasks (MCI: Mean ± SD = 9.40 ± 8.88, AD: Mean ± SD = 1.60 ± 4.58) [217]. Two other studies also found reduced recall performance in story-retelling tasks [207], [224]. The content qualifier and surrogate phrases ("I suppose", "I cannot remember", and "I do not know") were found to be more likely used by the MCI group than by healthy controls [45]. Also, the participant's year of schooling and age influenced their recall performance [203], [224]. Ash et al. (2007) suggested that speech narration in AD patients comprise a multifactorial deficits, including semantic impairment impacting word-finding difficulty and limited executive function and episodic memory decline associated with incoherent speech [61]. Other studies also reported that speech incoherence is related to disease status and create communication difficulties [144], [209].

In the analyses of conversational speech from AD participants, studies showed that false starts and repetitions were present in their speech content [32], [65], [116], [130], [168], [179], [204].

Some studies referred to these disfluencies as linguistic mazes [32]. Incomplete and repeated propositions in moderate and moderate-severe cognitive decline groups were also found in Brandão et al. (2009) [158]. Part-of-speech and lexical richness statistics could not discriminate between patient groups [179], [204]. Go-ahead utterances in speech conversations ("Uh, huh" or "Aww" to make other speaker continue their speech) were more in the conversation with AD patients compared to the healthy controls [179]. In addition, unintelligible content discriminated between healthy and patient groups [188].

Several studies have found evidence of an increase in sentence fragments, reduced utterance length, relatively simple utterances, and rate of open-class words (nouns, verbs, adjectives, adverbs) [47], [50], [57], [59], [111], [131], [137], [146], [161], [190], [195], [204], [216], [222], [225], [233]. However, other studies claim that syntactic complexity was relatively preserved, at least during the early stages of the disease [58], [151], [171].

3.1.7.3 Association of speech-based features with in-clinic tests

A few studies have reported associations of speech-based features with clinical diagnostic tests, including MMSE and MoCA. The severity of the cognitive impairment measured from spontaneous speech and the MMSE score showed a significant correlation (r = 0.33 - 0.59; sample sizes: 39 - 517; disease severity: healthy controls, MCI, and AD groups) [183], [206], [229]. The proportions of sentence fragments were negatively correlated (r = -0.32, N = 118) with the MMSE score [131]. Lexical and semantic features such as type-token ratio (r = .57), mean word frequency (r = -0.34) and the total number of words (*r* = -0.36) extracted from the AD group (N = 70) were significantly associated with the MMSE score; however, no such significant relationship was found in the healthy controls [186]. In a recent study, information units (r = 0.53, N = 111), word-finding difficulty (r = -0.48, N = 111), and participants' speech

rate ($\beta = 0.158$, $p = 0.03$, N = 521) were significantly associated with MoCA score [105], [218].

Information units extracted from the AD patient's speech and their cerebral blood flow measurement in the bilateral occipital lobes ($r = 0.76$, N = 23) and left thalamus ($r = 0.71$, N = 23) were highly correlated [139]. Cuetoes et al. (2007) found that the carriers of the E280A mutation in the Presenilin 1 gene produced fewer semantic units than noncarriers, [$F(1, 38) = 12.15$, $p = .001$, N = 40] [21].

3.2 Limitations in the existing literature

3.2.1 Lacking socio-demographic diversity

The number of participants in the studies was mostly small, with 43% of studies (range (n): 5-46) <50 participants, 32% of studies (range (n): 50-96) with participants between 50-99, and 25% studies (range (n): 100-521) with >= 100 participants. The studies with participants under 50 may be useful for a proof of principle for speech-based features. Studies did not report ethnicity or racial backgrounds. Validation of new approaches across different settings (population-based or memory clinics), age, education, and ethnic groups are required to understand whether a new method works.

3.2.2 Labour-intensive manual coding and feature extraction

The manual techniques used in the studies, such as manual transcription and counting the frequency of features, are feasible for a study with a small sample size. Most studies manually or semi-automatically extracted audio and language features for each recording, which is labour-intensive and time-consuming. Only a few studies report using an ASR system to obtain transcriptions from the speech recordings (see section 3.1.5.1).

3.2.3 Speech task for collecting comprehensive language samples

The most widely used task in analysing spontaneous speech was describing a picture. These single picture-based description tasks produced a short duration of speech recording. Discourse tasks such as "Tell me about your trip to New York City" or "Describe your happiest day" provide sufficient speech content (recommended minimum of 150 words) [216]. However, these tasks would vary individual speech duration, and the speech content could be too heterogeneous as experience varies from person to person. A standard task sensitive to structured narration and providing rich content for speech analysis that could assess audio, language and episodic memory would be beneficial for future research.

3.2.4 Limited ability to aggregate results across studies

3.2.4.1 Use of multiple terms for the same measure

Most of the studies in this review used multiple terms for the same measure (for example: "semantic units", "thematic units", "micro-propositions", and "content units"), making it difficult to compare results across studies. Also, studies used other terms such as "Idea units", "words per information units", and "lexical index" for idea density and "efficiency ratio", "narrative efficiency", "discourse effectiveness", "conciseness", "conciseness ratio" for discourse efficiency, making literature search more difficult.

3.2.4.2 Use of different statistical analysis methods

Table 3.3 shows the different statistical methods used by the studies, for example, ANOVA, t-test, Chi-square test, Inter Class Correlation, Logistic Regression, and classification methods (including Support Vector Machines, Random Forest, and Decision trees). The individual studies were fairly small (75% with < 100; largest study with 521 participants) and typically diverse in their metrics and types of participants, making any form of meta-analysis not

valuable. Also, the findings obtained from participants speaking other languages should be carefully interpreted as there might be various underlying effects of language in those results, limiting the generalisation of results to studies of English speech.

3.3 Discussion

The reviewed articles have shown that a simple and non-invasive speech elicitation task can yield rich information, including lexical, semantic, syntactic, speech fluency, spectral and voice quality features.

The majority of the reviewed articles demonstrate promising results in identifying cognitive functioning in different stages of AD based on audio and language features. However, the field is heterogeneous regarding the collected data (conversation, picture/map description, story retellings, biographical narratives, and answering open-ended questions), choice of text pre-processing units (syllable, phoneme, word, or sentence) and feature extraction methods, making it difficult to compare results across studies. This literature review highlighted a need for more standardisation in feature generation, as there were variances in the feature extraction process. Moreover, depending on the study, some features were grouped into different audio and language domains. In addition, the lack of transparency in the evaluation methods used may affect the reliability and generalisability of those methods.

Machine learning and artificial intelligence technologies require large amounts of data to work effectively, and the size of the datasets available poses restrictions in using these methods. Most of the studies used a manual method of transcription, which might have been feasible due to the small sample size. Moreover, data from different patient types and populations with

diverse ethnic and cultural backgrounds may offer greater versatility in monitoring cognitive health in varied contexts.

Some reviewed articles discussed the advantage of automation compared to manual approaches. It appears that despite the potential of computational methods in the assessment of AD, several factors may be contributing to the limited utilisation of computational methods in clinical practice. One reason may be the lack of interdisciplinary collaboration between clinicians and researchers with expertise in computational methods. Another reason may be the need for more robust and reliable research to demonstrate the effectiveness and utility of computational methods in clinical settings to build a stronger evidence base. Overall, the interest in using machine learning and artificial intelligence in AD research is growing, which may lead to the more effective translation of these techniques into clinical practice.

Based on the findings from the reviewed literature, some of the evidence gaps to be addressed by this thesis are described below.

a) There is growing interest in remote data collection, requiring the feasibility and utility of the methods used. The research focus should be on the earlier stages of the disease when the unobserved changes emerge. Also, research should consider conducting a large-scale longitudinal study that includes a healthy population at a higher risk of developing AD. This thesis will address this gap using a large-scale GameChanger study (Chapter 7).
b) Regarding speech collection methods, studies mainly used picture description tasks, monologues, dialogues, and non-spontaneous speech. This thesis will use the Story Time task, which is capable of measuring audio and language characteristics, including episodic memory.

c) Most of the studies either focused on language features or audio features alone. As these speech data are collected as audio recordings, it would be possible to transform speech into text. Using features from both audio and language in the analysis could lead to better accuracy in discriminating between AD group stages [199], [223]. A standardised feature set should be developed for AD-based speech analysis and tested further to create a gold standard feature set. NLP, ASP and ML have been applied in AD research, and more studies need to leverage these technologies. A feature extraction pipeline will be developed in this thesis using NLP, ASP and ML/AI to extract features related to audio, language, and memory (Chapter 5).

d) Studies should consider including demographic information (age, education) in the analysis, along with audio features. Adding these features has shown better performance in discriminating the disease groups [106]. The effect of age and education on audio and language features will be studied in this thesis (Chapter 7).

3.4 Summary

This literature review chapter on audio and language processing for cognitive assessment summarises the speech collection methodologies used, demonstrates the use of computational methods, presents the informative speech-based features, and reports audio and language characteristics of various AD stages and control groups. The finding showed that the audio and language features are informative enough to identify problems associated with word retrieval, semantic content, and speech fluency. Based on the findings from this literature review, the recommendation for future studies is to emphasise the pre-clinical stage of AD, the use of new speech narration tasks, the use of a standardised feature set with both audio and language features, and the use of computational methods for scalability and automation. The next

chapter, Chapter 4, will introduce PREVENT and GameChanger cohorts, which subsequent chapters will use for analyses to address the research gaps listed in Section 3.3.

Chapter 4. Design and cohort characteristics of the PREVENT and GameChanger studies

This chapter provides an overview of the design and cohort characteristics of the PREVENT and GameChanger studies. Participants from these studies produced speech recordings using the Story Time task available within the Mezurio smartphone app. The subsequent chapters will show the use of the subset of PREVENT data in developing and evaluating the feature extraction pipeline (Chapter 5) and piloting computational methods in the small-sized PREVENT dataset (Chapter 6). The evaluation of these techniques was extended to GameChanger, a population-based study where the feasibility and utility of the speech-based cognitive assessment were studied (Chapter 7).

The following sections describe the study design, recruitment methods, assessments, and baseline characteristics of the PREVENT and GameChanger study.

4.1 PREVENT

4.1.1 Study design

PREVENT [247] is a prospective study that aims to identify the earliest signs of dementia by investigating interactions between risk factors for dementia and traditional biomarkers in the middle-aged population [83]. This section describes the data collection, smartphone assessments, questionnaires at baseline, and follow-up.

The study protocol was reviewed and approved by the University of Oxford Medical Science Inter-Divisional Research Ethics Committee: R48717/RE001. The study was conducted in accordance with the ethical principle's compliant with the Helsinki Declaration of 1975.

4.1.2 Recruitment methods

The PREVENT study targeted to recruit at least 150 middle-aged participants [83]. This thesis includes twenty-five volunteers recruited by the Oxford Health National Health Service Foundation Trust site (in the year 2017) who provided Story Time audio data. Volunteers were invited to join the study via the study website and social media and the 'Join Dementia Research' and ConCERT-D database, with a further invitation to participate in this ancillary study communicated via email or post [248].

Participants were screened at the time of initial consent to ensure they met the following criteria:

(a) aged 40 to 60 years,

(b) no diagnosis of dementia according to the Tenth Revision of the International Statistical Classification of Diseases and Related Health Problems criteria,

(c) no known contradiction to a magnetic resonance imaging scan.

Although a family history of dementia was not an inclusion criterion, a high proportion of participants (72%) reported a first-degree relative with dementia. All the participants provided written consent upon entry to the study.

4.1.3 Baseline assessment

Eligible participants were invited to the study sites as part of the main study [83], where they were asked to complete a questionnaire on demographics, lifestyle, socioeconomic, and health information. Participants also completed two neuropsychological assessments, a)

Computerized Assessment of Information Processing (COGNITO) [249] and b) Addenbrooke's Cognitive Examination-Revised (ACE-R) [15], to test episodic memory and general cognitive ability.

4.1.4 Smartphone-based assessments

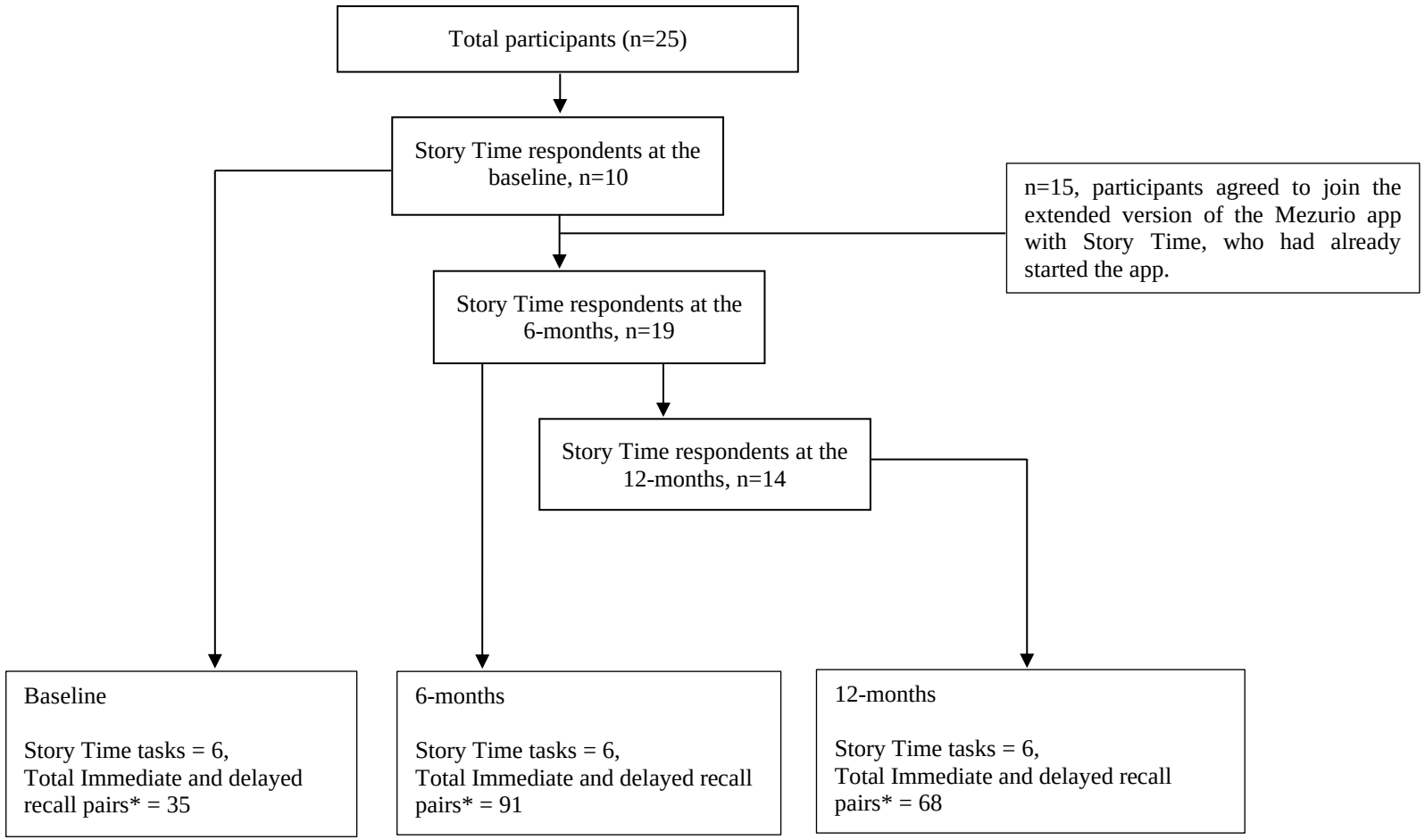
The research team gave participants an invite code and written instructions for downloading the Mezurio app to their iOS[4] or Android[5] smartphones. Loaned Android devices (n = 7) were offered to participants who did not have access to an appropriate personal device. Participants received notifications to complete the scheduled test; however, if a participant did not respond, a new notification was sent after 15 minutes, inducing variable delay between learning and delayed subtasks. The Mezurio app contained three cognitive tasks, Gallery Game, Story Time and Tilt Task, to assess the cognitive domains of episodic memory, speech and language, and executive functions. At the start of Story Time, the app checks the background noise and encourages users to move to a quiet location.

4.1.5 Repeated assessments

Participants were followed up at 6-months and 12-months. Story Time task was added to the Mezurio as an optional task on a more extended assessment schedule lasting up to 36 days. Fifteen participants who had already started the Mezurio schedule agreed to join the extended version, which was only available on their 6-months assessment. **Figure 4.1** shows the PREVENT participants, data availability, and the number of assessments.

[4] https://apps.apple.com/us/app/mezurio/id1116215871

[5] https://play.google.com/store/apps/details?id=com.oidh.biomarkers.gallery



*Recall pairs represent complete immediate and delayed recall response for a particular story from participants.

Figure 4.1 Participants flow diagram from the PREVENT study, data availability, and the number of assessments.

4.1.6 Baseline characteristics of PREVENT study participants

Overall, 25 participants (aged between 40 and 60) opted to attempt the Story Time task, of whom 84% (21) were women, and 72% (18) had a family history of dementia. There were 64% (16) participants under the age of 55 years, and the participants were primarily Caucasian, with only one participant from another ethnic category (**Table 4.1**). The overall mean (SD) age was 51.5 (5.37) years, with men aged between 42 and 59 years compared to women between 40 and 59 years. Out of 25 participants, 4 scored 100 on the ACE-R test [15], measuring global cognition. One participant scored 87, which was below the cut-off score of 88 for healthy participants. COGNITO [249] does not provide a total score like the ACE-R test. The narrative recall task from the COGNITO test battery was selected to compare with the Story Time task because this task was similar in construct with the Story Time. The overall mean (SD) of the COGNITO narrative recall task score was 15.44 (0.71).

Table 4.1 Summary of sample characteristics on the PREVENT dataset and performance (Mean ± SE) on in-clinic tests.

Characteristic	**n (%)**	**ACE-R (0-100)**	**COGNITO* (0-27)**
Overall	25 (100)	96.20 ± 3.21	15.44 ± 0.71
Demographic			
Sex			
Female	21 (84)	95.90 ± 0.75	15.14 ± 0.82
Male	4 (16)	97.75 ± 0.85	17.00 ± 0.91
Age Group (years)			
under 55	16 (64)	96.06 ± 0.92	16.13 ± 0.63
55 and above	9 (36)	96.44 ± 0.85	14.22 ± 1.63
Family History of AD			
Family history	18 (72)	96.00 ± 0.66	15.05 ± 0.96
No family history	7 (28)	96.71 ± 1.70	16.42 ± 0.61
Ethnicity			
Caucasian	24 (96)	96.45 ± 0.62	15.58 ± 0.73
Other	1 (4)	N/A	N/A
Socioeconomic			
Full-time education completed			
Under 16 years	13 (52)	96.31 ± 0.60	15.00 ± 1.22

16 years and over	12 (48)	96.08 ± 1.23	15.91 ± 0.72
Work			
Self-employed	5 (20)	94.60 ± 1.93	15.60 ± 1.03
Employed	20 (80)	96.60 ± 0.66	15.40 ± 0.86

* Narrative recall task within COGNITO; Abbreviation: AD- Alzheimer's Disease; ACE-R- Addenbrookes Cognitive Examination-Revised; COGNITO- Computerized Assessment of Information Processing.

4.1.7 Development of Manual Information Unit (MIU) scoring

Chapter 5 will discuss the development of scalable and automated feature extraction pipeline, which includes the automated information unit (AIU) scoring method. The quality of any automated approach needs to be compared with conventional methods. Psychological tests typically judge a participant's recall performance by manually listening to their response and scoring it against a set of expected "information units" [42], [52], [77].

Nicholas and Brookshire (1993) provided an Information Units (IUs) scoring method to quantify the informativeness and efficiency of speech samples [77]. This scoring method measures the quality of information an individual can provide from visual stimuli. Informativeness is calculated by comparing the descriptions provided by the individual to a list of information units (also known as semantic units, idea units, and thematic elements) that accurately convey relevant information on the eliciting stimulus, for instance: objects, actions and scenes from the pictures [42], [52], [77]. Studies have shown that AD participants produced about half the expected information units and mentioned fewer references on nouns and verbs, highlighting significant changes in semantic content and part of speech tags [46], [52]. IU scoring methods are also used in Story Retelling Procedures (SRP) in languages other than English [250], [251].

A protocol to score IUs on the Story Time data was developed where instead of listening to the recordings, text transcripts were manually read and then followed to provide a human-derived

measure of IU recall. The number of participants (n = 25) in PREVENT provided opportunity to extract such scores manually.

Manual Information Unit (MIU) scores were calculated by totalling the IU's produced within the speech sample and dividing it by the total number of Story IUs for each Story Time comic. IUs were counted only once per story by avoiding the repetition of the same IU, as suggested by McNeil et al., 2001 [252].

4.1.7.1 Protocol for scoring Manual Information Units (MIU)

Nicholas and Brookshire (1993) have provided rules for counting words and correct information units from the picture description task. The definition of information units, as stated by Nicholas and Brookshire (1993) is “the words that are intelligible in context, accurate in relation to the picture, and relevant to and informative about the content of the pictures” [77]. The predefined story IUs for one of the Story Time comics are provided in **Table 4.2**.

The following rules were followed in the Story Time task context to get MIU scores.

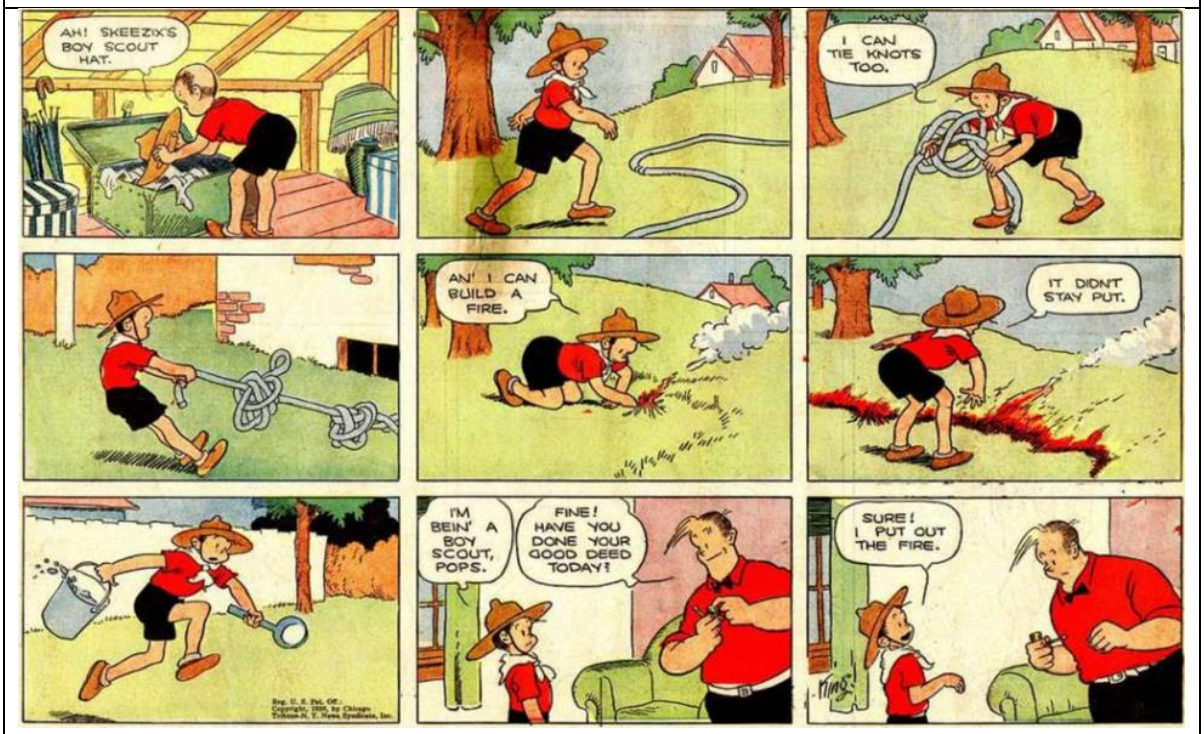
1. Remove the statements or phrases that the participants say before and after the task, for example: "Ok, let’s start today's story", "I think this is it", "This is the end of the story", etc. These statements do not provide any information on the Story Time comic itself but provide their readiness to begin or have completed the task.
2. Remove the statements that provide commentary to the task and lead-in phrase, for example: "In this first picture", "I cannot think of the name of that", or "No, that's unfair".
3. Remove dysfluencies and filler words such as um, er, uh, you know, well, I mean, okay, wow, etc.
4. Count all words that are nouns, adjectives, verbs, and adverbs and also intelligible, accurate, relevant, and informative in the context of the Story Time comic.

5. Do not count non-specific words or phrases, e.g., things, stuff, something, something like that, and so on.
6. Ignore grammar unless it leads to an incorrect interpretation of the story. However, if grammatically incorrect words cause confusion or misunderstanding, do not count them. Example of incorrect information unit: The boys and girls are eating ice lollies. (The picture shows only a boy and a girl). The boy walked away. (The picture shows a boy and his father walking).
7. Count the word with a regional reference. Example: "We will have it for the tea (dinner) tomorrow". Although 'tea' is usually a hot drink, here 'tea' is referred to as a cooked evening meal (British English).
8. Do not count if the participant is saying the word that represents an expected information unit but relates to their experience or performance, for example: "I was in a hurry to go to my mom, so I do not remember anything.", "I cannot walk that fast", "I remember my father used to ask me to go and catch a hen like in this picture" etc.
9. Count even if the incomplete sentence contains some information about the comic, for example, "The mother was. The red hen flew away.", contains incomplete response "the mother was" and includes a valid information unit - mother.
10. Words expressing legitimate uncertainty are counted as correct information units; for example, A man or maybe a neighbour asked what the dog's weight was.
11. Count the words that result in another correct English word and sound similar to the predefined information units (this might be due to the transcription error); for example, His father is sitting on a school (stool) and is very tired.

The total IU count is divided by the expected number of story IUs for that comic to get an MIU score.

Table 4.2 Example of a Story Time comic with the list of expected information units.

Story Time comic



List of expected information units

1. corky	16. fire
2. attic, loft	17. ground, field, yard
3. umbrella	18. stay-put, stop
4. lamp	19. bring
5. find, search	20. bucket
6. scout	21. water
7. hat	22. ladle, dipper, saucepan, scoop
8. box, trunk	23. sofa
9. wear	24. boy scout
10. neckerchief, necktie	25. father, pops
11. tie	26. ask
12. knot	27. deed, good-deed
13. rope	28. today
14. pull	29. put-out, extinguish, douse
15. build	

4.2 GameChanger

4.2.1 Study design

GameChanger [84] is a citizen science project launched in September 2018. This project, targeted at residents in the United Kingdom, uses a smartphone application named Mezurio containing various cognitive tasks designed to provide information on how the brain works and changes over time. Healthy participants aged eighteen years and over can sign up on the GameChanger website, after which they receive an email with a unique participant code that enables use of the Mezurio smartphone app, which is available to download from both Apple and Google app stores.

The GameChanger study was ethically approved by the University of Oxford Medical Sciences Inter-Divisional Research Ethics Committee (R58202/RE001) and complied with the Helsinki Declaration of 1975.

4.2.2 Recruitment methods

The volunteers were recruited to join this study via different routes, including the Alzheimer's Society's website, the Join Dementia Research database, email newsletter, social media, and the press, including local, regional, and national. The study was endorsed by celebrities and was promoted by the Alzheimer's Society. Volunteers were not compensated by any means for joining the study.

Participants had to register for the study on the GameChanger study webpage[6]. The webpage includes sections with important study information such as inclusion criteria, withdrawal from the study, and personal data privacy. The following were the inclusion criteria for taking part in the study:

a) having an iOS or Android smartphone

b) aged 18 years or older

c) not having a known diagnosis of dementia

As part of the registration process, participants had to complete two mandatory e-consent forms to ensure that they understood the information on the study and that they were happy for the Mezurio app to record their voice. In addition, four optional consent statements were presented relating to future linkage with other UK-based research studies and study communications about new tasks or blog posts. Following the registration, the participant's email address recorded during the registration process was used to send them an email containing an invite code for launching their month of GameChanger activity and a link to Mezurio in the App or Google Play store for downloading the app.

4.2.3 Demographics and other questionnaires

Basic demographic information collected at the beginning of the GameChanger schedule includes a) date of birth, b) gender, c) postcode, 4) ethnicity and 5) education levels. In addition, participants were asked to report whether they had an immediate family history of dementia and were diagnosed with psychiatric (e.g., depression or anxiety) or neurological conditions (e.g., stroke). Participants were allowed to skip any individual self-report items. Residential

[6] https://www.joingamechanger.org/

postcodes of participants were used to derive participant's relative deprivation based on Lower Layer Super Output Areas (small areas of around 1500 residents), which was defined by the Index of Multiple Deprivation (IMD) deciles [253]. IMD combines information from seven domains: employment, income, crime, health, education levels, availability of services and local environment.

The Mezurio app contains a subjective cognitive decline (SCD) questionnaire, a self-report measure designed to capture perceived cognitive decline within the past 24 hours and the past six months. The questions are based on difficulty finding the right words, forgetting the date, difficulty in recalling the name of a familiar person or object, difficulty in navigating familiar streets, losing the train of thought, and listening difficulties. **Table 4.3** lists all the questions for the daily and yearly self-report questionnaires.

Table 4.3 Subjective Cognitive Decline Questionnaire deployed within Mezurio in Game Changer study.

Daily Subjective Cognitive Decline Questionnaire Everyone experiences small 'slips' in cognition from time to time. Please tell us whether you have experienced the following events over the last 24 hours.		
Category	**Question**	**Options**
Words	In the last 24 hours, have you had difficulty finding the right words? E.g., feeling like a word is on the tip of your tongue.	Not at all, Once, A couple of times, Frequently
Date and Time	In the last 24 hours, have you forgotten what day of the week or month it is?	Not at all, Once, A couple of times, Frequently
Names	In the last 24 hours, have you had trouble recalling the name of a familiar person or object?	Not at all, Once, A couple of times, Frequently
Places	In the last 24 hours, have you had trouble navigating familiar streets?	Not at all, Once, A couple of times, Frequently

Thinking	In the last 24 hours, have you lost your train of thought?	Not at all, Once, A couple of times, Frequently
Listening	In the last 24 hours, have you had trouble understanding a conversation or simple instructions? Please exclude occasions where you couldn't hear clearly.	Not at all, Once, A couple of times, Frequently

Yearly Subjective Cognitive Decline Questionnaire

We all experience small 'slips' in our thinking abilities from time to time. We want to know how often these things have happened to you in the last 6 months?

Category	**Question**	**Options**
Words	Difficulty finding the right words. E.g., feeling like a word is on the tip of your tongue.	Never, Very rarely, Occasionally, Often, Very Often
Date and Time	Forgetting what day of the week or month it is.	Never, Very rarely, Occasionally, Often, Very Often
Names	Trouble recalling the name of a familiar person or object	Never, Very rarely, Occasionally, Often, Very Often
Places	Trouble navigating familiar streets	Never, Very rarely, Occasionally, Often, Very Often
Thinking	Losing your train of thought	Never, Very rarely, Occasionally, Often, Very Often
Listening	Trouble following a conversation or simple instructions	Never, Very rarely, Occasionally, Often, Very Often

4.2.4 Smartphone-based assessments

Participants enrolled in this study were asked to complete cognitive tests deployed within the Mezurio app, which can be completed remotely using the participant's own mobile device. The app consists of the tasks mentioned in section 4.1.4, with new questionnaires and additional cognitive tasks added for the GameChanger study. **Table 4.4** provides the assessments and cognitive measures from each of those assessments.

Table 4.4 List of assessments deployed within the Mezurio smartphone app for the GameChanger study.

Assessments	Cognitive measure
Story Time task	Connected language and episodic memory
Gallery Game	Measure of episodic memory
Flashcards	Measure of response time
Tilt task	Measure of executive function
Gallery Blast	Measure of memory

Note: Assessments other than Story Time and Gallery Game are not considered further in this thesis.

In this thesis, the outcome from the Gallery Game, which has been validated in a small PREVENT cohort [248], will be used in examining its relationship with the Story Time measure. Gallery Game has a repeated measure design where participants are asked to complete multiple learning tasks followed by a recognition and recall test (**Figure 4.2**). This task uses a set of professionally photographed image stimuli which are then used in creating confusable target-distractor pairs for the recognition test.

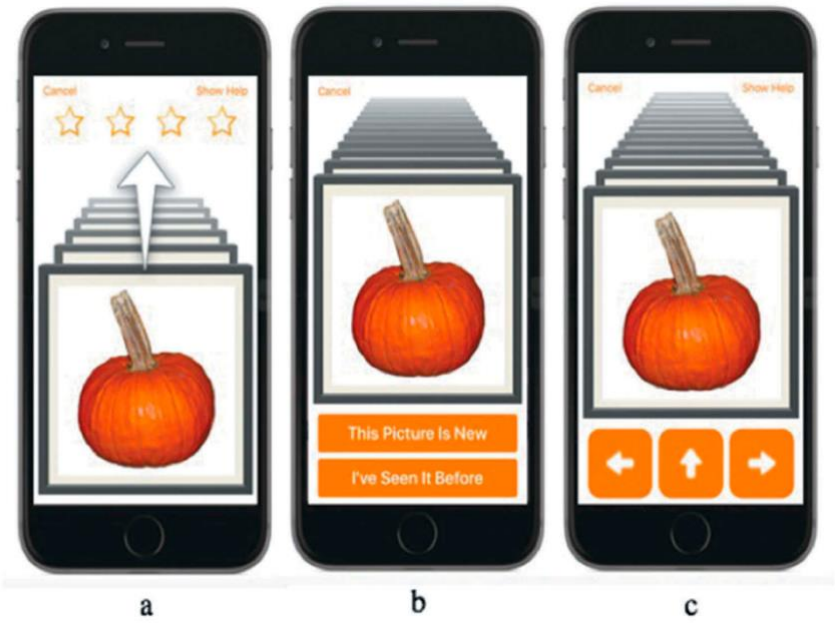


Figure 4.2 Representation of (a) learning, (b) recognition, and (c) recall (swipe direction: left, up, and right) in Gallery Game task within the Mezurio app. This image is reproduced from [248] with the permission from the author.

4.2.5 Repeated assessment

Participants who consented or participated in round 1 (Year-1) were followed-up for round 2 (Year-2). Participants can sign up for the GameChanger study at any time, even between assessment waves. However, due to COVID-19-related reasons, participants who joined the GameChanger study in late 2018 and early 2019 were followed up only in 2022 to participate in round 3 (Year-3). **Figure 4.3** shows the participant's flow diagram for the GameChanger study, data availability, and the number of Story Time assessments.

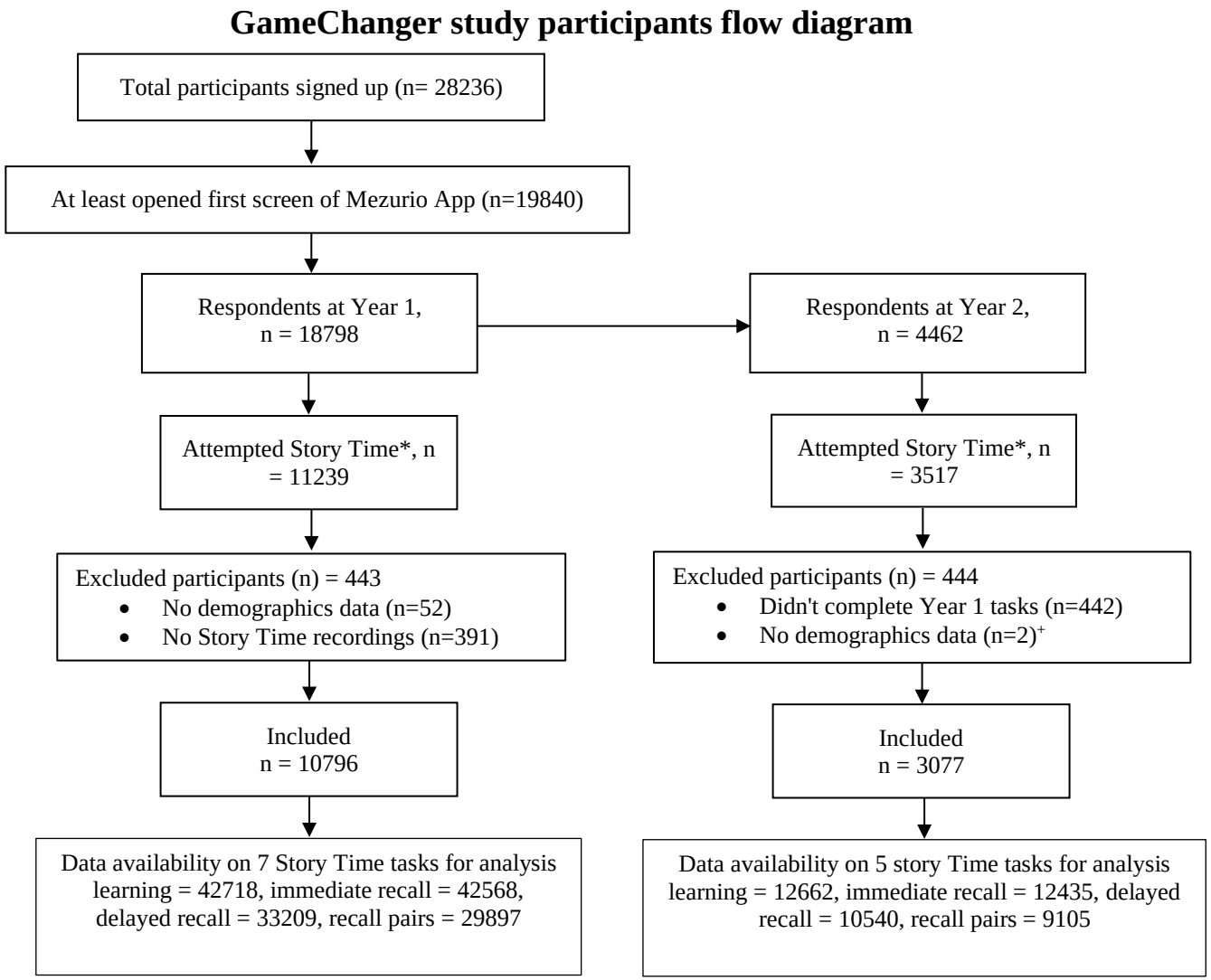


Figure 4.3 Flow diagram of participation in the GameChanger study and data availability.
* These participants were used for compliance and Story Time attempts analysis; + Out of 52 participants who did not report demographics data in Year 1, only 2 of them continued in Year 2; All the participants in Year 2 provided at least one Story Time recording. As discussed in section 1.3.2, in the learning sub-task, participants narrate the story looking at the comic strips presented to them, followed by immediate and delayed recall where they are asked to recall the story from their memory immediately and after 24-hr delay.

4.2.6 Baseline characteristics of GameChanger participants

Overall, 18798 participants responded to the GameChanger study, although 1042 participants who opened the app did not proceed further (**Figure 4.3**). There was an overrepresentation of female participants, with 77.5% of the participants being female (**Table 4.5**). This gender representation bias was also found in other citizen science studies [254] and the literature review (Chapter 2), showing that female participants are more likely to participate in the observational study than men. Some participants (n = 467) skipped the birth date, adding a default value for the year of birth (1990-01-01) in the demographic's questionnaire response. After excluding those default date of birth values, the overall mean (SD) age of the participants was 49.7 (14.4) years, where men were slightly older than women. 19.5% of the participants did not provide their postcode, the highest among the missing socio-demographic characteristics. The majority of participants were Caucasians (92.3%), followed by other ethnic groups (3.7%) thus lacking ethnic diversity. A higher number of participants (24.7%) joined the study from the postcodes with the highest deprivation decile indicative of a better standard of living. The GameChanger version-1 demographic questionnaire was missing the post-graduate category under education due to a technical error (fixed in the version-2), which might have led to a higher number of participants (34.2%) with undergraduate studies as the highest level of education in Year-1. Of 18798 participants, 11239 (60%) made it to the Story Time task. These Story Time attempters also had similar socio-demographic characteristics (**Table 4.5**).

Table 4.5 Sociodemographic characteristics of those taking part in GameChanger and those who attempted the Story Time task.

		GameChanger Participants, n (%)	Story Time Attempters, n (%)
Characteristics		18710* (100)	11239 (100)
Gender	Male	4120 (22.0)	2206 (19.6)
	Female	14506 (77.5)	8973 (79.8)
	Other	9 (0.0)	2 (0.0)
	Missing data	75 (0.4)	58 (0.5)
Age (years)	Mean (SD)	49.7 (14.43)	53.4 (13.09)
Family History	No	10135 (54.2)	5676 (50.5)
	Yes	7791 (41.6)	5109 (45.5)
	Not Sure	685 (3.7)	387 (3.4)
	Missing data	99 (0.5)	67 (0.6)
Education	Primary	47 (0.3)	22 (0.2)
	Secondary	2571 (13.7)	1548 (13.8)
	College	3492 (18.7)	1954 (17.4)
	Trade or training	2476 (13.2)	1471 (13.1)
	Undergraduate	6393 (34.2)	3854 (34.3)
	Graduate	3600 (19.2)	2310 (20.6)
	None of the above	6 (0.0)	4 (0.02)
	Missing data	125 (0.7)	76 (0.7)
Ethnicity	Caucasian	17261 (92.3)	10496 (93.4)
	Hispanic	50 (0.3)	26 (0.2)
	Black	86 (0.5)	37 (0.3)
	Indian Subcontinent	153 (0.8)	76 (0.7)
	Asian other	143 (0.8)	63 (0.6)
	Other	680 (3.6)	347 (3.1)
	Mixed	111 (0.6)	67 (0.6)
	Missing data	226 (1.2)	127 (1.1)
Deprivation Decile of England	1 & 2	1405 (7.5)	731 (6.5)
	3 & 4	2328 (12.4)	1300 (11.6)
	5 & 6	3050 (16.3)	1804 (16.1)
	7 & 8	3654 (19.5)	2350 (20.9)
	9 &10	4616 (24.7)	3067 (27.3)
	No data+	3657 (19.5)	1987 (17.7)
Condition Neurology	No	18191 (97.2)	10915 (97.1)
	Yes	419 (2.2)	256 (2.3)
	Missing data	100 (0.5)	68 (0.6)
Condition Psychology	No	13473 (72.0)	8504 (75.7)
	Yes	5095 (27.2)	2637 (23.5)
	Missing data	142 (0.8)	98 (0.9)

Data are expressed as numbers (%), values are mean (standard deviation); * 88 participants missing all the demographic information; + Contains missing data plus non-English residents.

Table 4.6 shows the sociodemographic characteristics of the participants who completed follow-up for GameChanger Year-2. The mean age (SD) was 54.98 (12.16), and 79.5% of females followed up for the Year-2 schedule. A higher proportion of participants who provided "yes" to the response on a family history of dementia question followed up compared to "no" and "not sure" on the follow-up Year-2. A higher proportion of participants with higher education (undergraduate and graduate) and the participants in the higher deprivation decile group continued in the follow-up study. The proportional participation on all the other characteristics remained similar in both Year-1 and Year-2.

Table 4.6 Sociodemographic characteristics of those taking part in year 1 and returning to year 2 of the GameChanger study.

Characteristics		**Participants**	
		Year-1, n (%)	**Follow up in Year-2, n (%)**
		(n = 18710*)	(n = 4450)
Gender	Male	4120 (22.0)	908 (20.4)
	Female	14506 (77.5)	3536 (79.5)
	Other	9 (0.0)	1 (0.0)
	Missing data	75 (0.4)	7 (0.2)
Age (years)	Mean (SD)	49.7 (14.43)	54.9 (12.16)
Family History	No	10135 (54.2)	2170 (48.8)
	Yes	7791 (41.6)	2126 (47.8)
	Not Sure	685 (3.7)	151 (3.4)
	Missing data	99 (0.5)	3 (0.1)
Education	Primary	47 (0.3)	5 (0.1)
	Secondary	2571 (13.7)	542 (12.2)
	College	3492 (18.7)	692 (15.6)
	Trade or training	2476 (13.2)	520 (11.7)
	Undergraduate	6393 (34.2)	1622 (36.4)
	Graduate	3600 (19.2)	1058 (23.8)
	None of the above	6 (0.0)	1 (0.0)
	Missing data	125 (0.7)	12 (0.3)
Ethnicity	Caucasian	17261 (92.3)	4245 (95.4)
	Hispanic	50 (0.3)	10 (0.2)
	Black	86 (0.5)	12 (0.3)
	Indian Subcontinent	153 (0.8)	26 (0.6)
	Asian other	143 (0.8)	14 (0.3)
	Other	680 (3.6)	95 (2.1)
	Mixed	111 (0.6)	32 (0.7)

Characteristics		Participants	
		Year-1, n (%)	Follow up in Year-2, n (%)
	Missing data	226 (1.2)	16 (0.4)
Deprivation Decile of England	1 & 2	1405 (7.5)	241 (5.4)
	3 & 4	2328 (12.4)	448 (10.1)
	5 & 6	3050 (16.3)	750 (16.9)
	7 & 8	3654 (19.5)	965 (21.7)
	9 &10	4616 (24.7)	1340 (30.1)
	No data+	3657 (19.5)	706 (15.9)
Condition Neurology	No	18191 (97.2)	4351 (97.8)
	Yes	419 (2.2)	92 (2.1)
	Missing data	100 (0.5)	7 (0.2)
Condition Psychology	No	13473 (72.0)	3447 (77.8)
	Yes	5095 (27.2)	985 (22.1)
	Missing data	142 (0.8)	18 (0.4)

Data are expressed as numbers (%), values are mean (standard deviation); * 88 participants missing all the demographic information; + Contains missing data plus non-English residents.

4.3 Summary

This chapter introduced two cohorts that were used in this doctoral research. PREVENT being small makes it feasible for manual transcription and scoring information units manually, both of which are essential for comparing with automated methods for scalability. PREVENT also provides neuropsychological in-clinic tests allowing investigation of the consistency of automated measures with clinical parameters. GameChanger is a large dataset with a wide range of age groups and education, a family history of dementia, deprivation decile indexes, and previous psychological and neurological conditions, providing an opportunity to explore the utility of speech-based cognitive assessment by studying the distribution of audio and language features in different socio-demographic groups and investigating if there is a relationship between Story Time outcomes and well-established socio-demographic variables associated with dementia risk. The next chapter, Chapter 5, will present a feature extraction pipeline to extract speech-based features from speech recordings provided by these cohorts for AD research.

Chapter 5. Developing a feature extraction pipeline to derive audio and language features from raw speech data

5.1 Introduction

Manual extraction of audio and language features can be time-consuming and labour-intensive, especially for larger datasets. It also requires expertise in audio and language processing techniques. Although manual methods have been used in various studies (Chapter 3), they are only applicable for analysing data with a small sample size.

For large-scale studies, there is a need to automate the feature extraction process without requiring manual intervention, including the automatic speech transcription reducing the possibility of human error and expenditure of resources (time and money). Developing a feature extraction pipeline would ensure that the method is consistent and repeatable, which is essential for the accuracy and reliability of the methods used. Also, the pipeline should be scalable, making it possible to run on large amounts of data.

Developing a feature extraction pipeline involves various steps. First, features that are most relevant for the early detection of AD need to be identified. Second, existing tools and algorithms for data processing need to be selected. Third, the set up and configuration of existing tools must be chosen. Fourth, the performance of tools and algorithms should be evaluated, where necessary. Finally, new algorithms should be developed when the existing ones are not satisfactory and code should be written to integrate algorithms into a complete pipeline to provide automation, consistency and scalability.

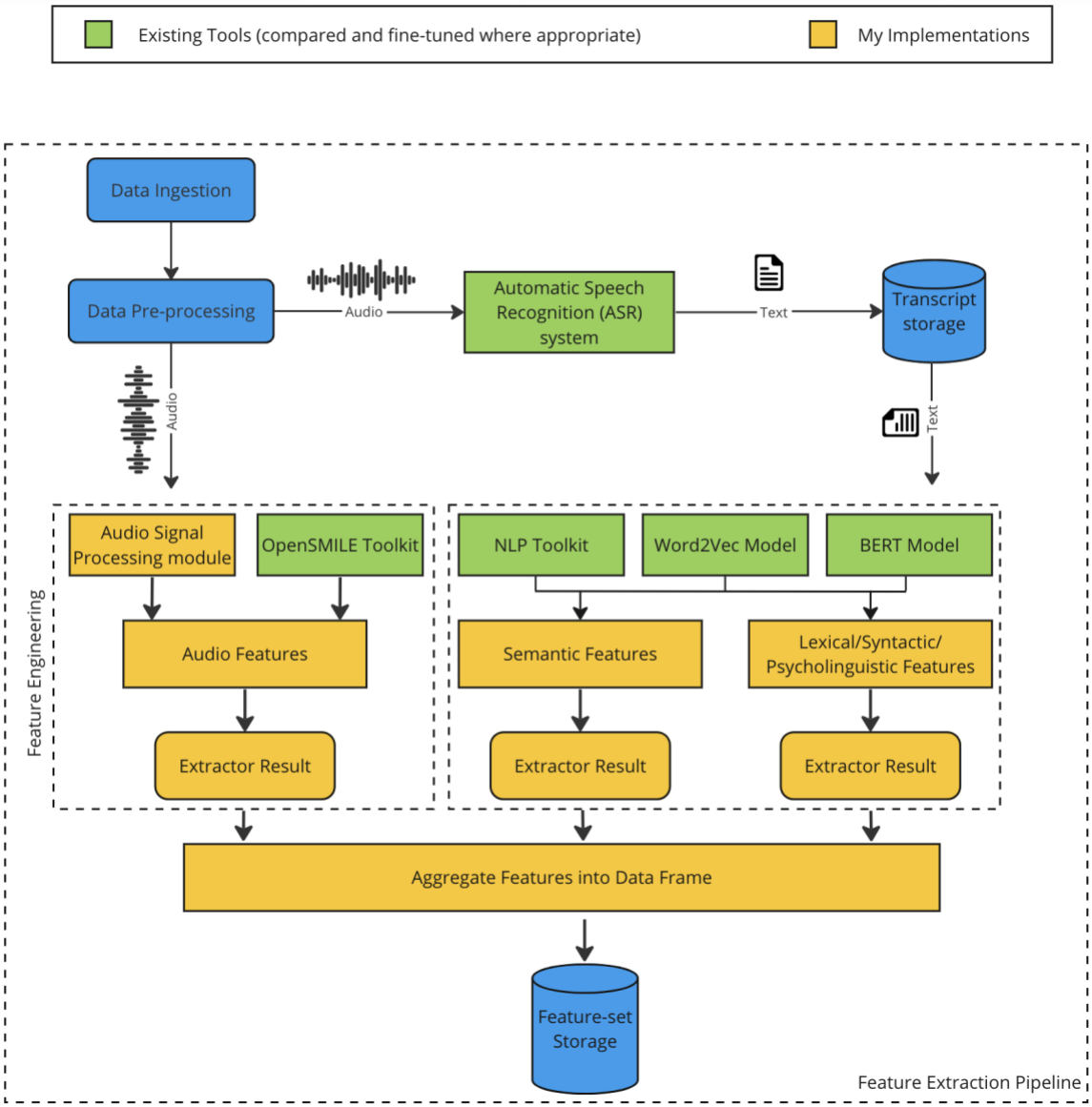


Figure 5.1 Flowchart of feature extraction pipeline developed for this doctoral research.

Abbreviation: NLP - Natural Language Processing; BERT- Bidirectional Encoder Representations from Transformers

This chapter introduces a new efficient and scalable feature extraction pipeline developed in this thesis to extract various audio and language (from text) features. This feature extraction pipeline was developed in Python because it has a large number of open-source libraries and tools which provide a broad range of functionalities for various tasks in data processing/analysis, machine learning, and deep learning. **Figure 5.1** represents the flowchart of the feature extraction pipeline, which converts any speech recording into a set of audio and language features.

The main components of the pipeline are Automatic Speech Recognition (ASR) system, Natural Language Processing (NLP) Toolkit, Word2Vec model, Bidirectional Encoder Representations from Transformers (BERT) model, Audio Signal Processing (ASP) module and OpenSMILE Toolkit.

Speech recognition is the process of converting spoken words into written or computer-readable text. The ASR system is responsible for this task. The quality of the output from the ASR system plays a key role in determining the quality of the language features derived from the speech. If the speech recognition system produces inaccurate transcriptions, the resulting text-based language features may also be inaccurate or misleading. In order to obtain high-quality transcripts, it is often necessary to compare and evaluate different ASR systems to determine which outputs the most accurate transcript compared to manual transcription in a similar task. Various commercial and non-commercial ASR systems are available, and each may have different strengths and weaknesses depending on the application's specific requirements. In this thesis, various ASR systems were systematically compared based on the accuracy of the transcripts with fewest errors (section 5.2).

The NLP toolkit is a set of tools and algorithms used to analyse and understand human language. These tools can be used to extract relevant language features by pre-processing the text into part-of-speech tags. In this thesis, text inputs are pre-processed using the NLP toolkit.

The Word2Vec model represents words in a vector space, with each word represented by a dense vector of numbers [255]–[257]. This representation allows capture of relationships between words and their meanings, which can be useful for scoring the story recall performance. Most Word2Vec models are trained on a general corpus dataset, including

Wikipedia and Google News. Therefore, to obtain the best results, it may be necessary to fine-tune or retrofit the model on a specific dataset to ensure that it is well-suited for the task at hand. This thesis uses the expected relationship between the words based on the Story Time task as a transfer learning approach to adapt the model (section 5.3).

The BERT model is a type of deep learning model that has been widely used for natural language processing tasks [124]. It processes input text and generates a representation of the text that can be used for downstream tasks such as text classification, question answering, and Named Entity Recognition. In this pipeline, BERT is used for processing the text to tag surrogate phrases. As discussed in the previous chapters, surrogate phrases[7] are found in story recall responses. For information unit scoring, these qualifiers or phrases do not convey important meaning on their own; however, they provide cues on forgetting and doubts during the narrative recall tasks. The BERT model is trained on a dataset of labelled surrogate phrases. The trained model is then added to the feature extraction pipeline to detect such phrases contained in the speech (section 5.4)

ASP involves the analysis and manipulation of audio signals, those produced by speech. In this thesis, the ASP module was developed to process audio to produce speech features using several pre-existing Python libraries to analyse and manipulate audio signals, including scipy [258], librosa [259], and pyAudioAnalysis [260]. The literature review (Chapter 3) showed that studies used PRAAT [261], an open-source software program to analyse and manipulate speech and voice signals. It is reported to be used semi-automatically in extracting speech

[7] In this thesis, the surrogate phrase includes content qualifiers ("I think", "I suppose", etc.) and memory words ("I can't remember", "I forgot", etc.) to avoid using multiple terms.

features such as pitch, voiced segments, formants, and intensity [96], [102], [227]. PRAAT was not directly applicable for use in the scalable pipeline as its application programming interface requires knowledge of the PRAAT scripting language. Therefore, the ASP module was developed in Python as a part of this feature extraction pipeline to integrate speech analysis with other data processing tasks for large-scale analysis. The OpenSMILE toolkit [243], a collection of tools and libraries for extracting features from audio data, was also integrated into the pipeline to extract spectral and voice quality features from audio recordings. The audio data was processed into chunks to avoid insufficient memory issues (commonly known as "out of memory" among software developers) arising when simultaneously processing a large dataset.

In the remaining sections of this chapter, each of the components of the feature extraction pipeline and the overall performance of the pipeline will be discussed in detail.

5.2 A systematic comparison of ASR services and evaluation of their transcription results on a subset of the PREVENT dataset

ASR systems have improved significantly over recent years [262], and they are now widely used in a variety of applications, including commercial services like Amazon Alexa and Google Home, as well as in medical research [263], [264].

Spontaneous speech is often more challenging for ASR systems to transcribe accurately than read-aloud speech. This is because spontaneous speech tends to be more unstructured and less predictable and may contain a greater variety of disfluencies, hesitations, and other non-lexical elements. In addition, spontaneous speech often has incomplete words and a higher proportion of out-of-vocabulary words, adding challenge to the ASR performance [265]. Studies in the literature review (Chapter 3) also reported other speech intelligibility issues, such as

background noise, speaker's articulation, and regional dialects, can make it more difficult for the ASR system to recognise and transcribe words correctly [209], [237], [266].

Here, the aim was to explore evidence on the performance of different ASR system providers on their base and optimised models to evaluate their quality of automated transcripts against the time-consuming manual approach and the data security and privacy of the audio data in using their services. A systematic comparison of six commercially or freely available ASR systems was conducted.

5.2.1 Automatic Speech Recognisers

The ASR systems that were compared are Amazon Transcribe, Deep Speech, Google Speech to Text, Watson Speech to Text, Kaldi, and Microsoft Azure Speech to Text. Python scripts were developed to make the appropriate Application Programming Interface call to ASR systems to test their performance.

Amazon Transcribe

Amazon ASR service is available under the name Amazon Transcribe [267]. Amazon provides pre-trained models trained in various languages. They have recently added the capability of recognising UK-based regional accents. Amazon also gives the ability to improve the transcription service by providing custom vocabulary (list of phrases) and training its custom language model.

DeepSpeech

DeepSpeech is an open-source ASR system developed using an end-to-end deep learning technique developed by Mozilla [268]. ASR model (version 0.9.3) available in their GitHub

code repository[8] was used for transcribing speech recordings. Their code repository also provides scripts for adapting the language model using domain-specific texts and adding domain-specific words.

Google Cloud speech to text

Google provides ASR service under the name Cloud Speech to Text (referred to as Google Cloud hereafter), a part of the Google Cloud platform [269]. They provide Software Development Kit (SDK) and pre-trained ASR models for both narrowband (audio recordings with 8 kHz) and wideband (audio recordings greater than 16 kHz). A wideband ASR model was used as the smartphone records higher frequency recordings. Google also provides a speech adaptation technique where users can list words and phrases. Tokens specific for the phone number, house number, month, year, and others that commonly occur in the user's use case can also be included in that list.

IBM Watson speech to text

IBM provides ASR service in the name of Watson Speech to Text (referred to as IBM Watson hereafter) [270]. Registering an account in IBM Cloud allows access to the transcription service. Like Amazon, IBM also gives speech adaptation techniques to list phrases or train their language model to improve their transcriptions.

Kaldi

Kaldi is popular among audio signal processing researchers and was developed as an open-source ASR toolkit to support speech recognition research [236]. It allows training an ASR

[8] https://github.com/mozilla/DeepSpeech

model from scratch with audio data and human transcription. The British Broadcasting Corporation (BBC) also uses Kaldi for transcription [271]. A request to the BBC was made to share their ASR model[9]; however, the request was declined stating that their model access is only available to BBC partners. Therefore, a pre-trained off-the-shelf model developed by Zamia-speech[10] was used to generate transcripts from speech recordings. Zamia-speech trained this model on voxforge, librispeech, cv-corpus, ljspeech, and m_ailabs_en corpus with additional 400 hours of noise-augmented audio. Zamia-speech also provides a Kaldi-adapt-lm library, allowing language model adaptation with in-domain text.

Microsoft Azure speech to text

Microsoft's ASR service is named Azure cognitive services: speech-to-text (referred to as Microsoft Azure hereafter) available within the Azure cognitive services [272]. After setting up an account in Azure, their service can be used. Microsoft provides three model adaptation techniques: a) providing domain-specific phrases, b) training a custom language model, and c) adapting both acoustic and language models (definitions in section 5.2.2) by providing human transcriptions and audio files.

5.2.2 ASR optimisation techniques

ASR systems contain two models: the acoustic and language models. The acoustic model is a statistical model used to convert the sequence of speech sounds into a sequence of linguist units (phonemes or words). It is trained on a large dataset of speech sounds and the corresponding transcriptions of those sounds, and it learns to recognise patterns in the speech signal that are

[9] I sent an email request to the BBC MakerBox tool (makerbox@bbc.co.uk).

[10] https://github.com/gooofy/zamia-speech

indicative of specific words or sub-words. On the other hand, the language model is a statistical model that is used to predict the probability of a sequence of words in a given context.

Most ASR systems allow for the use of custom vocabulary lists, building custom language models, and adapting acoustic models to improve their performance (**Table 5.1**). Although ASR systems use different terms, such as custom vocabulary, keywords, phrases, and tokens, they all use optimising techniques using vocabulary lists.

Table 5.1 Optimisation methods provided by ASR systems.

ASR services	**Optimisation methods**
Google Cloud	Speech adaptation by providing a list of - Single words - Multiple words - Class tokens (specific for a phone number, house number, month, year, and others)
Amazon Transcribe	Speech adaptation with - Custom vocabulary (list of phrases) - Custom language model (train a language model by providing domain specific texts)
IBM Watson	Speech adaption with - Keywords - Custom language model (train a language model by providing domain specific texts)
Microsoft Azure	Allow model adaptation of their base model with - Domain specific texts - Phrases - Audio recordings and their manual transcripts
Deep Speech	Option to train own model with audio recordings and their manual transcripts - Option to adapt their language model with domain specific texts
Kaldi	- Option to train own model with audio recordings and their manual transcripts - Option to adapt the language model with domain specific texts

The ASR optimisation methods are as follows:

a) Providing vocabulary/phrases as input to the base model

A vocabulary list contains words the ASR system is expected to recognise and transcribe. By providing a custom vocabulary list, the ASR system can be tuned to recognise and transcribe those specific words and phrases from the list better and making the transcription output more relevant to the task. Amazon Transcribe requires a text file containing custom vocabulary as a list or a table. All the ASR systems that were compared accepted words/phrases via their Application Programming Interface call, except Kaldi. Google and IBM Watson ask users to provide a threshold or boost value for each word/phrase. This value increases the probability that a specific phrase will be recognised over other similar-sounding phrases.

b) Building a custom language model

Text data from specific use cases need to be provided to the base model to train the custom language model. Building a custom language model is initiated by the commercial ASR system after the text data is provided on their transcription portal. These systems did not provide detailed documentation on how they train the language model. However, in Kaldi and DeepSpeech, the following steps must be completed, which are:

I. Add domain-specific texts: The base language model is created using a large number of sentences containing conversation, newspaper reports, online text, and other text resources. Domain-specific sentences (for example: possible stories (in text format) that would be produced from the Story Time task) need to be added to this set of generic sentences as a first step.

II. Limit the language model to the vocabulary that the ASR model supports (optional): This step ensures that the text that will be used to train the custom language model includes only the words and phrases that the acoustic model is able to recognise.

This can help to ensure that the language model is consistent with the vocabulary of the acoustic model and can improve the overall performance of the ASR system.

III. Train the custom language model: Using the data from 1) or 2), the custom language model is trained to estimate the statistical relationships to create a model that can predict the probability of a sequence of words given the context in which they are used.

IV. Adapt the base model using the new custom language model: This step involves fine-tuning the parameters of the base model using the custom language model to improve the model's ability to recognise and transcribe speech in the specific domain for which the custom language model was created.

c) Adapting an acoustic model

Microsoft Azure (among commercial ASR systems) and both open-source ASR systems (DeepSpeech and Kaldi) provided an option to adapt the acoustic model. This adaptation process modifies the parameter of a pre-trained acoustic model to better suit the characteristics of a specific speaker, recording environment, or language. It requires audio recording with scenarios like background noise, accented speech, and a diverse audio quality covering the full scope of speech and their accurate transcriptions as inputs. This optimisation method was not explored in this thesis as it requires a large number of audio recordings with manual transcriptions.

5.2.3 Methods

As mentioned in the previous section, four commercial and two open-source ASR systems were selected for this comparison. Audio recordings from PREVENT were available in two file formats, MP4 (iPhone) and 3GPP (Android). The ASR systems used for the comparison

did not support these file formats. Therefore, files were converted to WAV (Linear PCM format) with a sample rate of 16 kHz using a free and open-source FFmpeg tool[11].

5.2.3.1 Manual transcription of the recordings

Audacity [273], an open-source audio software, was used to play and listen to the audio recordings and manually record information. CHAT transcription guidelines [274] are widely used in spoken language research in dementia, traumatic brain injury, right hemisphere damage, Aphasia, and Autism. The CHAT guidelines[12] available on the talkbank.org website were adapted so that the manual transcripts would be easily compared with the ASR system's transcription outputs without further processing, and the procedures followed were to:

a. be accurate, capture every word and do not change any word

b. write speech contractions (don't, can't, it's) and do not modify

c. include every detail (false starts, fillers, disfluencies)

d. use proper spelling (British English) for the word or phrase

e. represent numbers in words (e.g., 10 as ten)

f. discard external sounds

In total, 25 audio recordings from the PREVENT dataset, including five recordings from each speaker, averaging approximately one minute and thirty seconds in duration (i.e., about 40 minutes in total) and containing an average of 176 (SD = 98) words, were manually transcribed. These manual transcripts were used as a gold standard (test text set) for ASR model comparison.

[11] https://www.ffmpeg.org/

[12] https://talkbank.org/manuals/CHAT.pdf

5.2.3.2 Preparing data for ASR model optimisation

Two different text datasets are needed to optimise the base ASR model. One which optimises the transcription by providing relevant vocabulary or phrases, and the other by providing a text file containing domain-specific text data related to the use case. For this, the following texts were compiled:

a) all information units and expected phrases for each of the Story Time stories in a text file with a single information unit per line (included as a supplement material - File B.1)

b) a total of 2803 lines containing different sentences in a text file after manually correcting transcription errors in the PREVENT transcripts generated by Kaldi (included as a supplement material - File B.2). This text file does not include the test text set described in section 5.2.3.1.

5.2.3.3 Model optimisation

Amazon Transcribe had a different approach to using custom vocabulary methods. The text file containing information units from all comics was uploaded to Amazon transcribe bucket. In their Application Programming Interface call, that text file name was provided as "VocabularyName" under "ModelSettings". In the rest of the ASR systems except Kaldi, vocabulary/phrases as a list of strings (with a boost value on Google and IBM Watson) were provided.

To build a custom language model, the text file containing sentences was uploaded to the ASR system provider's web portal, and after some time, a customised model was ready to be used. The portal provided the model identifier so they could be accessed using Application Programming Interface calls. In DeepSpeech and Kaldi, a custom language model was trained using the KenLM library [275], and the adaptation methods provided in DeepSpeech and Kaldi were utilised for model adaptation.

5.2.3.4 Analyses

Python scripts were developed for each ASR systems to request the transcription using their Application Programming Interface calls. Using the jiwer[13] 2.2.0 Python library, Word Error Rate (WER) was calculated as a performance metric to compare transcription accuracies [276]. WER is calculated as an edit distance between two transcripts, where each ASR generated transcript (hypothesis) is compared to the manual transcript as the gold-standard or reference. The edit distance comprises the number of insertions, deletions, and substitutions to transform reference text into the hypothesis text (Eq. 5.1).

$$WER = \frac{Insertion + Deletion + Substitution}{Number\ of\ words\ in\ reference\ text} \qquad Eq.\ 5.1$$

5.2.4 Comparison results

This subsection shows the comparison results based on transcription costs, data security, and the accuracy based on WER.

5.2.4.1 Transcription costs

Table 5.2 shows various ASR system providers (including a manual transcription provider) and their cost to transcribe audio of one-minute duration. There is no cost associated with open-source ASR systems, and commercial systems are also free for a limited amount of speech data. Transcribing manually using Rev.ai is, on average, 62 times more expensive than a commercial ASR system. Among commercial ASR systems, Microsoft Azure provided the cheapest transcription rate per audio minute.

[13] https://pypi.org/project/jiwer/

Table 5.2 Summary of the cost associated per audio minute using various transcription providers.

Transcription Providers	Type	Cost (per audio minute)
Manual (Rev.ai)	Commercial	$1.25
Google Cloud	Commercial	First 60 minutes free $0.024
Amazon Transcribe	Commercial	First 60 minutes free * $0.024
IBM Watson	Commercial	500 minutes free ***** $0.02
Microsoft Azure	Commercial	5 audio hours free per month £0.012
Deep Speech	Open source (research)	Free
Kaldi	Open source (research)	Free

* There is a cost associated with optimising the provider's base ASR model using domain-specific data to obtain a custom language model.

5.2.4.2 Data processing procedures and data security measures

Audio data contain features that can reveal sensitive information about the speaker's age, gender, language, and accent [277]. Audio features extracted from these data, including spectral and voice quality features (formants, MFCC and voice quality features), are also used in speaker recognition systems to identify the speaker [278]. Therefore, it is crucial to consider data processing procedures and data security measures when evaluating transcription services for clinical research. **Table 5.3** summarises data processing and security measures in place by the ASR system providers. Google Cloud, Amazon Transcribe, and IBM Watson allow users to opt out of using the data to improve their transcription service and delete all the data stored in their cloud. On the other hand, Microsoft Azure does not retain any audio data, and the data is encrypted during the transfer to their server, ensuring data privacy protection. All the commercial ASR systems were General Data Protection Regulation (GDPR) compliant, and to

use their transcription service, their data centre based in the UK (geographical location) was selected to adhere to GDPR.

Table 5.3 Summary of the data processing procedures and data security measures used by the ASR system providers.

Transcription providers	Data and privacy
Google Cloud	Google processes the audio data in memory; however, it logs some metadata related to the transcription request. The transcript output is stored for approximately five days. Users should opt out of data logging and make sure that Google does not use the input data.
Amazon Transcribe	The data is stored in the Amazon Cloud. By default, it uses the input audio to improve their service; however, they allow customers to opt out of having their data used by Amazon by using an AWS Organisations out-out policy.
IBM Watson	The data is encrypted at-rest, and the user can delete all the audio and transcript data.
Microsoft Azure	Audio input data is encrypted in-transit and processed only on Azure's server memory for real-time transcription. The audio and transcriptions are not stored by their service.
Deep Speech	No data privacy concerns, as the model is deployed offline.
Kaldi	No data privacy concerns, as the model is deployed offline.

5.2.4.3 WER Analyses

Table A.2 shows the WER for each of the Cloud-based ASR and open-source ASR outputs compared to manual transcriptions. The lowest WER for each story comes from Microsoft Azure's base (Mean ± SD = 0.17 ± 0.1) and custom language model (Mean ± SD = 0.16 ± 0.8). **Figure 5.2** compares reference transcripts (manual method) and automatic transcripts from

ASR services (hypothesis). Microsoft Azure's confidence interval is narrower than all other ASR systems.

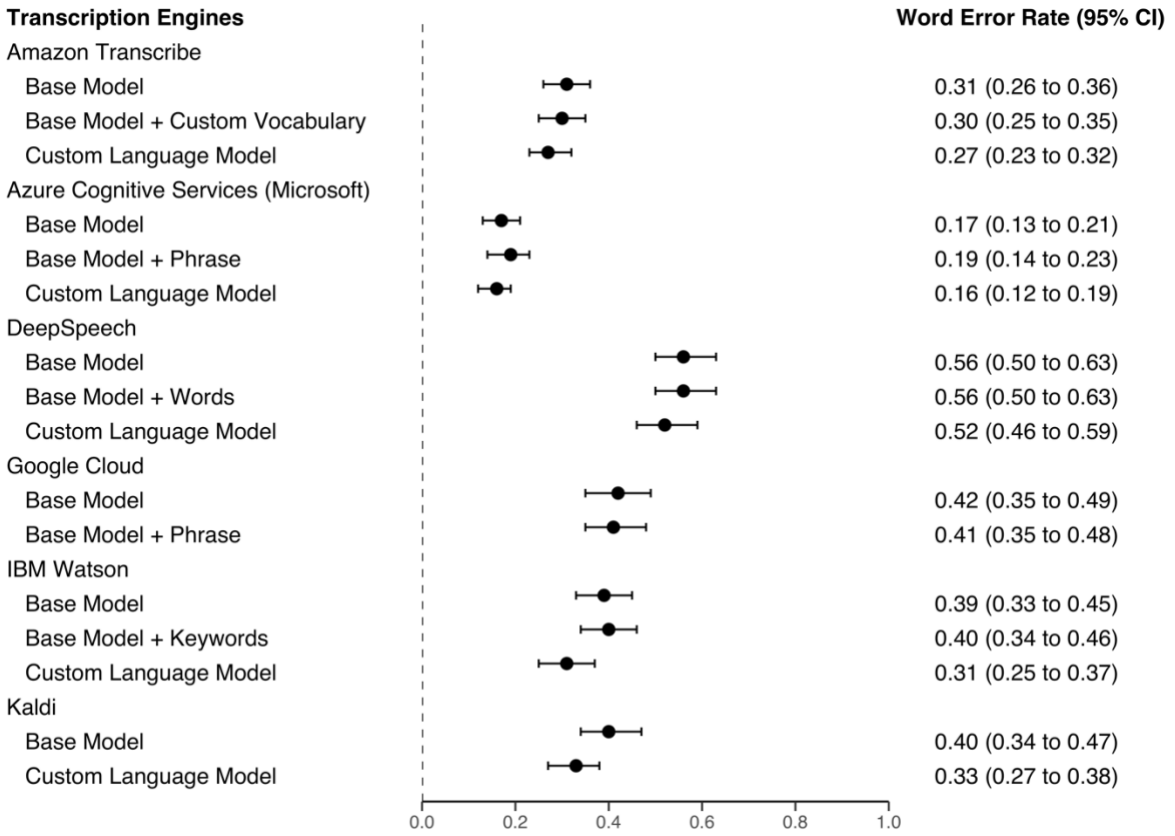


Figure 5.2 Plot shows the average Word Error Rate (WER) obtained using different ASR systems and manual transcription outputs. 95% confidence intervals are calculated from t-distributions. Abbreviation: CI - Confidence Interval.

ASR optimisation method that used vocabulary/phrases/keywords did not always produced transcripts with lowest WER compared to their base model. For example, the average WER utilising such optimisation method in Microsoft Azure (0.19) and IBM Watson (0.40) was higher compared to their base model (0.17 in Microsoft Azure and 0.39 in IBM Watson) suggesting such optimisation does not always improve transcription accuracy. Transcription outputs obtained from these optimisations were further investigated. The word or phrases appeared in the transcripts even though they did not occur in the speech recording because this optimisation process tried to find words/phrases provided in the list by splitting the word, merging two words, or removing repetitions which added noise (example: "there was" to

"nose", "for this" to "flowers", "be a" to "beard", "shorts" to "Shorty" (Corky's friend), "red top" to "red rose top", "tree" to "street", etc.). However, proper nouns such as Corky and Judy were transcribed correctly.

The custom language model optimisation method always provided lower average WER than other models. Based on the above accuracy and data security measures comparison, a custom language model using Microsoft Azure was chosen as the ASR system for the feature extraction pipeline.

5.2.5 Benefits and limitations of different ASR systems and the utility of Word Error Rate (WER) as a transcription accuracy measure

Commercial ASR systems offer benefits in terms of accuracy, scalability and accessibility for transcribing speech data in AD research. However, they also have limitations, including limited optimisation options, the cost associated with using these systems, and privacy concerns about the use of personal data. On the other hand, open-source ASR systems provide flexibility and optimisation opportunities; however, they require technical expertise for setup, optimisation and maintenance.

Although WER is a commonly used metric, it is important to consider its limitations and implications in the context of AD research. AD patients may exhibit speech patterns that differ from the general population, making accurate transcription and word-level evaluation challenging [209]. Phoneme and utterance-level accuracy is recommended in using an ASR system to extract audio features (for example, speech rate, articulation rate, and others) [227]. However, this thesis focuses on WER because word level accuracy was most relevant for the extraction of semantic and language features (in sections 5.5.1 and 5.5.2).

5.3 Evaluation of the performance of retrofitted word embedding models using a subset of the PREVENT dataset

Word2Vec models are trained on large generic corpora, such as Wikipedia or news articles, designed to capture semantic relationships between words [120]. However, when these models are used on a narrow corpus or a specific domain, they may show limited performance because the context in which the words are used might be different [279], [280]. Lexical variants from the Story Time task were identified (included as a supplement material - File B.3) and counter-fitted using the algorithm provided by N. Mrkšić et al. (2016) [280] into the generic British National Corpus (BNC) Word2Vec model [281]. This process improved semantic distances between related words (**Table 5.4** and **Figure A.1**).

Table 5.4 Top 5 words that are semantically close to the word "boy" before and after the domain adaptation.

Generic model		**Adapted model**	
Words	**Cosine similarity**	**Words**	**Cosine similarity**
girl	0.725	lad	0.974
lad	0.439	chap	0.973
kid	0.421	guy	0.968
man	0.414	child	0.968
chap	0.412	youngster	0.967

5.4 Fine-tuning the BERT model for surrogate phrase detection and evaluating its performance on a subset of the PREVENT dataset

Surrogate phrases (such as "I do not remember", "I guess", and "I think") are studied in identifying MCI from healthy adults [45]. In addition, such phrases are recommended to be removed before estimating word count and information units (IUs), as stated by Nicholas and Brookshire (1993). These phrases are detected manually or automatically using lexical rules, part-of-speech (POS) tags and list-matching methods [45], [141]. The rule-based approach

follows a supervised machine-learning technique that employs lexical, syntactic and semantic features [282].

Recently, pre-trained language models have significantly improved the performance of various NLP tasks with the limited amount of labelled data for the supervised learning [283]. In this doctoral research, an automatic detection of surrogate phrases using the BERT model [123] as a sequence labelling problem using story recall transcripts was developed, which is described in this section.

5.4.1 Methods

This section presents the methodology used in detecting surrogate phrases.

5.4.1.1 Preparing a surrogate phrase corpus using story retelling scenario

Surrogate phrases that occur in story retellings were identified from the Story Time task and the literature [45], [141], which are shown in **Table 5.5**. These include phrases that are used in indicating difficulty to recall, e.g., "I guess", "all I can remember", or "I cannot recall". For this study, a story-retelling corpus containing surrogate phrases was prepared using retellings from the MC Test challenge dataset [284] and story narrations prepared from the Story Time task.

Table 5.5 Examples of surrogate phrases that indicate difficulty in retelling stories.

Surrogate phrases
somewhere, something, that, thing, maybe, "let me think", "I can't remember", "I just can't remember", "that thing", "I believe", "I think", "I am guessing", "I guess", "I am confused", "all I can remember", "I am not sure", "I am trying to think", "I do not recall", "I do not remember", "something like that", "I do not know", "I forgot", "as far as I can remember", "I know it didn't say", "I can't think", "I am gonna go with", "I don't", "I mean", "I suppose", "I can't tell", "what was that", "I know it was not", "some kind of", "something similar to"

This story corpus with surrogate phrases was automatically split into 60% training, 20% validation, and 20% test set. The division of surrogate phrases into the three sets is shown in **Table 5.6**. Some of the surrogate phrases in the training set were also in the validation and test set; however, the validation and test sets contained unseen stories by the trained model, and the context in which these surrogate phrases were used may be different.

Table 5.6 Surrogate phrases in the training, validation, and test set after the split.

Training	Validation	Test
all I can remember	all I can remember	as far as i can remember
as far as I can remember	I am confused	as you know
I am confused	I am gonna go with	I am not sure
I am trying to think	I am guessing	I am confused
I believe	I am not sure	I am gonna go with
I can't remember	I am trying to think	I am trying to think
I can't tell	I believe	I am unsure
I do not know	I can only remember	I believe
I don't really remember	I can't really remember	I can't remember
I don't recall	I can't remember	I can’t think of that word
I don't remember	I can't tell	I didn't say
I forgot	I can't think	I do not know
I forgot this	I can't think of that word	I don't remember
I guess	I cannot remember	I don’t know
I just can't remember	I do not know	I don’t recall
I know	I do not recall	I dunno
I mean	I do not remember	I forgot
I see	I don't know	I guess
I think	I don’t recall	I imagine
I'm confused	I don’t recall anything	I just can’t remember
I'm gonna go with	I don’t remember	I really do not know
I'm guessing	I guess	I reckon
I'm not quite sure	I know it didn't say	I suppose
I'm not sure	I know it was not	I think
I've forgotten	I mean	I think that
let me think	I suppose	let me think
oh I see	I think	maybe
oh god	I'm confused	one of those things
oh no	I'm not sure	some kind of
one of those things	I've forgotten	something
some kind of	let me think	something like that
something	maybe	somewhere
something like that	one of those things	that thing
something similar to	something along that line	this stuff
that thing	something like that	you know
you see	that thing	
	you see	

5.4.1.2 Real story recall datasets for testing the fine-tuned model

Two story recall datasets that contained story recall responses were used for testing. One was a subset of Story Time retellings from PREVENT data set. This subset contained story retellings from all the healthy participants with 310 files that included both immediate and delayed recall.

PREVENT cohort contains only healthy participants, and because the lack of story retelling dataset containing MCI/AD groups available in English, dataset named Datasets of Neuropsychological Language Tests in Brazilian Portuguese obtained from an online repository[14] was used as an additional test dataset. This test set contains transcripts of speech data from participants collected by researchers in clinical or academic studies [285]. One of the stories included in this dataset is the Wallet story from Arizona Battery for Communication Disorders (ABCD), which has immediate and delayed retellings from 23 elders with MCI and 12 healthy ageing adults. In this experiment, the Wallet retellings were translated to English from Portuguese using Google Translate[15]. There were seventy files, including both the immediate and delayed recall, with 24 files from controls and 46 from MCI participants.

5.4.1.3 Labelling surrogate phrases in the corpus and story recall datasets

Surrogate phrases in the corpus were labelled using the BIOES tagging scheme [286] used in sequence labelling tasks (i.e., Named Entity Recognition).

Table 5.7 describes BIOES tags. An example of a story with BIOES tags is shown in **Table 5.8**.

[14] https://github.com/nilc-nlp/DNLT-BP

[15] https://translate.google.com/

Table 5.7 Description of BIOES tags used in labelling the data.

Abbreviation	Description
O	Outside of a phrase
B-Phrase	Beginning of a phrase
I-Phrase	Inside of a phrase
E-Phrase	End of a phrase
S-Phrase	Single phrase

Table 5.8 Surrogate phrase corpus preparation using BIOES tagging scheme.

Story	once upon a time there was a boy named jack he wanted to play with his dog I believe his name is max jack finds something he throws the that thing across the yard max sees a ball and brings the ball to jack they play fetch for a long time I think jack then wanted to take max for a walk he looked for max's leash I don't remember what he did after that
BIOES Tags	O O O O O O O O O O O O O O O O O B-phrase E-phrase O O O O O O S-phrase O O O B-phrase E-phrase O O O O O O O O O O O O O O O O O O O O O B-phrase E-phrase O O O O O O O O O O O O O O B-phrase I-phrase E-phrase O O O O O

5.4.1.4 Fine-tuning BERT model

Deep learning state-of-the-art algorithms require a large amount of data for training. Due to data scarcity to train the model from scratch, the pre-trained language model (trained using generic corpus - Wikipedia, Google news) can be fine-tuned with a limited amount of domain-specific data (also discussed in section 5.3), which has proven to be effective in NLP tasks [283]. Pre-trained BERT models ($BERT_{BASE}$ and $BERT_{LARGE}$) hugging-face[16] [123] were fine-tuned[17] using the surrogate phrase story corpus developed in section 5.4.1.1.

[16] https://huggingface.co/transformers/

[17] Both BERT models were fine-tuned using Tesla K80 GPU (on Google Colab) with a batch size of 15.

5.4.1.5 Evaluation of fine-tuned BERT model

The performance of fine-tuned BERT models was evaluated using Positive Predictive Value (also known as Precision), Sensitivity (also known as Recall) and F1 metrics.

Positive Predictive Value: The proportion of surrogate phrases that were identified by the model that are indeed surrogate phrases (Eq. 5.2).

$$Positive\ Predictive\ Value\ = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive} \qquad Eq.\ 5.2$$

Sensitivity: The proportion of surrogate phrases that are correctly identified by the model out of the total surrogate phrases available in the data (Eq. 5.3).

$$Sensitivity\ = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative} \qquad Eq.\ 5.3$$

F1: The harmonic mean of PPV and sensitivity scores [287] (Eq. 5.4).

$$F1\ = \frac{2\ (Positive\ Predictive\ Value\ *\ Sensitivity\ )}{Positive\ Predictive\ Value\ +\ Sensitivity} \qquad Eq.\ 5.4$$

5.4.2 Results

The positive predictive value indicated that $BERT_{BASE}$ predicted 88% of the correct labels in the evaluation and 87% in the test set (**Table 5.9**). Additionally, the sensitivity value in $BERT_{LARGE}$ model showed that the model correctly identified 91% of surrogate phrases in the evaluation set and 83% in the test set. Both of these models show fairly similar results. It is also important to acknowledge that the training corpus size was small.

Table 5.9 Comparison of positive predictive value, sensitivity and F1 scores result on BERT models using surrogate phrase corpus developed for this study.

Model	Validation			Test		
	PPV	**Sensitivity**	**F1**	**PPV**	**Sensitivity**	**F1**
$BERT_{BASE}$	0.88	0.89	0.88	0.87	0.84	0.86
$BERT_{LARGE}$	0.88	0.91	0.89	0.87	0.83	0.85

Abbreviation: PPV- Positive Predictive Value

Table 5.10 shows the evaluation of Story Time and Wallet story recalls. The positive predictive value was higher than the sensitivity value (a very slight difference with Story Time), indicating a few false positives compared to false negatives.

Table 5.10 Comparison of positive predictive value, sensitivity and F1 scores on real story recall test datasets.

Model	Story Time Recalls			Wallet Story Recalls		
	PPV	**Sensitivity**	**F1**	**PPV**	**Sensitivity**	**F1**
$BERT_{BASE}$	0.76	0.75	0.75	0.93	0.81	0.87
$BERT_{LARGE}$	0.79	0.78	0.79	0.94	0.80	0.87

Abbreviation: PPV- Positive Predictive Value

The frequency of use of the surrogate phrase "I think" was the largest in Story Time retellings, and the phrase "I don't know" was frequently detected in Wallet story retellings, as seen in the word clouds of surrogate phrases in two datasets (**Figure 5.3**). **Table A.3** shows a few examples of predicted surrogate phrases from the story retellings where the ones highlighted in bold with underlined text were not used as labels to fine-tune the model.

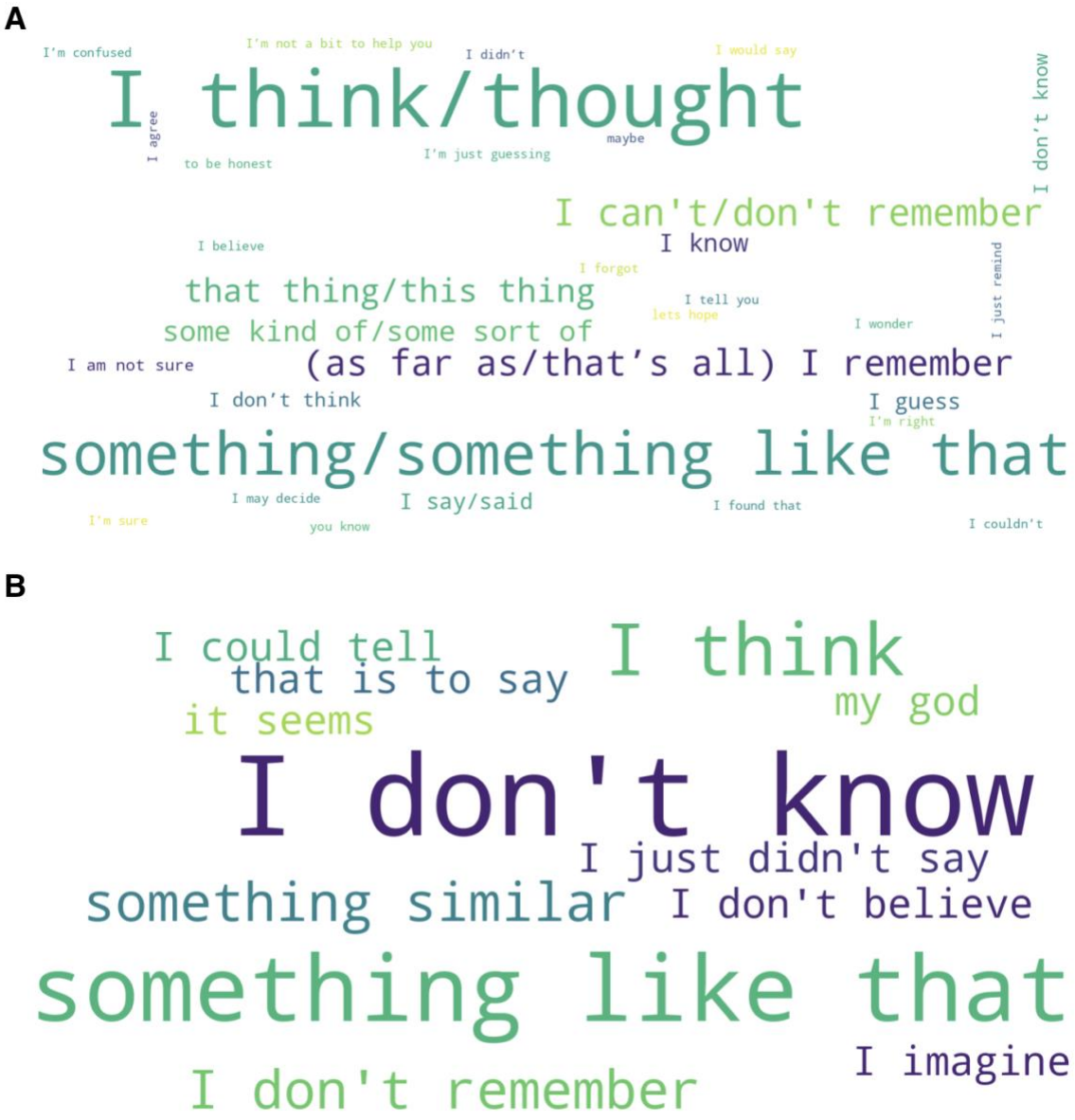


Figure 5.3 Word cloud of surrogate phrases used in Story Time (A) and wallet story (B) retellings. Surrogate phrases including "to be honest", "some sort of", "I may decide", "let's hope", "I'm not a bit to help you", "that is to say", "I just didn't say" were not used in training model (Table 5.5); however, the model captured these phrases showing the robustness of this model. In the world cloud, the text size of the phrase is directly proportional to the frequency of its appearance in the test corpus.

5.5 Feature engineering: deriving features using a combination of computational methods

This section discusses audio and language features derived using a combination of computational methods and Python libraries. The features are organised according to semantic similarity-based, language-based and audio-based features.

5.5.1 Semantic similarity-based Automated Information Unit (AIU) memory score

The pre-processed audio signal is passed to the Microsoft Azure ASR system, which then provides speech-to-text. Using the BERT model, these texts are further pre-processed to determine the surrogate phrases that could appear in the speech responses. Also, the repetition of words or phrases is determined using the n-gram approach [225], [288], where unigram is used for identifying word repeats, and bigrams and trigrams are used for identifying multi-word/phrase repeats. These surrogate phrases and repetitions are removed from the transcription before passing through the information unit scoring Python module. The Natural Language Processing Toolkit provides a list of non-informative words, such as articles, prepositions, and conjunctions, known as stop words. NLP literature recommends removing these stop words in the pre-processing step [289]. Text tokens are extracted from the text after removing the stop words. The steps that are followed after these text pre-processing are listed below:

a) After text pre-processing, words that are POS tagged with nouns, verbs and adjectives are retained.
b) Words from a) are converted to their vector space representation using the Word2Vec model.

c) A manually compiled list of information units is also represented as word vectors.

d) For comparison, each word vector from b) is compared with the word vectors from c) using cosine similarity (Eq. 5.5), where a higher value implies that the words are semantically closer. One point is awarded if the similarity value exceeds the threshold value of 0.6; otherwise, zero is awarded. This process is repeated until the last word in b). Different threshold values (0.9, 0.8, 0.7, 0.6, 0.5) were empirically evaluated on cosine similarity between the word pairs to determine the best threshold value. A threshold value of 0.6 was found to preserve the semantics of the word-pairs for the purpose of information unit extraction (see **Figure A.1**).

e) Steps c) and d) are repeated for each story included in the Story Time task to find whether a previously learned story is recalled instead of the presented story for the day (incorrect recall). Suppose the total score was higher than 50% for another story than the actual story presented to the participant; it was hypothesised that the participant recalled a different story and scored 0.

f) The expected information units can also take form of multi-word phrases, for example: the word "remove" can also be expressed as "come off" or "wipe off" and the word "stop" as "put out". To find such information units expressed as multi-word, n-grams are extracted from the text before the pre-processing in step a). These n-grams are then searched within the list of expected information units, and one point is awarded for each n-grams that is found.

g) The final score is calculated by summing all awarded points and dividing by the total number of information units in c) to standardise the final score in the range of 0-1.

$$Cos(v_1, v_2) = \frac{v_1 . v_2}{||v_1|| ||v_2||} \qquad Eq.\ 5.5$$

where v_1 and v_2 are word vectors from speech token and manual information unit token, respectively.

Another feature that is also extracted is information density. The information unit score from e) is divided by the total words used in story narration to measure the information density of participants' responses.

5.5.2 Lexical, syntactic, and audio features from transcripts

5.5.2.1 Lexical and syntactic features

As discussed in Chapter 2, lexical richness features such as type-token ratio, moving average type-token ratio, Brunet index, and Honore's stats are extracted from the transcripts after labelling Part-of-Speech tags using Stanford Part-of-Speech tagger[18]. Other features, including proportions of nouns, verbs, adverbs, adjectives, and past tense verbs, are also extracted from the Part-of-Speech tags.

5.5.2.2 Psycholinguistic features

Literature shows that a semantic impairment may be indicated by an increased reliance/dependency on highly familiar words [127], [128]. Bristol and Gilhooly-Logie norms [129] provide subjective ratings of words for AOA, familiarity, imaginability and concreteness. A higher proportion (98%) of expected Story Time words[19] were present in the AOA rating list, which was not surprising, as the contents of stories in the Story Time task are quite simple,

[18] https://nlp.stanford.edu/software/tagger.shtml

[19] Words including "centre", "odour", "quiff", "realise", "tache", "tash", and "tyre" were not available in the AOA rating list; however, American English words such as "center", "odor", "realize", "tire" were available and were replaced by their British English equivalent words.

involving domestic scenes. To generate psycholinguistic scores, the average rating for each feature is computed on all nouns and verbs present in the speech response.

5.5.2.3 Audio features from transcription output

Microsoft Azure ASR provides the timestamp of each of the words, which was used to extract a few audio features such as words per minute, average word duration (in seconds), total long pause (>= 2 sec), total short pause (<2 sec), average pause duration (sec), and hesitation ratio. The detail of the features is provided in **Table A.4**.

5.5.3 Audio features

As the pre-processing step, the audio data is converted to 16-bit, 16 kHz sample rate linear PCM format. In addition, the audio data is also converted to mono from stereo (2 channels), which is a requirement for most feature extraction algorithms. The audio data is sent to two different Python modules a) to extract speech fluency features and b) to extract OpenSMILE GeMAPS features.

5.5.3.1 Speech fluency features

Speech fluency features are extracted by implementing a Python module inspired by the PRAAT script developed by de Jong & Wempe, 2009 [290] and based on the evidence from Giles et al. [145], that the number of syllables produced in the picture description task by the moderate AD group was significantly lower compared to the control group. The number of syllables is extracted from the voiced segment using audio intensity (dB). Intensity finds peaks in the energy contour, as a vowel within a syllable has higher energy than surrounding sounds. Intensity contour ensures that the intensity between the current and preceding peaks is sufficiently low, which helps multiple peaks within one syllable to be removed. The features such as speech rate, articulation rate, and average syllable duration are extracted using the

number of syllables in the speech utterances. Also, the silent pauses are extracted by comparing the audio intensity from the intensity frames. The voiced block with a duration of fewer than 0.1 seconds between two silent blocks is merged into the silent blocks. **Figure 5.4** shows the audio feature extraction steps from a sample of a 20-second audio recording.

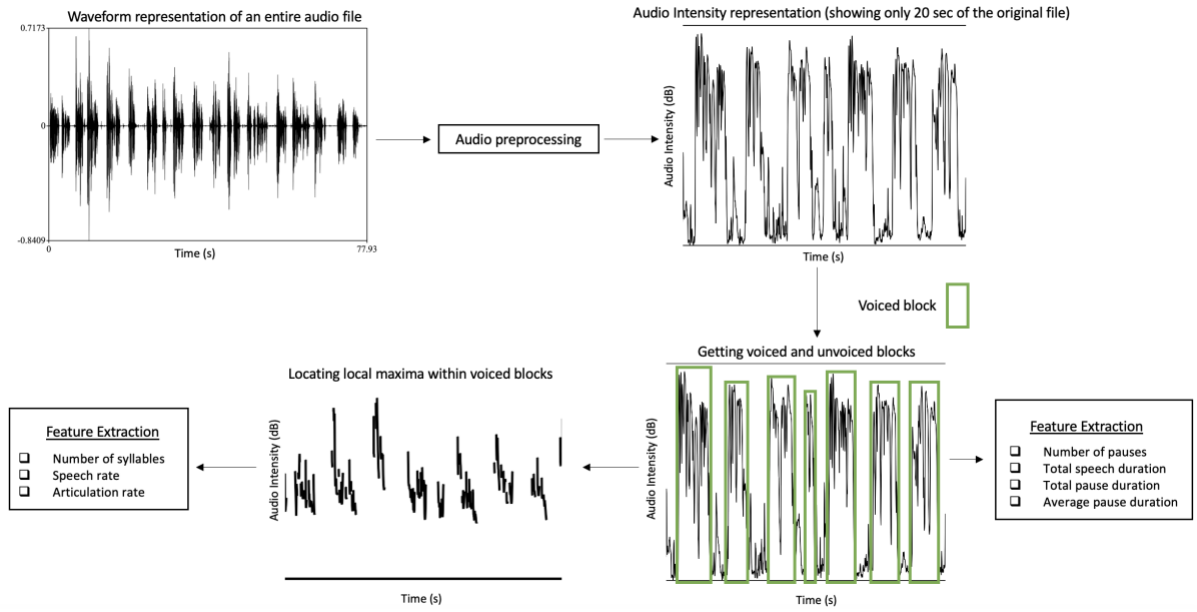


Figure 5.4 Audio feature extraction procedure undergoing various steps such as audio pre-processing, calculating audio intensity, obtaining voiced and unvoiced blocks and locating local maxima within the voiced blocks.

5.5.3.2 OpenSMILE GeMAPS features

GeMAPS features [103] are extracted from the OpenSMILE Python library, which provides 88 feature sets for audio data on a pandas data frame. Recently, there has been a growing interest in using OpenSMILE to extract spectral and voice quality features for AD research [99], [105], [106], [180], [244]. GeMAPS feature set provides features related to frequency (pitch, jitter, shimmer, loudness, harmonics-to-noise ratio, and first four formants - F0, F1, F2, F3) and spectral features (alpha ratio, Hammarberg index, spectral slope, and MFCC features), as defined in section 2.2.2. The voice quality, such as jitter, shimmer and speech fluency features, have shown a significant difference between MCI and healthy controls [93]. Spectral

analysis of speech consists of computation of the MFCC (Mel-frequency cepstral coefficients) that estimates frequency to match more closely what humans hear, identifying irregularities between patients and healthy individuals that can be tracked over time [42].

The detail of all the audio features related to speech fluency, spectral and voice quality is provided in **Table A.5**.

5.6 End-to-end feature extraction pipeline

An end-to-end feature extraction pipeline is a series of steps or processes followed to extract features from raw speech data. This involves data collection, data pre-processing, feature extraction, and post-processing of the extracted features.

5.6.1 Data ingestion

Data ingestion is the first layer of the big data architecture, where raw data and metadata containing information about the data are collected. The data from the Mezurio smartphone application is written in the protocol buffer format, which is further processed using the pyzurio pipeline[20], which outputs data for each participant in an open source, column-oriented data file format named parquet, capable of efficient data storage and retrieval [291].

5.6.2 Data pre-processing

The parquet file created in the data ingestion layer, contains audio in byte format (useful for transmitting from phone to the server) and the associated metadata, including the participant's id, story name, time at which the story was attempted, and operating system of the smartphone

[20] Pyzurio python pipeline is developed by Dr Chris Hinds and stored in a private GitHub repository.

device which recorded the audio. The audio data obtained from the data ingestion layer is further processed (based on the smartphone used for recording speech data) so that it can be passed to the other layers of the pipeline. Smartphones use different audio data formats, mainly .3gp in Android and .m4a in iOS operating systems. These audio formats cannot be used in most audio signal processing libraries and ASR systems. The speech recognition systems are typically trained on the narrow band (8 kHz) or wideband (16 kHz) signal. Therefore, to make the audio data compatible, the audio is converted to the linear PCM mono format with 16 kHz sampling frequency and 16-bit resolution, as the speech characteristics are contained on the higher frequency band, which is known to be sufficient for speech signals [292].

5.6.3 Feature extractor

The feature extractor component of the pipeline contains a Python sub-module for extracting various audio and language features, as discussed in detail in section 5.5. After all of these features are extracted, they will be post-processed, which will be discussed in the following subsection.

5.6.4 Postprocessing and reformatting, and storage of features

The features from the feature extractor, along with the other variables such as participant id, the type of the task (learning, immediate recall, and delayed recall), the time at which the task was completed, and the story identifier, are merged together. All the features are inserted per participant, and their number of rows represents the number of sub-tasks they have completed. The final feature set is converted into the parquet data format, ready for further analysis. A total of 138 different audio and language features are provided by the pipeline (**Table 5.11**).

Table 5.11 Audio and language features that are produced from the pipeline. The number in the parenthesis denotes the number of features for that feature type.

Audio and language feature types					
Audio (n = 20)	Lexical (n = 15)	Syntactic (n = 7)	Semantic (n = 4)	Psycholinguistic (n = 4)	OpenSMILE (n = 88)

5.7 Improving performance and scalability with parallel and distributed computing

Python programming language is increasingly used in Data Science due to the availability of computational libraries such as SciPy, NumPy, pandas, and scikit-learn. These libraries are designed to run on a single core which will not be able to cope with extremely large data resulting in memory problems. Dask is a parallel computing library that allows parallel processing by distributing the data across multiple processors/machines, providing ways to scale the pandas, scikit-learn, and NumPy workflows natively [293]. The larger computation is broken down into smaller ones through a task scheduler and task workers. This is advantageous as Dask can easily integrate into the existing Python code. Therefore, only minimal further code-writing is required.

Dask was implemented into the Python feature extraction pipeline. Performance was compared based on the total time required to finish the entire feature extraction on two available datasets with 462 and 2952 audio recordings, respectively. **Figure 5.5** shows the decrease in total computational time with increased task workers to compute all audio recordings in both sets on a machine with 16 cores and 32 GiB memory. There was a 6-fold decrease in total time required to process the data when 4 four workers were used.

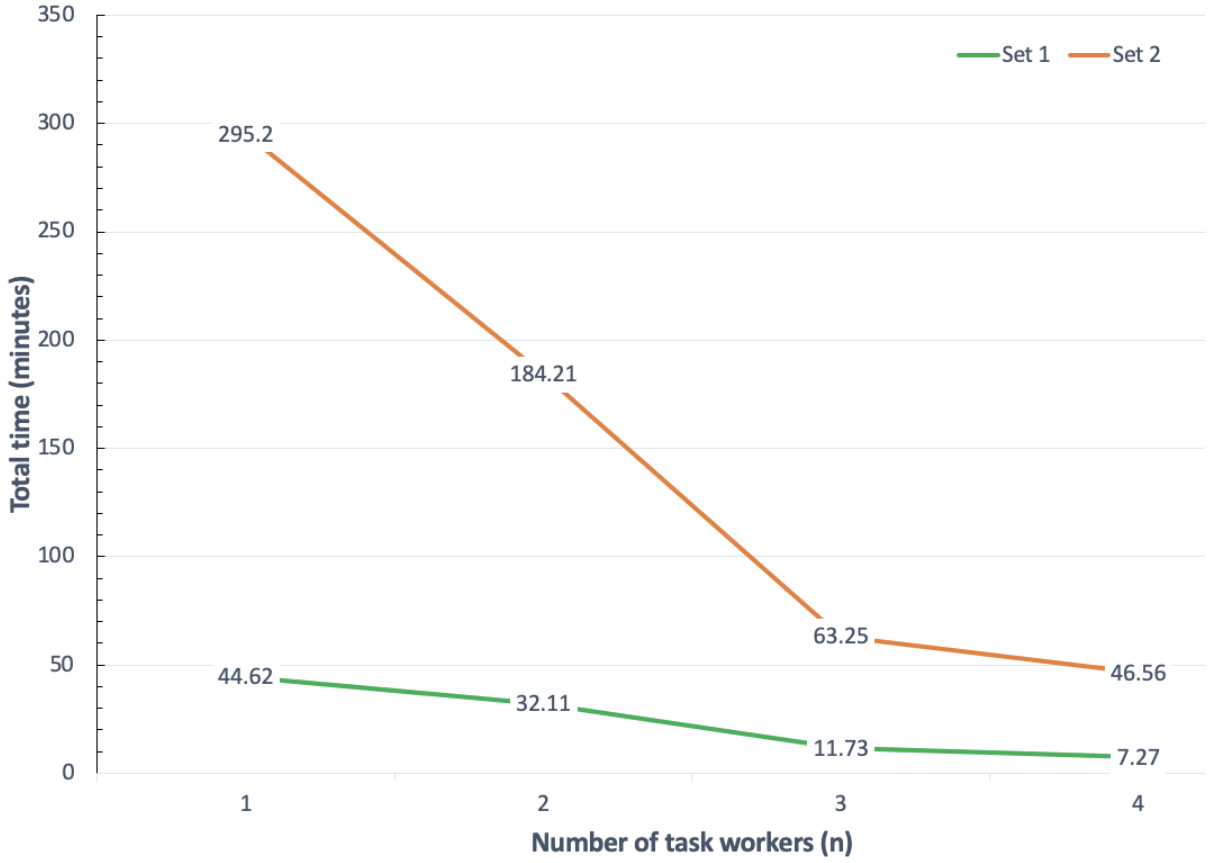


Figure 5.5 Comparison of parallel and distributed computing performance using Dask on a machine with 16 cores and 32 GiB memory. Set 1 consists of 462 and Set 2 contains 2952 audio recordings.

5.8 Summary

This chapter showed the work required to develop an automatic and scalable feature extraction pipeline that extracts various audio and language features from speech samples. The Microsoft Azure ASR system provided the most accurate transcripts compared to the manual transcriptions with their base model. Further optimisation using a custom language model improved the transcripts compared to the base model supplied by the ASR systems. Another optimisation technique utilising a list of phrases or vocabulary did not always optimise the transcription output obtained from ASR systems for story narration tasks. The Word Embedding model included in this pipeline was improved with the counter-fitting technique

[280] developed by N. Mrkšić et al. (2016). A Deep Neural Network based method was used to detect surrogate phrases in story retelling. This method showed robustness in detecting those phrases in story recall tasks, as it did not require a look-up table method as used in previous studies [45], [141]. This method also found other phrases the model did not see in the training step.

This pipeline provides an extensive set of lexical, syntactic, semantic, speech fluency, spectral, and voice quality features, which is beneficial for audio and language analysis as the analysts do not have to develop scripts to obtain specific features using separate libraries. Currently, language feature extraction is tuned for the English language. Substantially more work, including best ASR system selection and domain adaptation of the word-embedding model, is necessary to include other languages in future. If required, the pipeline also allows extracting only audio features from non-English language recordings, as speech fluency, spectral, and voice quality feature extraction works independently of transcripts. The pipeline is also optimised for processing larger datasets utilising parallel computing using Dask.

The literature review in Chapter 3 suggested features useful in audio and language analysis of spontaneous speech in AD research. The PREVENT and GameChanger study (Chapter 4) included healthy participants (no biomarker or disease status); therefore, the machine learning approach was not used. A future study with disease status could use all the features employing the machine learning approach, as these approaches can automatically identify relevant features. Accordingly, about 50 pipeline features were selected based on the literature to simplify the analysis and manage the impact of multiple comparisons. The next chapter, Chapter 6, will evaluate these pipeline features in a small-sized PREVENT cohort before using them in a larger GameChanger study.

Chapter 6. Piloting the methods on the PREVENT dataset

6.1 Introduction

The previous chapter showed that various audio and language features could be extracted from a single audio recording by employing NLP, ASP, ASR and AI/ML technologies. Among the cohorts introduced in Chapter 4, the PREVENT dataset was available from the start of this doctoral research is smaller in size (n = 25) than the GameChanger (n = 18710), contains in-clinic measures (COGNITO and ACE-R), provides an important opportunity to pilot the methods identified for capturing features of speech. Specifically, a) how accurate is the automated method compared to the manual approach for producing speech features based on the transcribed text? and b) are speech features produced by the automated pipeline consistent with clinical parameters? In large-scale studies, speech and language analysis will only be possible if the automated production of language features, such as information units, is accurate compared to manual approaches. This pilot study may highlight improvements in terms of accuracy and efficiency, which will then be followed through before using them in a large study, GameChanger (Chapter 7).

6.2 Methods

6.2.1 Study population

Data from twenty-five middle-aged healthy adults from the PREVENT study [83], self-selected to be at elevated familial risk of AD, highly relevant to eventual screening utility, described in section 4.1, will be analysed. The analyses include all individuals who have provided audio recordings by completing the Story Time task within the Mezurio schedule.

The baseline assessment of the in-clinic tests, namely Computerised Assessment of Information Processing (COGNITO) [249] and Addenbrooke's Cognitive Examination-Revised (ACE-R) [15], were available for all the participants. Associations of these tests with scores from the Story Time task were investigated in this population.

6.2.2 Pipeline features derived from the Story Time task

Chapter 5 discussed the automatic feature extraction pipeline using NLP and ASP techniques to extract semantic (MIU and AIU score), language, and audio features, which are analysed in this chapter. **Table 6.1** lists key measures, including the extracted features using the automatic feature extraction pipeline.

6.2.3 In-clinic neuropsychological tests

The main PREVENT study required in-clinic neuropsychological tests to be completed during the baseline visit for each enrolling participant [83]. This study used the total score on ACE-R [15] as the general cognitive ability. The narrative recall subset from the COGNITO test battery [249] was selected for comparison with the Story Time task. This test contains a story narration with 27 story elements, and participants listen to the story and recall it from their memory.

6.2.4 Selection of confounders

The baseline variables age, sex, and education were identified as potential confounders of the association of speech-based measures with cognition because they are known to be the determinants or risk factors of cognitive decline [29].

6.2.5 Statistical analyses

This chapter uses the key measures extracted from the Story Time task and the in-clinic tests shown in **Table 6.1**. The Story Time measures were extracted using the feature extraction pipeline developed in Chapter 5. This section outlines the statistical methods used in this chapter.

Both the Story Time task and the AIU scoring method are novel; hence, the feasibility and utility of both remain to be established before using them in a large study. The feasibility of the Story Time task was assessed by examining the compliance with the task schedule, calculated for each participant as the proportion of task schedules they attempted up to the moment the participant left the study. Additionally, the participant's task completion is calculated as a ratio of the number of completed Story Time tasks to the total number of available tasks during the given assessment period (baseline, 6-month, and 12-month). Spearman's rank correlation coefficient was used to evaluate the associations of compliance with age and the year of education. A non-parametric Chi-square test was used in evaluating the association of compliance with sex and family history of dementia.

The potential utility of manual and automated information unit scores was evaluated by examining the distribution of IU scores on the immediate and delayed recall sub-tasks of Story Time, the influence of those scores on various Story Time tasks, and their associations with audio features and in-clinic measures as follows:

- The AIU score extracted using the feature extraction pipeline (Chapter 5) was compared with the MIU score using correlation analysis and visual analysis using Bland-Altman limits of agreement plot [294]. The Bland-Altman method examines the average difference

between the two scoring methods by considering the variability in those differences across the participants.

- To investigate the loss of information in storytelling ability between immediate and delayed recalls, the delta features (delta MIU and delta AIU) were calculated by subtracting delayed features from immediate features.
- The association of delay and AIU recall score was investigated using a repeated measures mixed effect model with the participant and story as random effects as shown in Eq. 6.1.

$$y_{ij} = \beta_0 + \beta_1 IU_j + \beta_2 Recalls_{ij} + \beta_3 X1_i + \ldots + \beta_n XN_i + \gamma_{0i} + \gamma_{1j} + \varepsilon_{ij} \qquad Eq.\ 6.1$$

where,

y_{ij} represents, 1) Manual Information Unit (MIU) score for the ith participant on jth story

2) Automated Information Unit (AIU) score for the ith participant on jth story

3) Delta Manual Information Unit (MIU) score for the ith participant on jth story

IU_j represents story IUs for jth story

$Recalls_{ij}$ represents recall type (immediate and delayed) for the ith participant on jth story

$X1_i \ to \ XN_i$ represent other adjustors and predictors (age, gender, education, family history of dementia etc.) for the ith participant

γ_{0i} represents participant as a random effect

γ_{1j} represents story as a random effect

ε_{ij} represents the random error term

- The association of audio and language features with AIU score was investigated using the mixed model that is described in Eq. 6.1.
- Spearman's rank correlation coefficient was used to evaluate the associations between Story Time scores and in-clinic test scores from the PREVENT dataset. The associations of Story

Time scores and both age and education were also investigated using the Spearman's rank correlation coefficient.

All statistical analyses were conducted in R version 4.1.0 (R Foundation for Statistical Computing, Vienna, Austria). Linear mixed effect model analyses were performed using 'lmer' package. The 'lmerTest' package was used to calculate significance, which estimates degrees of freedom using Satterthwaite's method (using a linear combination of the variance-covariance matrix of the fixed and random effects) and calculates p-values for linear mixed models [295]. The method is available in the article by Satterthwaite (1941) [296]. All the plots were produced using R packages - 'sjPlot', 'ggplot2', and 'ggpubr'.

Table 6.1 Definitions of key measures for the analyses.

Measures	**Definitions**	**Analysis level**
Compliance	Compliance is calculated for each participant as the proportion of task schedules they attempted up to the moment the participant leaves the study	P
Completion	Successful completion is one with an audio recording with at least five words in transcription output	P, S, R
Semantic features		
MIU score	Proportion of Manually extracted information units to total expected Information Units for a Story Time comic (see section 4.1.7)	P, S, R
AIU score	Proportion of Automated Information Units to total expected Information Units for a Story Time comic (see 5.5.1)	P, S, R
Information density	Ratio of Information Units and total number of words in the response (see section 2.3.4)	P, S
AIU recall score	The average of the immediate and delayed AIU score	P, S, R
Delta AIU score	The difference in the immediate and delayed AIU score	P, S, R
Lexical features		

Measures	Definitions	Analysis level
Unique words	Total number of unique words (non-repeated words) (see section 2.3)	P, S, R
Moving average type-token ratio	Ratio of the number of unique words to total words (see section 2.3.1)	P, S, R
Brunets' index	Lexical richness measure (see section 2.3.1)	P, S, R
Honore's stats	Lexical richness measure (see section 2.3.1)	P, S, R
Average word length	Average number of letters in the words	P, S, R
Total repeats	Total number repeats (one, two, three, four-word repeats)	P, S, R
Syntactic features		
Open class words	Proportion of sum of nouns, verbs, and adverbs to total words	P, S, R
Close class words	Proportion of sum of function words (i.e., prepositions, determiners, and conjunction)	P, S, R
Pronoun index	Proportion of pronouns by the total number of words	P, S, R
Verbs per utterances	Proportion of verbs per total words	P, S, R
Past tense verb	Proportion of past tense verbs to total words	P, S, R
Number of prepositions	Proportion of prepositions (i.e., in, at, on, of, to, by) to total words	P, S, R
Number of determiners	Proportion of determiners (i.e., your, my, her, this, that) to total words	P, S, R
Psycholinguistic features		
Age of Acquisition (AOA)	The age at which a word was learnt (see section 2.3.5)	P, S, R
Familiarity	A measure of the frequency of exposure to a word (see section 2.3.5)	P, S, R
Concreteness	The degree to which a concept refers to a perceptible entity (see section 2.3.5)	P, S, R
Imageability	A semantic variable that measures how easy it is for a word to arouse mental images (see section 2.3.5)	P, S, R
Speech fluency features		
Speech duration	Total time taken to complete a task	P, S, R
Number of syllables	Total number of syllables produced in a task	P, S, R
Number of pauses	Total number of pauses in speech response	P, S, R

Measures	Definitions	Analysis level
Average pause duration	Sum of pause durations divided by total number of pauses	P, S, R
Speech rate	Number of syllables per second (file duration) (see section 2.2.1)	P, S, R
Articulation rate	Number of syllables per second (speech duration) (see section 2.2.1)	P, S, R
Average word duration	Sum of word durations divided by total number of words (using word level timing obtained from ASR output)	P, S, R
Average syllable duration	Sum of syllable duration divided by total number of syllables	P, S, R
Hesitation ratio	Ratio of total duration of pauses to total speech duration (based on pauses calculated from voiced activity detection within ASP module) (see section 2.2.1)	P, S, R
Spectral and voice quality features		
RMS energy	Root mean square of the total magnitude of the audio signal	P, S, R
Loudness	Approximation of perceived signal intensity in decibel (see section 5.5.2.6)	P, S, R
Spectral flux	Difference of the spectra of two consecutive frames - also known as voice timbre. (see section 2.2.3)	P, S, R
Energy ratio	Energy ratio of the first harmonics of F0 to the second harmonics (see section 2.2.3)	P, S, R
F0	Mean fundamental frequency in semitone range	P, S, R
F1	Mean of centre frequency of the first formant	P, S, R
F2	Mean of centre frequency of the second formant	
F3	Mean of centre frequency of the third formant	
Hammarberg index	The ratio of energy maxima between 0 to 2 kHz and 2 to 5 kHz (see section 2.2.3)	P, S, R
Alpha ratio	Ratio between the sum of energy in lower frequency region (50 Hz to 1 kHz) and higher frequency region (1 kHz to 5 kHz) (see section 2.2.3)	P, S, R
MFCC1	Mel-Frequency Cepstral Coefficients 1 in voiced regions (see section 2.2.3)	P, S, R

Measures	Definitions	Analysis level
MFCC2	Mel-Frequency Cepstral Coefficients 2 in voiced regions (see section 2.2.3)	P, S, R
MFCC3	Mel-Frequency Cepstral Coefficients 3 in voiced regions (see section 2.2.3)	P, S, R
MFCC4	Mel-Frequency Cepstral Coefficients 4 in voiced regions (see section 2.2.3)	P, S, R
Jitter	Variation of frequency from one period to the next period (see section 2.2.2)	P, S, R
Shimmer	Variation of amplitude from one period to the next period (see section 2.2.2)	P, S, R
HNR	Relation of energy in harmonic components to energy in noise-like components (see section 2.2.2)	P, S, R
Voiced segments	Mean length of voiced segments	P, S, R
Loudness	Loudness measures as number of peaks per second (see 5.5.2.6)	P, S, R
In-clinic test scores		
COGNITO score	Number of correctly recalled narrative elements in the narrative recall subtask within COGNITO test battery	P
ACE-R score	The total score on ACE-R as the general cognitive ability	P

Note: Analysis level is represented as P - Participation Level, S - Story Level, and R- Repetition Level; MIU - Manual Information Units; AIU - Automated Information Units; HNR- Harmonics to Noise Ratio; RMS - Root Mean Square; MFCC- Mel-Frequency Cepstral Coefficients; COGNITO - Computerised Assessment of Information Processing; ACE-R - Addenbrooke's Cognitive Examination Revised

6.3 Results

In this PREVENT study participants who provided Story Time audio recordings, the mean (SD) age was 51.5 (5.4) years, 84% were women, and 60% completed fifteen years or more of full-time education (**Table 6.2**). The mean overall ACE-R score was 96.20 ± 3.21, slightly higher among men than women (97.75 ± 0.85 vs 95.90 ± 0.75). Similarly, the overall mean COGNITO score was 15.44 ± 0.71, slightly higher among men than women (17.00 ± 0.91 vs 15.14 ± 0.82). There was no statistically significant correlation between COGNITO narrative recall score and age (r = -0.24, p = 0.25). The correlation between ACE-R score and age was also statistically non-significant (r = -0.057, p = 0.79) (**Figure A.2**), as it was designed more

for screening clinical cognitive impairment [15]. Associations of the years of education with the ACE-R ($r = 0.065$, $p = 0.76$), and the COGNITO score ($r = 0.19$, $p = 0.37$), were non-significant (**Figure A.3**).

Table 6.2 Summary of sample characteristics on the PREVENT dataset and performance (Mean ± SE) on in-clinic tests.

		In-Clinic Tests	
Characteristic	**n (%)**	**ACE-R (0-100)[+]**	**COGNITO (0-27)[+]**
Overall	25 (100)	96.20 ± 3.21	15.44 ± 0.71
Age Group (years)			
<55	16 (64)	96.06 ± 0.92	16.13 ± 0.63
≥55	9 (36)	96.44 ± 0.85	14.22 ± 1.63
Sex			
Female	21 (84)	95.90 ± 0.75	15.14 ± 0.82
Male	4 (16)	97.75 ± 0.85	17.00 ± 0.91
Family history of dementia			
Family history	18 (72)	96.00 ± 0.66	15.05 ± 0.96
No family history	7 (28)	96.71 ± 1.70	16.42 ± 0.61
Ethnicity			
Caucasian	24 (96)	96.45 ± 0.62	15.58 ± 0.73
Other	1 (4)	N/A	N/A
Full-time education completed			
<15 years	10 (40)	96.00 ± 0.60	14.17 ± 1.49
≥15 years	15 (60)	96.33 ± 0.89	15.93 ± 0.57
Work			
Self-employed	5 (20)	94.60 ± 1.93	15.60 ± 1.03
Employed	20 (80)	96.60 ± 0.66	15.40 ± 0.86

[+] Higher the score, the better the cognitive performance; ACE-R- Addenbrookes Cognitive Examination-Revised; COGNITO- Computerized Assessment of Information Processing

6.3.1 Participant's engagement with the Story Time task

The feasibility information is provided in **Table 6.3** and **Figure 6.1**. There were only ten participants in the first assessment period, where the average completion of the schedule of learning subtasks was 75% (SD = 18.9). Due to operational reasons in the trial administration (see section 4.1.5), fifteen new participants were allowed to join the 6-month assessment. Of those fifteen participants, two started the Story Time task on the 12-month assessment only.

The overall mean completion on both delayed (Mean = 71.3%, SD = 13.6) and immediate recalls (Mean = 82.7%, SD = 13.7) were less than on learning subtasks (Mean = 85.8%, SD = 14.5).

Table 6.3 The proportions of participants completing each Story Time learning, immediate, and delayed recall distributed across the PREVENT assessment periods (baseline, 6-months, 12-months).

Assessment period (No. of participants [a])	**Story Index [b]**	**Completeness [c] (%)**		
		Learning	**Immediate Recall**	**Delayed Recall**
Baseline (10)	1	100	100	90
	2	100	90	70
	3	70	70	60
	4	70	70	60
	5	60	60	50
	6	50	50	40
Average (Baseline)		75	73.3	61.7
6-months (21)	7	95	95	86
	8	100	100	90
	9	95	90	81
	10	86	86	76
	11	86	86	71
	12	86	76	57
Average (6-months)		91.3	88.3	76.8
12-months (17)	13	100	100	70
	14	76	76	70
	15	94	82	82
	16	100	94	88
	17	88	88	71
	18	88	76	71
Average (12-months)		91	86	75.3
Average (Overall)		85.8	82.7	71.3

Values represent the proportion of participants completing each of the Story Time learning, immediate recall, and delayed recall; [a] These are all the participants who attempted the Story Time task, but they might not have completed the task successfully, as stated in [c]; [b] Stories as they appear in the Mezurio schedule; [c] Completion is calculated for each participant as the proportion of task schedules they completed. Successful completion involves an audio recording with at least five words in transcription output.

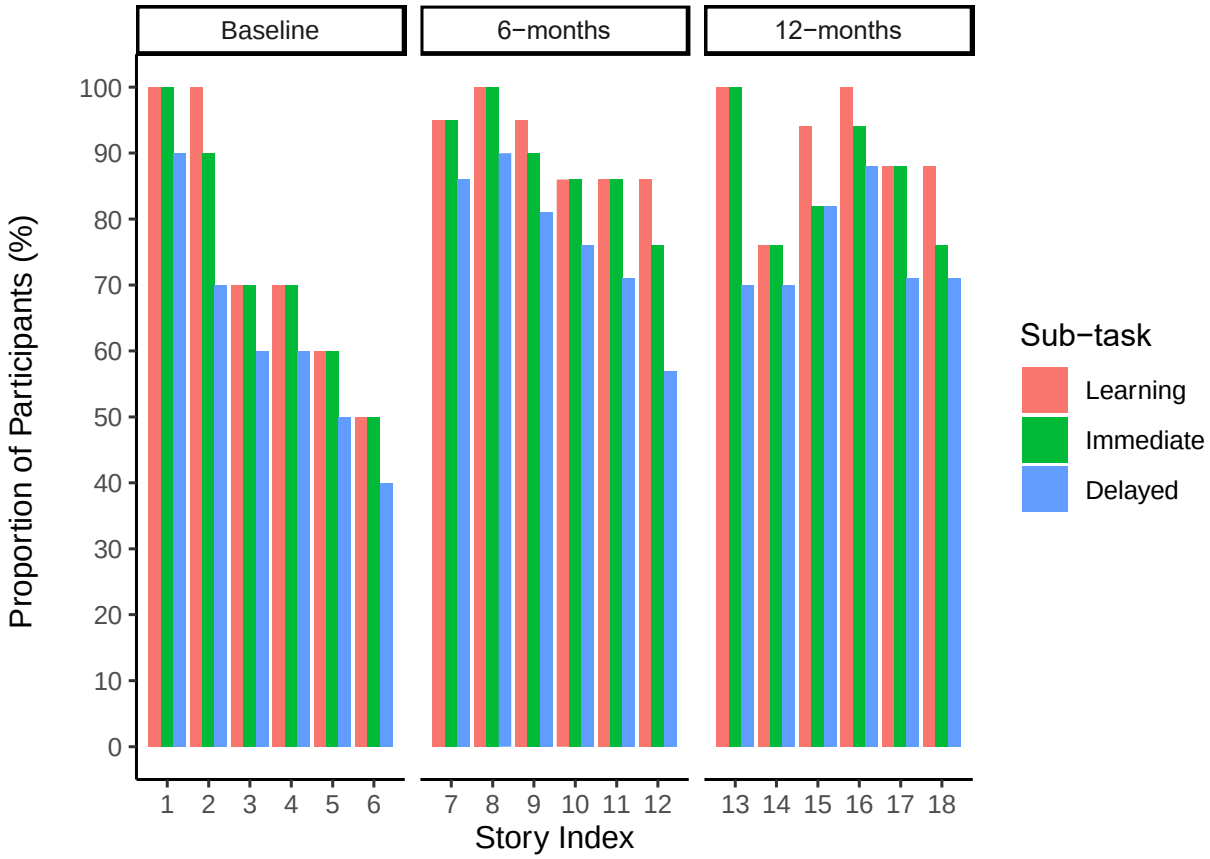


Figure 6.1 Proportion of participants completing Story Time subtasks stratified by assessment periods. The number of active participants in each baseline, 6-months, and 12-month were 10, 21, and 17, respectively.

There was a positive association between age and compliance (r = 0.48, p = 0.02), showing that the participants with increasing age were more committed to the task schedule (Figure 6.2). There were no statistically significant associations between gender (p = 0.37), education (p = 0.20), immediate family history (p = 0.39) and compliance with the Story Time task.

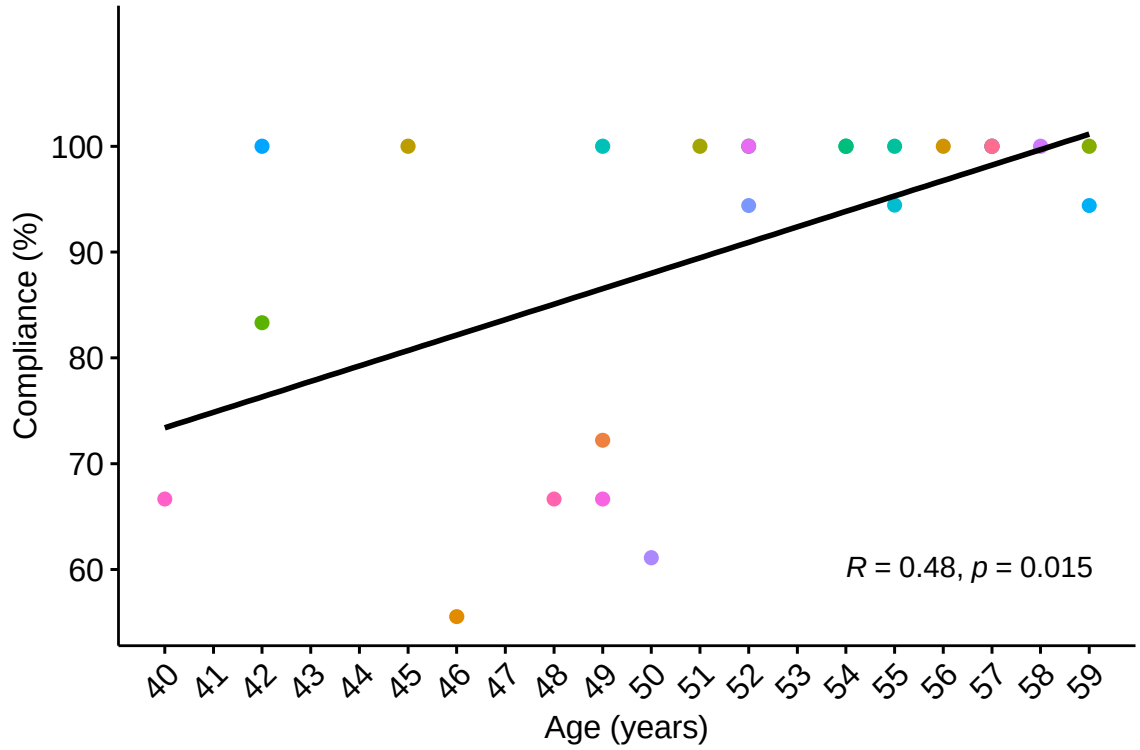


Figure 6.2 Scatterplot showing the association between compliance and the age of the participants. Spearman rank correlation was used in calculating the correlation coefficient and the p-value. Compliance is calculated for each participant as the proportion of task schedules they attempted up to the moment the participant leaves the study.

6.3.2 Distributions of MIU scores on recall tasks

MIU scores were calculated from Story Time's immediate and delayed recalls. The distribution of MIU scores on immediate recall (M = 0.68, SD = 0.16) and 24-hr delayed recall (M = 0.59, SD = 0.22) from Story Time tasks showed variability in the number of information units recalled (**Figure 6.3**). No participants could recall all the information units (score < 1.0) from the Story Time comics. Indeed, during one test of immediate recall, instead of retelling the newly learned story in the immediate recall, one participant recalled the story from the previous day, suggesting that the person either did not understand the task or had a problem remembering the recently learnt story. On a few occasions, participants (n = 7) were unable to recall the story (score of 0) after 24-hrs.

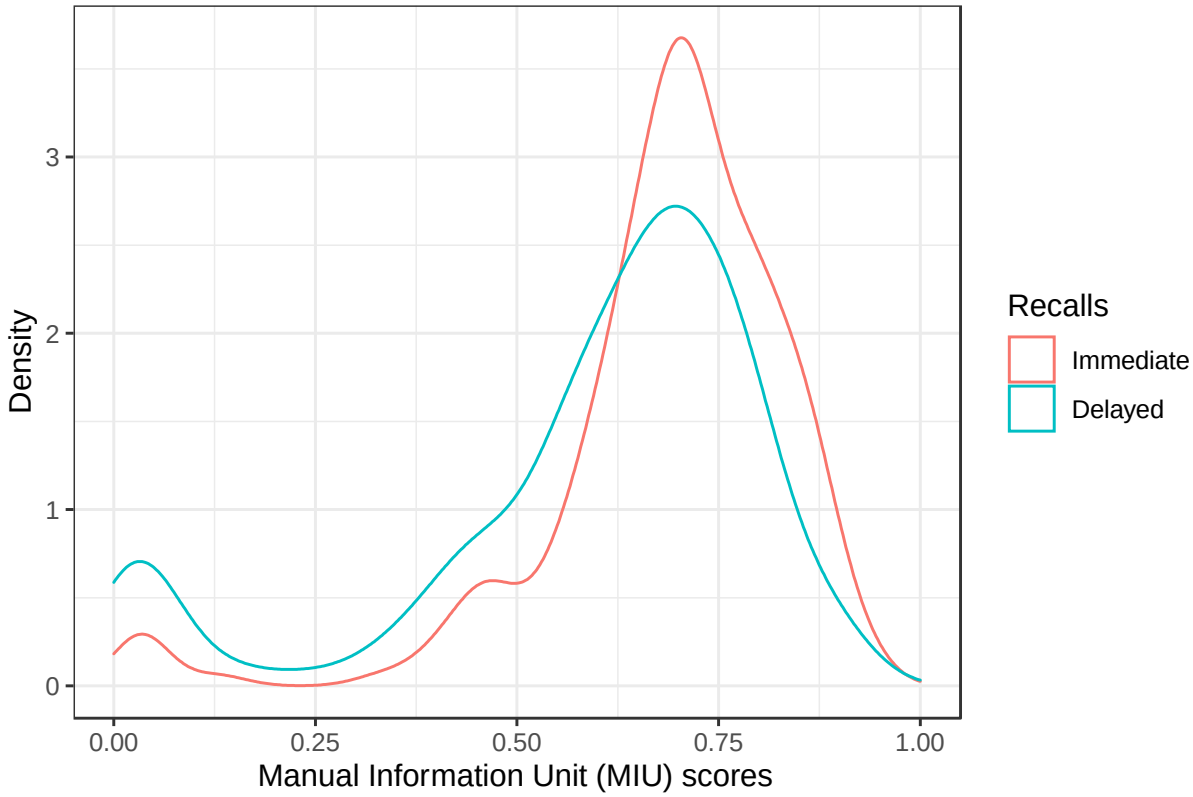


Figure 6.3 Density plot depicting the distribution of MIU scores extracted from both Story Time immediate and delayed recall tasks where the y-axis shows the probability density estimates.

6.3.3 Influence of the story stimuli on MIU recall scores

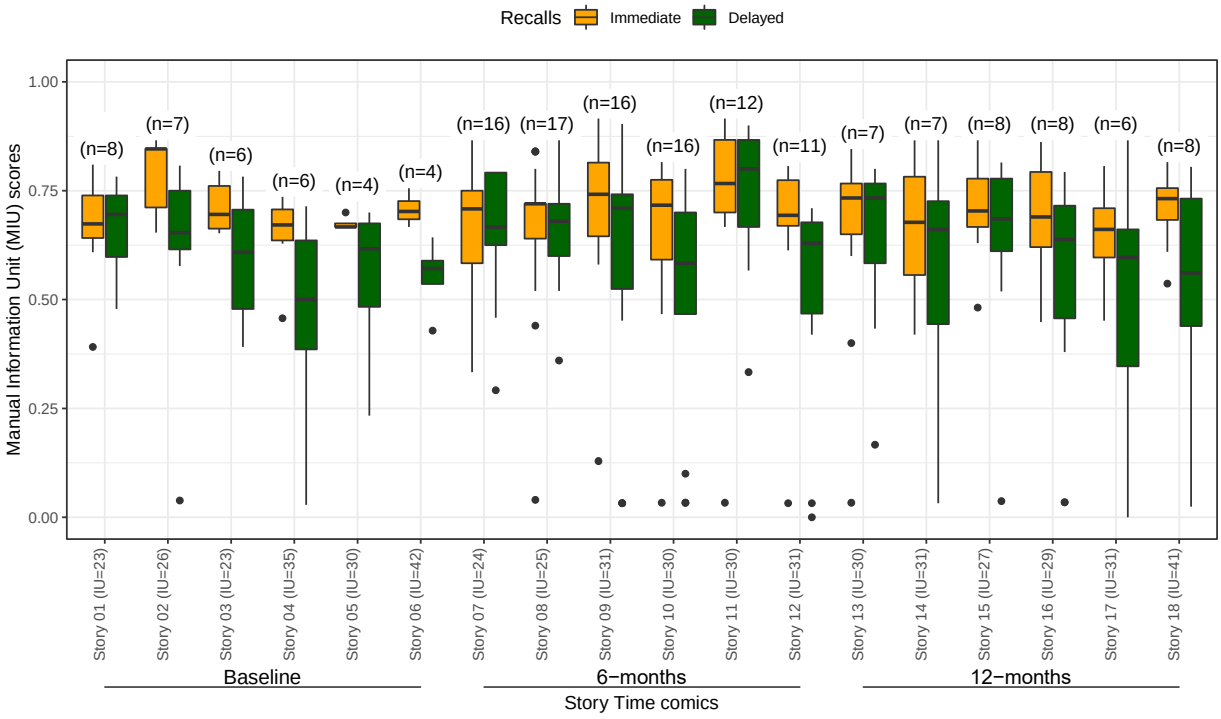


Figure 6.4 Plot showing the distribution of Manual Information Unit (MIU) scores on immediate and delayed recalls across Story Time. The edge of the box represents the first quartile (Q1) and third quartile (Q3), the box represents the interquartile range (IQR), the horizontal line inside the box represents the median value, whiskers represent minimum (Q1 -1.5*IQR), and maximum (Q3 + 1.5*IQR) and dots represent the values outside the range of minimum and maximum. The number of Information Units in each story is labelled as 'IU' and the number of participants completing the story as 'n'.

The mean MIU scores on immediate recall were better than delayed recall on most stories (**Figure 6.4**). In this group of healthy participants, there was a suggestion of an atypical, delayed recall performance on Story 4 (M = 0.46, SE = 0.10), Story 14 (M = 0.58, SE = 0.07), Story 17 (M = 0.49, SE = 0.09), and Story 18 (M = 0.51, SE = 0.09) with more dispersed (wider interquartile range) and larger distribution (more scattered data) of MIU scores compared to the immediate recall and distribution on other stories. These stories had a greater density of information units relative to others demanding higher cognitive effort.

6.3.4 Influence of 24-hr delay between recalls on MIU scores

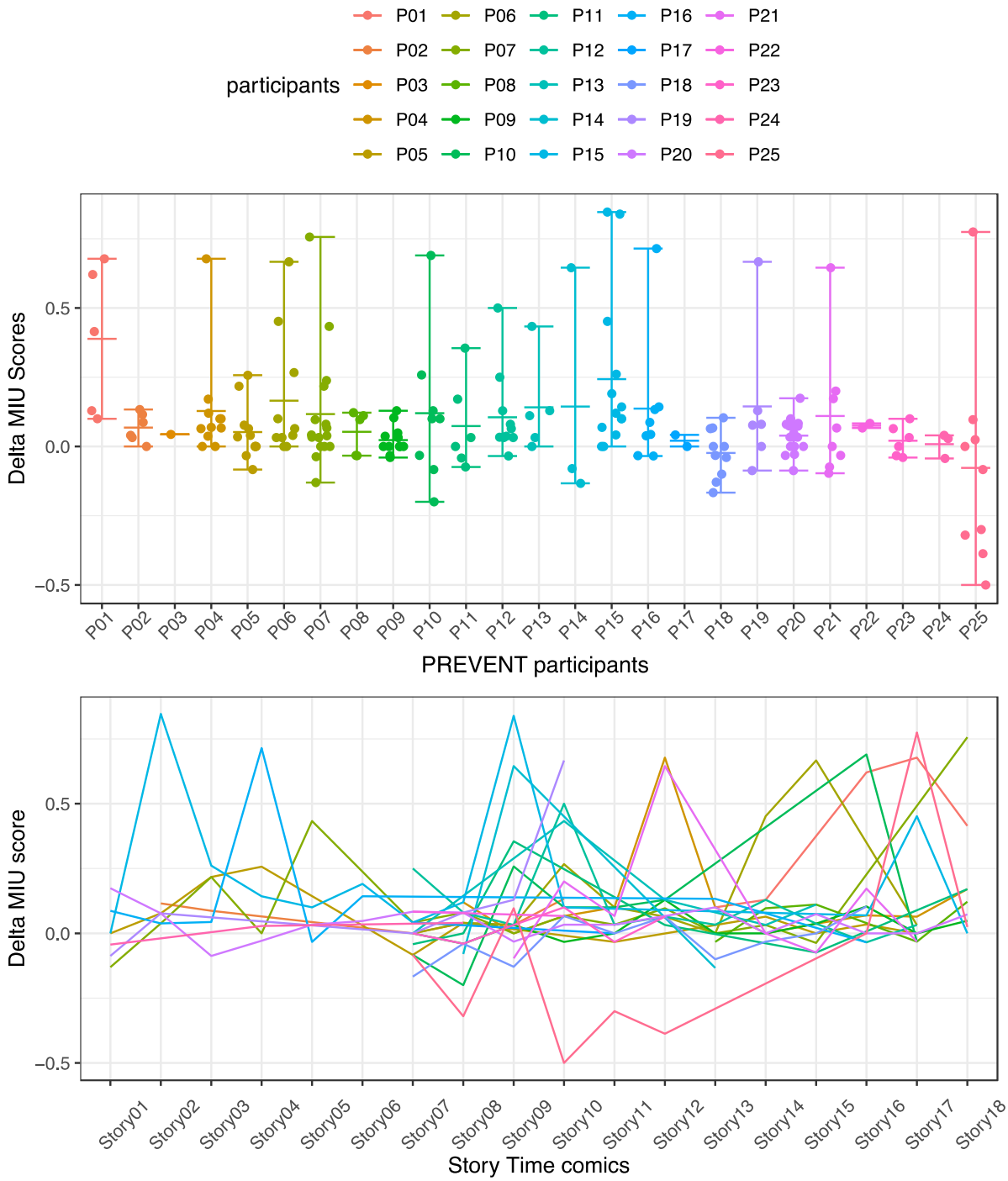


Figure 6.5 Plots showing the influence of 24-hr delay between immediate and delayed recall tasks. Top: Scatter plot showing delta MIU scores between immediate and delayed recall (middle line represents the mean score, and whiskers represent the minimum and maximum scores) for each completed Story Time task. Bottom: Spaghetti plot showing delta MIU scores for each story completed from PREVENT data where each colour represents a participant.

Delta MIU score shows the loss in score between immediate and delayed recall. **Figure 6.5** shows the distribution of delta MIU scores stratified by participants (top) and the story (bottom). Visual examination showed small delta MIU scores on the first completed task. For each individual, the delta MIU scores for later tasks were not consistently better than the first, suggesting an absence of a learning effect. Some participants recalled more information units in their 24-hr delayed recall than immediate recall, with delta MIU scores less than 0, suggesting that they remembered more in the delayed recall subtask. Also, the large negative delta MIU scores by the participant (P25) was because that person recalled the incorrect story in the immediate recall itself, which led to the immediate recall MIU score being zero for those stories.

The association of delay with MIU score was investigated using a linear mixed model with participants and stories as random effects and recall as fixed effect adjusted for story IUs as the story stimuli influenced recall scores (**Figure 6.6**). Additional adjustments for gender, age, and education were evaluated, and their associations were non-significant in this dataset with limited participants. Within this model, the change of scores between delayed and immediate recall was significantly negative (Estimate (SE) = -0.094 (0.02), $p = 7.3 \times 10^{-8}$), showing that the participants scored higher in immediate recall compared to delayed recall.

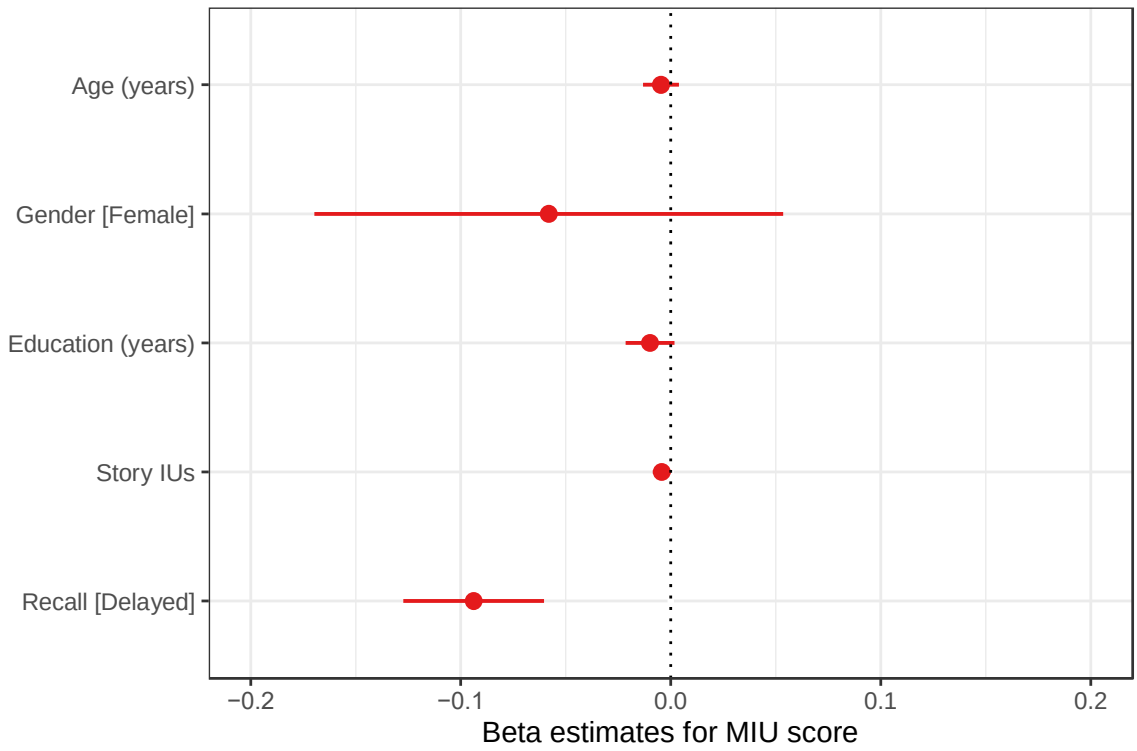


Figure 6.6 Plot showing the beta estimates obtained by regressing MIU scores on age, gender, education, story IUs and recalls using a linear mixed effect model with participants and stories as a random effect. The magnitude of the effect is indicated on the x-axis, and whiskers represent 95% CI.

6.3.5 Association between MIU and AIU scores

As discussed in section 4.1.7, during in-clinic tests, participant's picture description or story recall performance is usually judged by manually listening to their responses and scoring it against a set of expected "information units". The quality of the automated approach needs to be compared to the conventional methods to determine the accuracy of the automated process for producing information units from the audio recording.

Before finalising the feature extraction pipeline presented in Chapter 5, automated methods were evaluated in an iterative process for improvement. The information units were automatically extracted using the transcripts obtained from the open-source KALDI ASR, mainly used in the reviewed articles for automatic speech recognition (Chapter 3). Pearson's

correlation coefficient between MIU and AIU scores was r = 0.73 (**Figure 6.7**). After investigating fifty observations far from the regression line, discrepancies such as transcription errors, semantic similarity, and automated and human judgements were found (**Table 6.4**).

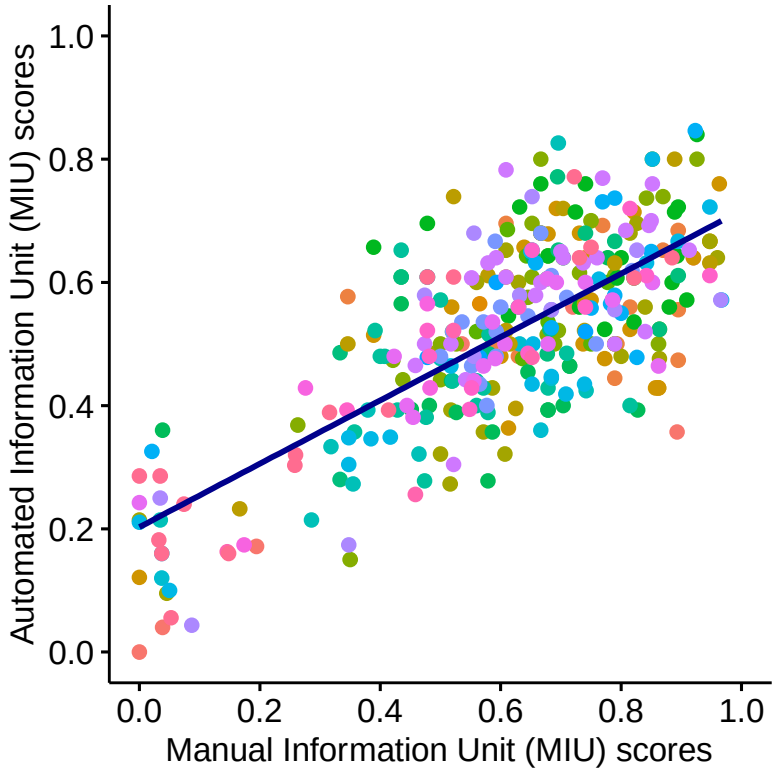


Figure 6.7 Scatter plot of the Manual Information Unit (MIU) scores and Automated Information Units (AIU) scores estimated from PREVENT Story Time data using the KALDI ASR system for transcript generation. The association of these two scores is displayed with a regression line (r = 0.73, p = 7.3 x 10^{-59}). Each colour in the plot represents a participant.

The discrepancies indicated that the main issue was the transcript quality, with 77% of the files including transcription errors. Therefore, various commercial and non-commercial ASR systems were systematically compared to select the system that provides transcripts with better accuracy compared to manual transcription (see section 5.2).

Table 6.4 Issues with Automated Information Unit extraction, shown as - transcript output (expected information unit) in 50 different story transcripts.

Issues	Description	Examples	% of files
1. Transcription errors	ASR system produces transcription errors that can be picked up by humans.	putting on his bit (bed), brings see it (sheet), lurch (lunch), postal (poster), wing (weighing), kicking (chicken), that (dad)	76%
2. Semantic similarity	The semantic similarity between some of the word pairs was too small to be considered a match.	barn (shed), write (paint/draw), rummage (search)	4%
3. Automated judgements	The automated method could not make a judgement on information units as humans do.	last long (comes off), busy working all day (hard work), throw colours (paint)	12%
4. Human judgements	Raters could identify retellings that either belongs to a different story or contains information relating to a different context by reading the transcript. However, the automated method compares semantic distances between words for scoring.	*"an apology I can't remember the story time from yesterday um the last one I can remember is the scout hat but I'm sure I repeated that one so I think I was just in too much of a rush to go see my mom and dad can't remember what the story was."* - In this example, "dad" and "rush" used in another context were expected information units.	44%

As discussed in Chapter 5, in the feature extraction pipeline, other discrepancies were handled in the AIU scoring algorithm by providing synonyms/phrases to the list of Story IUs (Issue 2), using the *n*-gram approach to find multi-word information units (Issue 3), and checking

whether the participant recalled an incorrect story (Issue 4). Using the transcripts obtained from the Microsoft Azure ASR system (the most accurate transcription with the lowest WER) (Issue 1), AIU scores were calculated again.

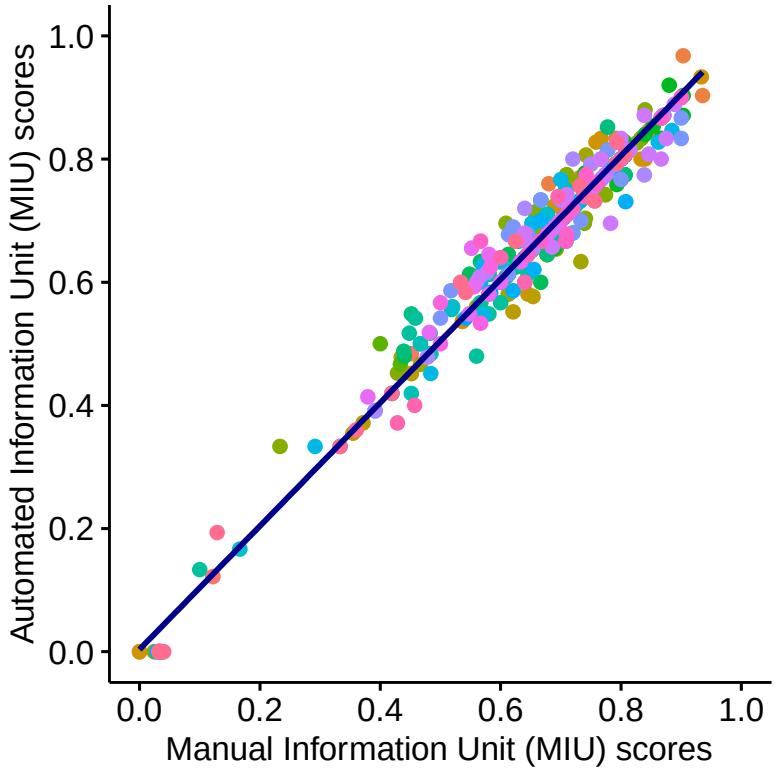


Figure 6.8 Scatter plot of the Manual Information Unit (MIU) scores and Automated Information Units (AIU) scores estimated from PREVENT Story Time data using the transcripts obtained from the Microsoft Azure ASR system. The association of these scores is displayed with a regression line ($r = 0.98$, $p = 1.8 \times 10^{-304}$). Each colour in the plot represents a participant.

The comparison between MIU and the new AIU scores using Pearson's correlation coefficient demonstrated a very high correlation coefficient of $r = 0.98$, as shown in **Figure 6.8**. However, it is important to note that the correlation analysis highly depends on the choice of the subjects, which in this case were healthy participants. Furthermore, this method only quantifies the degree to which the variables are related and does not compare their values or difference between them [297]. To determine whether the two methods are in agreement, the Bland-Altman method of agreement is a more appropriate approach and was used for the analysis.

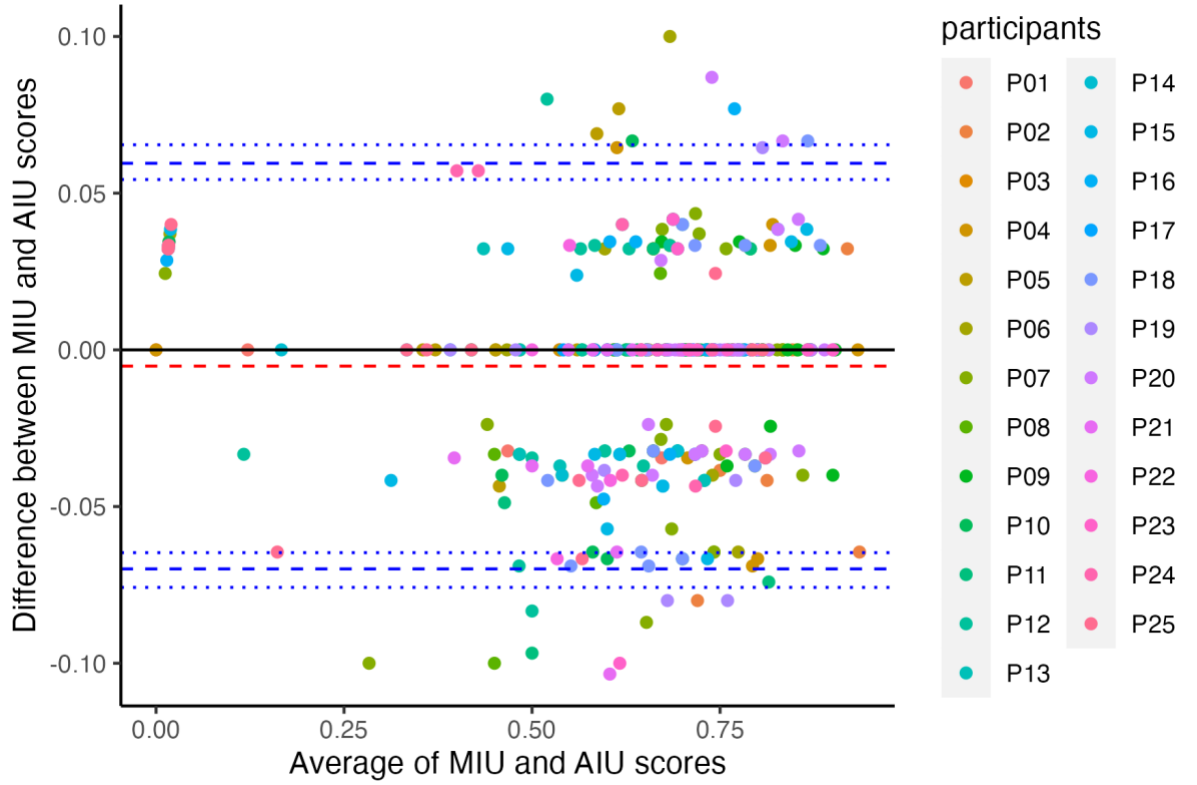


Figure 6.9 Bland-Altman agreement analysis plot on multiple observations per participant. The point shown corresponds to the individual pairs of observations. Each colour in the plot represents a participant. Dashed lines show the mean of the differences between the two measurements (red) and the limits of agreement (blue). The dotted lines (blue) are bootstrapped 95% confidence intervals.

Bland-Altman plot [298] is a visualisation method to plot difference scores of the two measurements (the 'bias') against the mean for each subject. This plot also provides 95% limits of agreement as the mean difference of two standard deviations, with the expectation that 95% of the differences between the scores from the two measurement methods lie within the 95% limits of agreement. This method was used to assess the level of agreement between two scoring methods (MIU and AIU). Story Time data consists of repeated measures, and it is desirable to use all the data for this analysis. While the traditional Bland-Altman method was developed for comparing two sets of measurements on a single occasion, **Figure 6.9** was plotted using the Bland-Altman method in comparing two measurement scores in the context

of the repeated measures [294]. The mean difference between the two scores was -0.005, while the limits of agreement were -0.069 and 0.059, showing 95% of the differences between these two measurements fall within this range, indicating good agreement between the two approaches.

6.3.6 Effect of 24-hr delay between recalls on AIU scores

The association of delay with AIU score was investigated using a linear mixed model with participants and stories as random effects and delay as a fixed effect adjusting for age, gender, education, and story IU. Similar to the result obtained with MIU scores (section 6.3.4), the change in scores (delayed recall - immediate recall) was significantly negative (Estimate (SE) = -0.088 (0.02), $p = 7.9 \times 10^{-7}$), with participants scoring higher in immediate recall compared to the delayed recall (**Figure 6.10**).

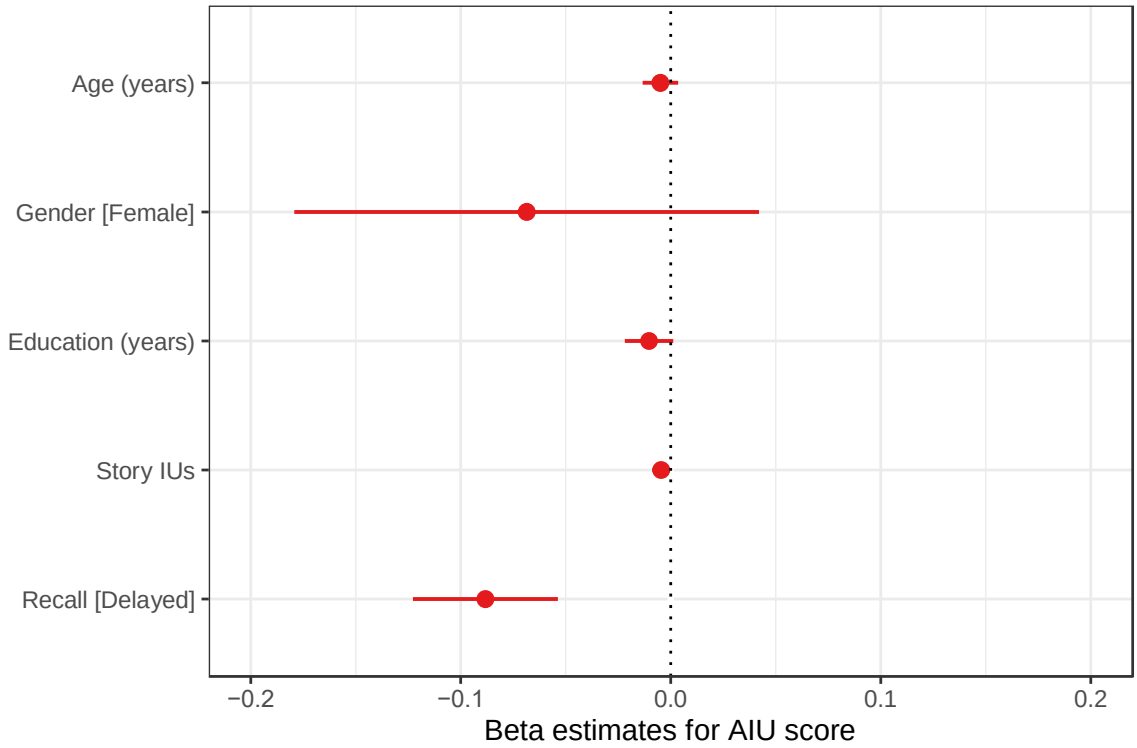


Figure 6.10 Plot showing the beta estimates obtained by regressing AIU scores on age, gender, education, story IUs and recalls using a linear mixed-effect model with participants and stories as random effects. The magnitude of the effect is indicated on the x-axis, and whiskers represent 95% CI.

6.3.7 Exploring associations between AIU recall scores, audio, and language features

In this section, the association of audio and language features with the AIU scores obtained from immediate and delayed recalls were evaluated using linear mixed models with participants and stories as random effects adjusted for age, sex, education, recall type (immediate and delayed) and story IUs.

6.3.7.1 Lexical and syntactic features

The joint associations of lexical and syntactic features (as described in section 2.3.1) on the recall performance with AIU scores are shown in **Table 6.5**. In this cohort of healthy participants (n = 25), higher use of unique words, higher Brunet's index, and lower verbs per utterance were significant predictors for obtaining higher AIU recall scores.

Table 6.5 Joint association of lexical and syntactic measures with immediate and delayed AIU recall scores.

	Automatic Information Unit (AIU) scores			
Predictors	**SD**	**Estimate (SE)**	**Estimate per SD (SE)**	**p**
Brunets' index	1.03	0.060 (0.016)	0.062 (0.017)	0.0002
Unique words	38.25	0.002 (0.001)	0.064 (0.019)	0.001
Verbs per utterances	0.03	-1.094 (0.338)	-0.031 (0.010)	0.001
Pronoun index	0.03	0.962 (0.382)	0.027 (0.011)	0.01
Open class words	0.03	0.645 (0.313)	0.020 (0.010)	0.04
Honore’s stats	281.08	0.000 (0.000)	0.017 (0.011)	0.10
Total repeats	3.16	-0.005 (0.004)	-0.016 (0.012)	0.17
Proportion of prepositions	0.02	-0.614 (0.645)	-0.013 (0.013)	0.34
Average word length	0.19	-0.023 (0.051)	-0.004 (0.010)	0.65
MATTR	0.02	-0.176 (0.421)	-0.004 (0.009)	0.68
Past tense verbs	0.2	0.011 (0.043)	0.002 (0.009)	0.80

Close class words	0.03	0.054 (0.478)	0.002 (0.017)	0.91

Please see Table 6.1 for definitions of the predictors; $N_{participants}$ = 25; $N_{stories}$ = 18; $N_{Observations}$ = 362; Restricted to complete information on age, sex, education, immediate and delayed recalls; The likelihood ratio test was used to determine the predictors to be included in the model; SE - Standard Error; MATTR - Moving Average Type Token Ratio; SD - Standard Deviation; SD is calculated using immediate recalls.

6.3.7.2 Psycholinguistic measures

Psycholinguistic measures extracted from Story Time recall task did not show any significant associations with AIU recall scores in this PREVENT cohort (**Table 6.6**).

Table 6.6 Joint association of psycholinguistic measures with immediate and delayed AIU recall scores.

	Automatic Information Unit (AIU) scores			
Predictors	**SD**	**Estimate (SE)**	**Estimate per SD (SE)**	**p**
Familiarity	25.17	-0.003 (0.002)	-0.074 (0.057)	0.19
Age of acquisition	0.32	-0.025 (0.035)	-0.007 (0.011)	0.41
Imageability	27.45	0.003 (0.003)	0.084 (0.103)	0.47
Concreteness	25.74	0.0001 (0.003)	0.003 (0.066)	0.96

Please see Table 6.1 for definitions of the predictors; $N_{participants}$ = 25; $N_{stories}$ = 18; $N_{Observations}$ = 362; The order of features in the table is based on the p-value; Restricted to complete information on age, sex, education, immediate and delayed recalls; SE - Standard Error; SD - Standard Deviation; SD is calculated using immediate recalls.

6.3.7.3 Fluency-related audio features

In cognitively healthy PREVENT participants, there were both within and between individual variations on the audio feature scores (**Figure A.4**). The joint association between speech fluency features and AIU scores showed that participants with longer speech duration and lower average word duration were significant predictors of obtaining higher AIU scores (**Table 6.7**).

Table 6.7 Joint association of speech fluency features with immediate and delayed AIU recall scores.

	Automatic Information Unit (AIU) score			
Predictors	**SD**	**Estimate (SE)**	**Estimate per SD (SE)**	**p**
Speech duration (sec)	39.57	0.002 (0.000)	0.091 (0.006)	2.2×10^{-31}
Average word duration (sec)	0.03	-0.653 (0.180)	-0.021 (0.006)	0.0004
Hesitation ratio	0.07	-0.282 (0.113)	-0.021 (0.008)	0.01
Speech rate (Syllable/sec)	0.53	-0.019 (0.023)	-0.010 (0.012)	0.46
Average pause duration (sec)	0.26	-0.022 (0.030)	-0.006 (0.008)	0.46

Average syllable duration (sec)	0.21	0.020 (0.039)	0.004 (0.008)	0.61
Articulation rate (Syllable/sec)	0.69	0.018 (0.017)	0.013 (0.012)	0.29

Please see Table 6.1 for definitions of the predictors; $N_{participants}$ = 25; $N_{stories}$ = 18; $N_{Observations}$ = 362; The order of features in the table is based on the p-value; Restricted to complete information on age, sex, education, immediate and delayed recalls; The likelihood ratio test was used to determine the predictors to be included in the model; SE - Standard Error; SD - Standard Deviation; SD is calculated using immediate recalls.

6.3.7.4 Spectral and voice quality features

In this cohort of healthy participants, there were no significant associations between these spectral and voice quality features and the AIU recall scores except with MFCC2, warranting the evaluation of these features in a larger population (**Table 6.8**).

Table 6.8 Joint association of spectral and voice quality features with immediate and delayed AIU recall scores.

	Automated Information Unit (AIU) scores			
Predictors	**SD**	**Estimate (SE)**	**Estimate per SD (SE)**	**p**
MFCC2	7.39	0.0063 (0.0015)	0.0466 (0.0109)	2.9×10^{-5}
Jitter	0.02	-2.8689 (0.9556)	-0.0574 (0.0191)	0.003
Voiced (segments/sec)	0.46	0.0484 (0.0208)	0.0223 (0.0096)	0.021
Energy ratio	3.68	-0.0038 (0.0024)	-0.0140 (0.0087)	0.106
MFCC4	10.83	0.0016 (0.0010)	0.0170 (0.0111)	0.125
F1 (Hz)	85.12	0.0005 (0.0004)	0.0449 (0.0304)	0.141
Spectral flux	0.05	-0.0810 (0.0576)	-0.0040 (0.0029)	0.160
Loudness (dB)	0.08	0.1799 (0.1334)	0.0144 (0.0107)	0.178
Hammarberg index	4.96	-0.0053 (0.0045)	-0.0263 (0.0224)	0.240
Shimmer	0.19	0.0837 (0.0718)	0.0159 (0.0136)	0.245
RMS energy	0.02	0.7232 (0.6456)	0.0145 (0.0129)	0.264
Loudness (peaks/sec)	0.58	0.0122 (0.0111)	0.0071 (0.0064)	0.272
HNR	2.97	-0.0137 (0.0149)	-0.0407 (0.0441)	0.357
F0 (Hz)	4.52	0.0062 (0.0069)	0.0280 (0.0312)	0.371
MFCC1	4.06	0.0024 (0.0033)	0.0099 (0.0133)	0.458
Alpha ratio	4.54	0.0014 (0.0022)	0.0063 (0.0100)	0.530
F2 (Hz)	118.28	-0.0001 (0.0002)	-0.0125 (0.0285)	0.660
MFCC3	10.99	0.0001 (0.0012)	0.0008 (0.0128)	0.948
F3 (Hz)	137.67	0.0000 (0.0002)	0.0025 (0.0284)	0.959

Please see Table 6.1 for definitions of the predictors; $N_{participants}$ = 25; $N_{stories}$ = 18; $N_{Observations}$ = 362; The order of features in the table is based on the p-value; Restricted to complete information on age, sex, education, immediate and delayed recalls; The likelihood ratio test was used to determine the predictors to be included in the model; SE - Standard Error; SD - Standard Deviation; SD is calculated using immediate recalls.

6.3.8 Construct validity of Story Time task

A comparison of MIU and AIU scores showed a high correlation, suggesting that analysis using these scores will produce similar results. Based on this finding, analyses in this section use AIU scores.

6.3.8.1 Association between AIU and in-clinic test scores

Recall AIU and delta AIU scores were averaged over stories for each participant and compared against the available in-clinic test measures (**Figure 6.11** and **Figure 6.12**). Although COGNITO narrative recall and Story Time are different tasks, they measure the same cognitive domain of semantic and episodic memory. There was a moderate correlation ($r = 0.47$, $p = 0.02$, $N = 25$) in the anticipated direction between COGNITO Narrative recall and AIU recall scores. Excluding the participant who performed poorly in COGNITO (2/27), the relationship was still in the expected direction ($r = 0.35$, $p = 0.09$, $N = 24$). Similarly, the correlation between COGNITO and delta AIU score was also moderate, in the anticipated direction ($r = -0.40$, $p = 0.047$, $N = 25$). Excluding the participant who performed poorly in COGNITO (2/27), the relationship was still in the expected direction ($r = -0.32$, $p = 0.131$, $N = 24$).

ACE-R contains various attention, orientation, memory, visuospatial, and visual perceptual skills tests and provides an overall cognitive score [15]. The positive correlation between ACE-R and recall AIU score ($r = 0.26$, $p = 0.22$, $N = 25$) and negative association between ACE-R and averaged delta AIU scores ($r = -0.29$, $p = 0.16$, $N = 25$) were in the anticipated direction, although there was no statistically significant association.

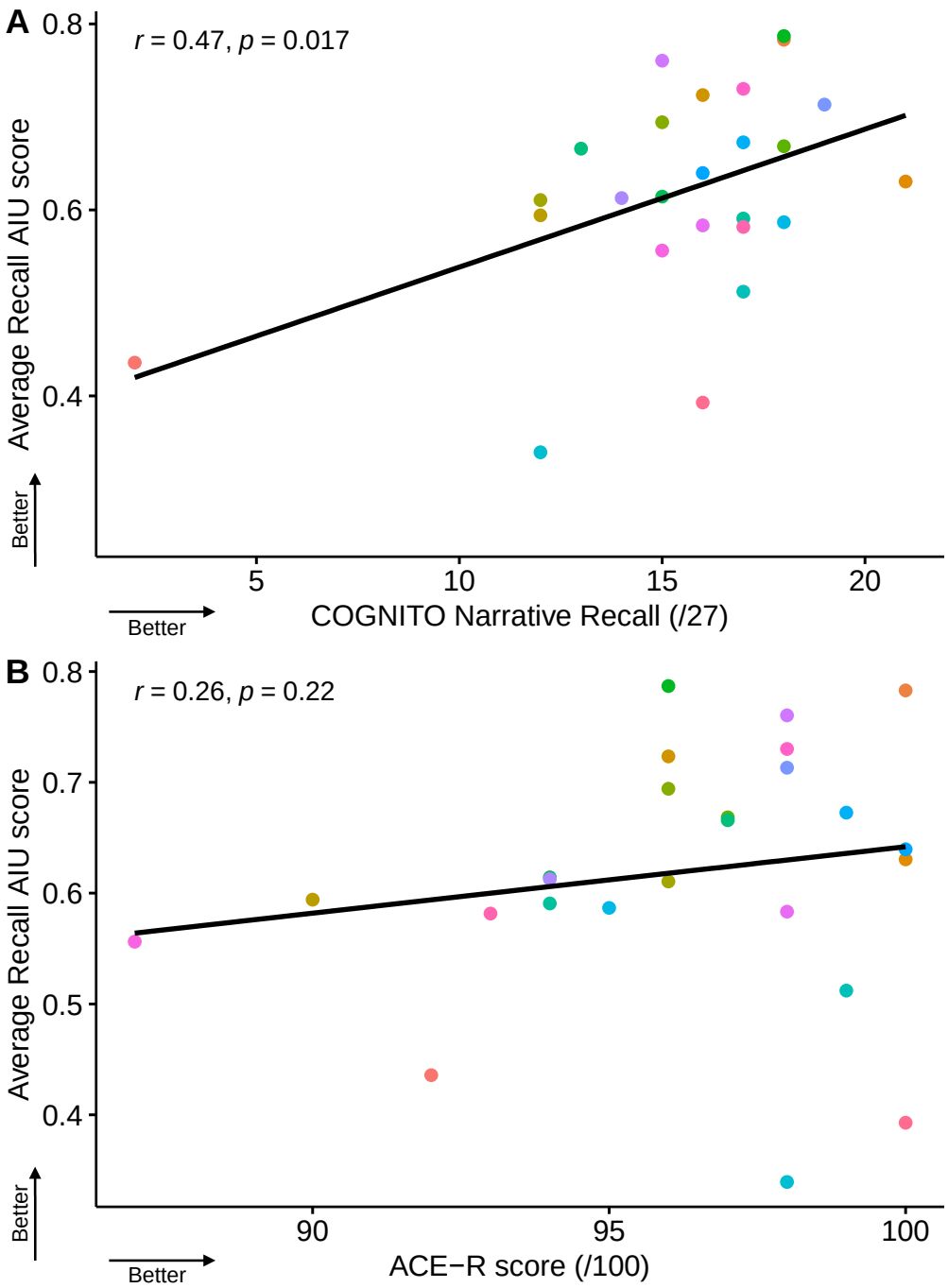


Figure 6.11 Scatterplot shows associations between AIU recall scores and in-clinic tests (A) COGNITO and (B) ACE-R from the PREVENT dataset. Scores are shown in different colours to represent participants.

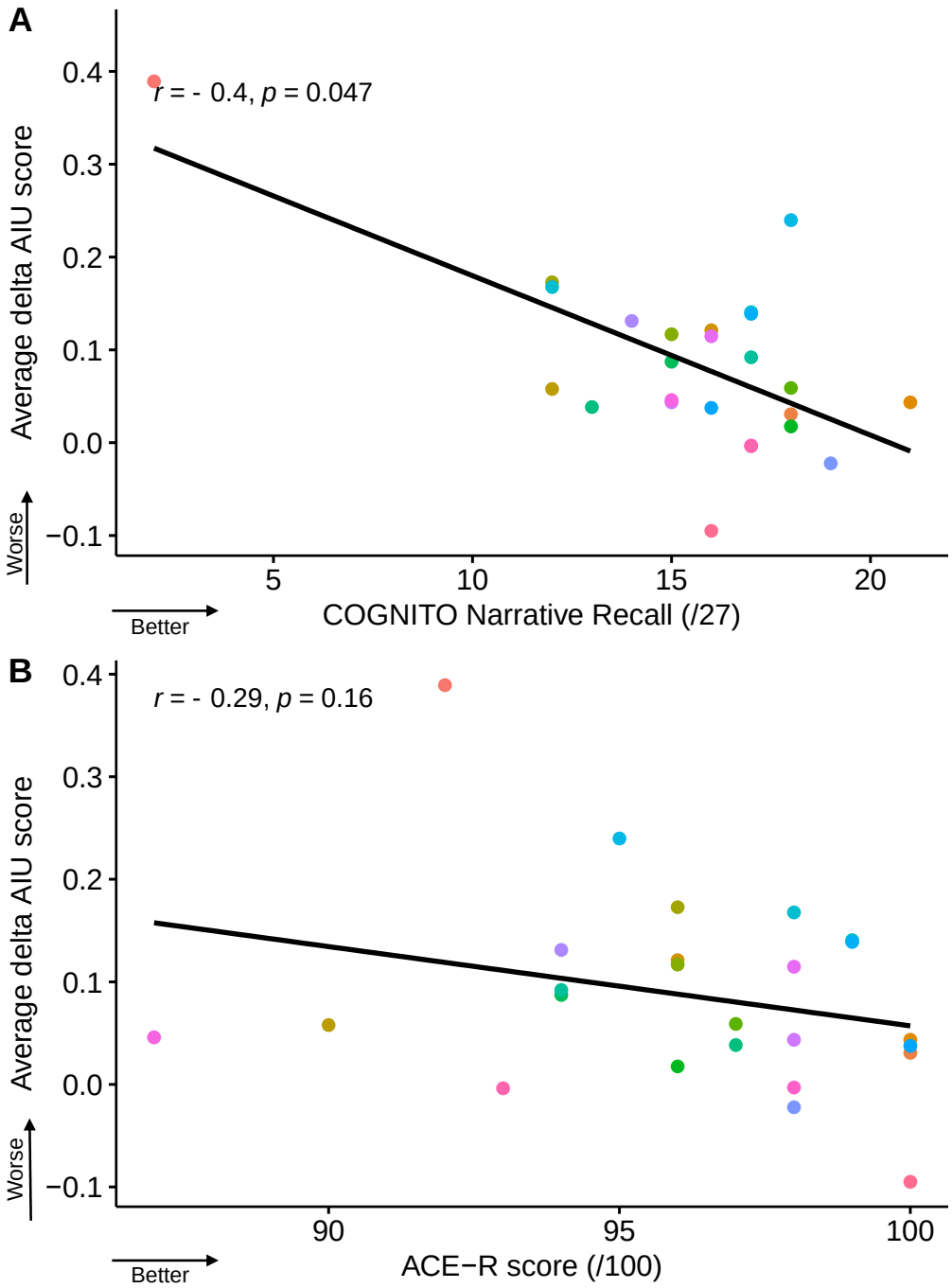


Figure 6.12 Scatterplot shows associations between delta AIU and in-clinic tests (A) COGNITO and (B) ACE-R from the PREVENT dataset. Scores are shown in different colours to represent participants.

6.3.8.2 Associations of AIU scores with age and education

The distribution of AIU recall and delta AIU scores were also examined in each age category (40-44, 45-49, 50-54, 55-59) and years of education category (<15 and 15+) which is shown

in **Figure A.5**, **Figure A.6**. **Figure A.7**, and **Figure A.8**. There was no association between average recall AIU score and age (r = -0.05, p = 0.81). The correlation between average delta AIU score and age was in the expected direction (r = 0.37, p = 0.07), suggesting that with increasing age, increased deterioration in memory after delay as measured with delta AIU score (**Figure 6.13**).

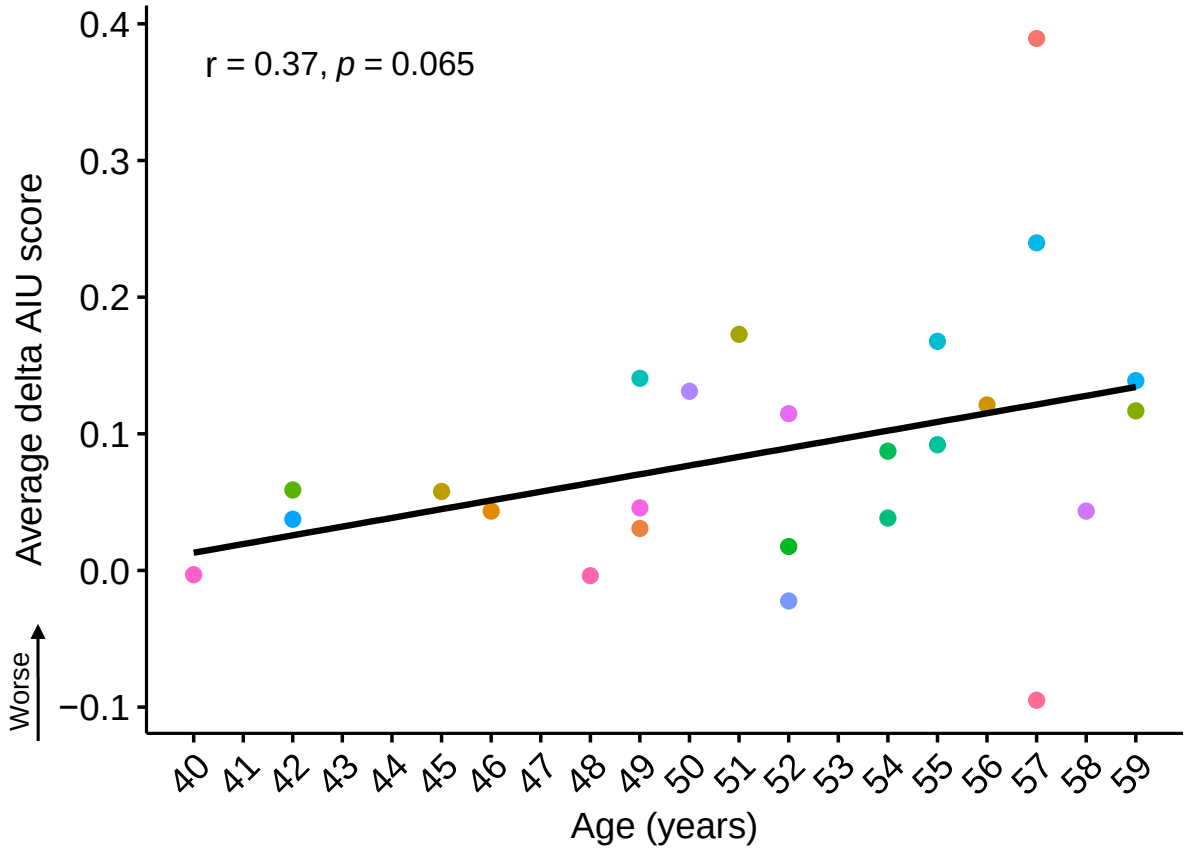


Figure 6.13 Scatterplot showing associations between participants' age and average delta AIU. Scores are shown in different colours to represent participants.

6.4 Discussion

The main goal of this pilot study was to evaluate methods used in analysing speech recordings collected using the Story Time task. The presented analyses investigated Story Time performance in a sample of middle-aged adults (40-59 years), where a high proportion of enrolled participants had a family history of dementia.

6.4.1 Overall findings

Participants contributed to the research from the home environment by engaging with the Story Time task (compliance: Mean = 90.2%, SD = 14.9), making this task a feasible method of data collection. Although the age range among the participant was small (40-59), participants with older ages showed more commitment to the task. The older participants in the study may be more aware that the risk of developing AD (72% of the participants with a family history of dementia) and other forms of dementia increases as people get older.

Evaluation of manually extracted information units demonstrated a reduction in recall after a delay, as evidenced by previous research where the score was higher in immediate recall compared to delayed [217], [224] supporting the utility of the Story Time task as a measure of episodic memory. On a few occasions, participants recalled more at the delayed recall than at the immediate recall. Consistency of storytelling does not necessarily reflect episodic memory, as people may consolidate and refine their newly learned stories with retelling [299]. The participant's performance differed by stories, where participants tended to produce fewer information units in the delayed recall of the story, which contained more information in the comic stimuli. Those stories required greater cognitive demands than others, showing that the participants have difficulties encoding lots of new information and consolidating episodic memory. Even in this population of healthy adults, retelling some stories was cognitively demanding, thus suggesting story retelling tasks might provide subtle differences in scores between healthy controls and potential pre-clinical populations.

The ability to automate the process of scoring is critical to scalability. Initially, the AIU score showed the influence of the poor quality of transcripts (obtained from Kaldi ASR with a higher word error rate), which has also been reported by other studies [209], [237]. In the refined

method, the AIU score extracted from transcripts produced by the Microsoft Azure ASR system demonstrated a significant correlation ($r = 0.98$) and an agreement with the MIU score, suggesting that a lower word error rate is important for the extraction of the AIU score. This result is better compared to the correlation of 0.89 between manual and n-gram based scoring methods by Hakkani-Tur et al. (2010) [300]. Since the last study, the rapid improvement of ASR and NLP methods might have contributed to the higher correlation obtained in this study. This result supported the hypothesis that automation has the potential to replace manual approaches.

Participants produced fewer information units (both MIU and AIU scores) in delayed compared to immediate recall. A similar effect of delay on memory was also seen in the study by Dunn et al. (2002), where both healthy and impaired participants obtained fewer story IUs in the delayed recall task compared to the immediate recall [239]. This suggests that AIU can be used as a measure of memory function.

Reduced speech fluency is known to be associated with multiple cognitive deficits, such as lexical, semantic, and executive function deficits [301]. In this healthy small PREVENT cohort, fluency-related features, including speech rate and hesitation ratio, did not show any significant association with AIU recall scores, which may be due to the small sample size. There was a significant negative association between recall AIU score and average word duration. In a study by Yeung et al. (2021), an increase in average word duration was associated with incoherent speech [302]. The negative association between the lower average word and higher AIU score suggested that coherent speech is more informative.

The association between AIU and in-clinic COGNITO narrative recall task was moderately strong, demonstrating that Story Time offers potential for scalable and sensitive cognitive assessment. Although the influence of age and education on most neuropsychological tests is well-established, showing a positive association of cognitive scores with age and years of education [303], [304], the COGNITO, ACE-R, and AIU recall scores in this sample showed a weak association, which may be due to the limited number of participants in PREVENT study (n = 25) and the very limited age range (40-59). On the other hand, age showed a positive correlation with the average delta AIU score, which could suggest a possible loss of information units with higher age. This result suggests that there might be value in having an extended schedule to capture the memory difference to monitor the participant's performance.

6.4.2 Strengths and limitations

The presented analyses of the evaluation of speech-based measures from the Story Time task in PREVENT have several strengths. This includes investigating audio and language measures derived using natural language and audio signal processing techniques. The smaller sample size allowed manual extraction of the semantic and memory-based features from the speech samples, essential in accessing the quality of automatically extracted scores. Moreover, the manual and automated scores highly correlated with each other, showing the feasibility of extracting these measures automatically from the larger population. The in-clinic measures in PREVENT allowed investigation of the construct validity of the Story Time task where the association between the Story Time measures and the in-clinic measures in the anticipated direction.

While the results show promise, it is worth noting that the small sample size in the PREVENT study limited the power of the analyses. Therefore, the results presented in this chapter are

preliminary and suggestive rather than conclusive. The other limitation of this study is that the Story Time task was optional. Some participants only started the Story Time task during the six-month assessment period due to the operational reason in the trial administration. Finally, the participants who were enrolled in this study showed a ceiling effect (Mean = 96.20, out of 100) in the ACE-R scores. They were highly educated and aware of research with a high commitment to the dementia study, which may have influenced the generalizability of the current findings. It is also necessary to correctly characterise acceptability affected by socio-demographic variables in a large dataset. In the small PREVENT dataset, participant's performance on the Story Time task showed no statistical significance on age, gender, and years of education, suggesting further exploration of these variables in a large dataset.

6.5 Summary

The results support the design of this novel, remote, and smartphone-based assessment, showing the value of testing audio and language-based measures on participants with no diagnosis of cognitive impairment. The agreement between the manual and the automated method and the high correlation between the scores justified the usability of the computerised method instead of the laborious and time-consuming manual approach for information unit extraction from the participant's speech elicitation. The automated measures were consistent with clinical parameters providing the preliminary validation of Story Time, a self-administered smartphone task, measuring long-term episodic memory, semantic memory, audio, and language functions suitable to be deployed in a larger population. In the next chapter, Chapter 7, these features and methods will be evaluated in a large-scale GameChanger study.

Chapter 7. Evaluating methods in GameChanger: a population-based study

7.1 Introduction

In this chapter, the automated and scalable feature extraction methods developed in Chapters 5 and 6 will be evaluated on the GameChanger cohort. The sub-tasks within the Story Time task have different roles. The learning sub-task, similar to picture description, measures the retrieval ability of appropriate lexical items from visual stimuli [183], whereas recall sub-tasks provide the measure of episodic memory [252]. This chapter aims to investigate the compliance and completeness with the task schedule, the attrition (dropouts) of the participants over time, and the impact of socio-demographic factors on task engagement. The further objective is to investigate the influence of regional accents on AIU scores and the relationship between AIU scores and remaining pipeline features. In addition, the association of the Story Time pipeline features and two other Mezurio tasks (section 4.2.4), Gallery Game and Subjective Cognitive Decline (SCD) scores, are evaluated. Finally, the association of socio-demographic variables, including age, education, family history of dementia, and deprivation, with Story Time pipeline features are evaluated. These socio-demographic variables are known to be associated with the risk of dementia and Alzheimer's disease [304], [305]. In addition, associations are assessed using only audio features since they could be readily applied to a non-English cohort.

7.2 Methods

7.2.1 GameChanger

This work used the GameChanger dataset, described in detail in Chapter 4. GameChanger is a citizen science study with over 18000 UK-based participants. Participants enrolled in this study were asked to complete various cognitive tasks and self-report questionnaires within the Mezurio app. Section 4.2 described the study design, participant recruitment methods, smartphone-based assessments, data availability and baseline characteristics in detail. Compliance with the study schedule was evaluated on all the GameChanger participants. In subsequent analyses, only participants who provided Story Time audio recordings were included.

7.2.2 Digital features

The feature extraction pipeline developed in Chapter 5 was applied to the Story Time data of GameChanger participants to extract speech-based features. This chapter focuses on analysing these audio and language features along with the outcomes of two other measures: Subjective Cognitive Decline (described in section 4.2.3) and Mezurio Gallery Game (described in section 4.2.4) (**Table 7.1**).

Table 7.1 Definitions of key outcomes from Mezurio, Story Time, Gallery Game, and Subjective Cognitive Decline (daily and six months).

	Outcomes	**Definitions**	**Analysis Level**
Mezurio	Compliance	Compliance is calculated for each participant as the proportion of task schedules they attempted up to the moment, the participant leaves the study	P
Story Time	Completion	Successful completion is one with an audio recording with at least five words in transcription output	P, S
Pipeline features	Semantic features Lexical features Syntactic features Psycholinguistic features Speech fluency features Spectral and voice-quality features	Please see **Table 6.1** for detailed definitions of each measure under lexical, syntactic, psycholinguistic, semantic, speech fluency, and spectral and voice quality features.	P, S, R
Gallery Game	Recall Accuracy	Proportion of correct recall trials	P
Subjective Cognitive Decline (SCD)	Subjective Cognitive Decline score (daily)	Average of subjective cognitive decline score recorded for eight consecutive days	P
	Subjective Cognitive Decline score (yearly)	Average of subjective cognitive decline score recorded at each assessment periods	P

Note: Analysis level is represented as P - Participation Level, S - Story Level, and R- Repetition Level

This approach involves a large number of statistical tests (about 50 pipeline features); therefore, the significance of results is interpreted using an approximate Bonferroni corrected p-value ($\frac{0.05}{50}$) to allow for multiple testing ($\alpha < .001$).

7.2.3 Selection of confounders

The baseline variables, including age, sex, and education, were identified as potential confounders of the association of speech-based measures with cognition, as discussed in Section 6.2.4.

7.2.4 Statistical analyses

In this section, the statistical methods that have already been used in Chapter 6 will be briefly mentioned, and new methods used in this chapter will be discussed in detail.

7.2.4.1 Participant's compliance with the task schedule

The feasibility of the Mezurio tasks was assessed by examining the compliance and completeness of the task schedule, as discussed in section 6.2.5. In addition, the mean values and proportions of the participant's characteristics were calculated within stages of continued participation in the Story Time task within the Mezurio schedule, adjusted for age (5-year age group).

7.2.4.2 Influence of regional accents on AIU scores

The association of region with AIU scores was investigated using a linear mixed model (as described in Eq. 6.1) with participants and stories as random effects and official regions of England as a fixed effect adjusted for age, sex, education, and story IUs. The nine official regions, including East Midlands, East of England, London, North East, North West, South East, South West, West Midlands, and Yorkshire and the Humber, are used as defined by the UK government [306]. χ^2 value for heterogeneity was calculated on the base model (without region) and the full model as described above.

7.2.4.3 Test-retest reliability of AIU scores

The test-retest reliability of AIU scores (the consistency of AIU scores for story across participants), including learning, immediate, delayed, average of immediate and delayed, and delta AIU scores obtained across different stories from the Story Time sub-tasks in year 1, were estimated separately using the Cronbach alpha coefficient. The average immediate and delayed recall scores provide an estimate of overall memory, considering both immediate and delayed

memory abilities. Delta scores measure the loss of information in storytelling ability between immediate and delayed recalls, estimated by subtracting delayed features from immediate features (as discussed in section 6.2.5). To assess how many stories are adequate for good reproducibility, Cronbach alpha was also calculated considering all 7 stories and considering only the first 2, 4, and 6 stories within the Story Time task. The formula to calculate Cronbach alpha coefficient, as stated in Bland and Altman, 1997 [307], is:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_T^2}\right)$$

Where k is the number of stories in the Story Time task, s_i^2 is the variance of the ith story and s_T^2 is the variance of the total score summing all the stories. The acceptable values of alpha reported in the literature range from 0.70 to 0.95 [308].

7.2.4.4 Influence of 24-hr delay and time of the day on AIU recall scores

For analyses examining the effect of 24 hr-delay, the AIU score was regressed using a linear mixed model with participant and story as random variables (as described in Eq. 6.1). The independent variables were age, sex, education, the story IUs, and recall types. Literature shows the fluctuations in cognitive performance at the time of the day at which the test is performed [309], [310]. For analyses examining the effect of time of the day on recall performance, time of the day was coded into four groups: morning (5 am - 12 pm), afternoon (12 pm - 5 pm), evening (5 pm - 9 pm), and night (9 pm - 5 am). Both immediate and delayed AIU scores were regressed separately using linear mixed models with participant and story as random effects and independent variables as age, sex, education, the story IUs, time of the day at which the task was performed, and the variable delay interval between two recalls (only on delayed recall). The beta estimates from the linear mixed model were plotted for visual analysis with the model fit and 95% confidence interval. χ^2 value for heterogeneity and trend were calculated.

7.2.4.5 Associations of AIU recall scores with audio and language features

For analyses examining the association of AIU recall score with pipeline features, AIU recall score was regressed on the pipeline features using linear mixed models (as described in Eq. 6.1) with participant and story as random variables adjusted for age, sex, education, the story IUs, and recall type (immediate and delayed). The standard deviation for each pipeline feature was calculated from the immediate recall values and multiplied with the model estimates to obtain estimates per SD. The combination of remaining pipeline features, most strongly predicting AIU recall scores, was investigated using stepwise regression with a significance level for entry and exit of 0.001 (Bonferroni correction for multiple testing of 50 features). Further sensitivity analyses were performed restricting to audio features only.

7.2.4.6 Association of pipeline features with other Mezurio scores

The Mezurio Gallery Game task (see section 4.2.4 for more detail) was scheduled for fourteen days, with retention intervals between learning and memory test ranging from 24-hr to 13 days. To compare Gallery Game scores with Story Time, tasks with a 24-hr delay interval between learning and recall, similar to Story Time's immediate and delayed interval, were considered. Gallery Game recall accuracy scores for those days were averaged to get an average score per participant. Spearman's rank correlation coefficient was used to investigate the relationship between Gallery Game recall scores and recall and delta AIU scores separately. In addition, these Gallery Game recall scores were grouped into quartile scores, and their relationships with the mean recall and delta AIU score were plotted to visualise the shape of the relationship.

Participants' subjective cognitive decline (SCD) questionnaire responses per module (daily and yearly) were scored (total score of 24) with equal weight on each category of questions (see **Table 4.3**). Spearman's rank correlation coefficient was used to investigate the relationship between SCD scores and recall and delta AIU scores separately. These SCD scores were further

grouped into quartile scores, and their relationship with the mean recall and delta AIU score were plotted to visualise the shape of the relationship.

The combination of Story Time pipeline features, most strongly predicting Gallery Game and SCD scores, was investigated using stepwise regression with a significance level for entry and exit of 0.001 (Bonferroni correction for multiple testing of 50 features). Further sensitivity analyses were performed restricting to audio features in these models.

7.2.4.7 Relationships between participant's socio-demographic factors and Story Time pipeline features

The association of socio-demographic factors (age, education, family history of dementia, and deprivation) with each Story Time pipeline features extracted from recalls and their delta were evaluated univariately using a linear mixed model (as described in Eq. 6.1) with participant and story as random variables. As an attempt to consider possible data reduction, Principal Component Analyses (PCA) on the feature groups (see **Figure 2.2**) were evaluated and the first dimension of the principal component (PC1) was extracted based on it and then evaluated using regression models. Participant's deprivation deciles were coded into most deprived (1-5 decile) and least deprived (6-10 decile) groups.

All statistical analyses were conducted in R version 4.1.0 (R Foundation for Statistical Computing, Vienna, Austria). Age-adjusted analyses were conducted using the 'DirectStandardisation' package, linear mixed effect model analyses were performed using the 'lmer' package, Cronbach alpha was calculated using the 'ltm' package, and principal component analysis was performed using the 'princomp' function within the 'stats' package. P-values for linear mixed models were estimated using the 'lmerTest' package [295], as discussed in Chapter 6. All the plots were produced using R packages - 'sjPlot', 'ggplot2', and 'ggpubr'.

7.3 Results

7.3.1 Participant's compliance with the Mezurio schedule

The temporal schedule of Mezurio influenced task commencement of the Story Time (59.8%), which appears later in the 30 days study schedule (see **Figure A.9**) compared to SCD (83.1%) and Gallery Game (86.4%) (**Table 7.2**). Participant's compliance was calculated as the proportion of the task schedules they were active in up to the moment they left the study. Learning and recall subtasks within Gallery Game and Story Time are based on stimulus sets. If someone missed the stimulus for the learning subtask, they are not offered recall for that stimulus. Participants completed a smaller proportion of the tasks while they were active in the task (Year-1 compliance: Story Time 67.1%, Gallery Game 72.3%, and SCD (Daily) 82.6%). 497 participants (out of 11239) opened the Story Time task in Year-1 but did not provide a single audio recording, which may have contributed to the lower compliance for this task. Among 4462 respondents in Year-2, the task uptake and compliance were higher for both Story Time and Gallery Game tasks compared to Year-1, showing a follow-up by a smaller number of more engaged participants in Year-2.

Table 7.2 The mean compliance, task uptake, and the number of participants on each GameChanger Mezurio task across Year-1 and Year-2 assessment periods.

	Number of Participants [a] (n)		Task commencement [b] (%)		Mean compliance [c] (%)	
Mezurio Tasks	Year-1	Year-2	Year-1	Year-2	Year-1	Year-2
SCD (Yearly)	14586	3980	77.5	89.2	100.0	100.0
Gallery Game	16250	4128	86.4	92.5	72.3	75.4
Story Time	11239[d]	3517	59.8	78.8	67.1	71.2
SCD (Daily)	15628	3223	83.1	72.2	82.6	83.6

[a] These are the number of participants who provided data in the Mezurio tasks listed in the table; [b] Task commencement is calculated based on the number of respondents to any of the Mezurio tasks in Year-1 (n = 18789) and Year-2 (n = 4462) (refer Figure 4.3); [c] Compliance is calculated for each participant as the proportion of tasks they opened up to the moment they left the study; [d] Includes participants who opened the task but did not provide any audio data; Abbreviation: SCD - Subjective Cognitive Decline.

Commitment to the Mezurio schedule was evaluated based on their involvement in Story Time learning sub-tasks in Year-1 and Year-2. **Table 7.3** provides characteristics based on those individuals with no Story Time data, those who dropped out before the last story in the Story Time task in that assessment period and those who continued to other Mezurio tasks available on the 30th day GameChanger schedule. Older participants (60-95) had a higher commitment to the Mezurio schedule, with a smaller proportion (8.5%) dropping out during the Story Time task and higher proportions (65.6%) continuing beyond the Story Time task compared to other age groups (overall 11.3% dropped out during Story Time and 46.3% continued further to other tasks). There was a similar pattern of dropouts and continuation in Year-2. The proportions of dropouts and continuation of the study schedule did not vary by any other demographic variables in either year.

Table 7.3 Characteristics of participants on their commitment to Mezurio schedule based on the audio recordings available for the Story Time learning sub-tasks during Year-1 and Year-2 assessment.

	Year-1				Year-2			
Characteristics [a]	**Overall** [b]	**No Story Time** [c]	**Dropped During Story Time** [d]	**Continued beyond Story Time** [e]	**Overall**	**No Story Time** [c]	**Dropped During Story Time** [d]	**Continued beyond Story Time** [e]
n (%)	18642 (100)	7900 (42.4)	2109 (11.3)	8633 (46.3)	4446 (100)	935 (21.0)	229 (5.2)	3282 (73.8)
Age (years), mean (SD)	50.2 (14.2)	45.1 (14.5)	49.1 (12.8)	54.9 (12.5)	54.9 (12.2)	50.3 (13.1)	49.6 (12.2)	56.7 (11.4)
Age Group n (%)								
16-39	4307 (24.1)	2776 (64.0)	484 (11.3)	1047 (24.7)	529 (12.6)	206 (38.1)	44 (8.6)	279 (53.3)
40-59	8558 (47.8)	3350 (38.0)	1123 (13.1)	4085 (49.0)	2017 (48.1)	433 (20.8)	124 (5.9)	1460 (73.3)
60-95	5032 (28.1)	1313 (26.0)	437 (8.5)	3282 (65.6)	1648 (39.3)	228 (14.0)	48 (2.8)	1372 (83.2)
Sex n (%)								
Female	14505 (77.9)	5866 (39.1)	1708 (11.7)	6931 (49.2)	3536 (79.6)	717 (19.7)	188 (5.2)	2631 (75.1)
Male	4120 (22.1)	2023 (50.2)	399 (10.1)	1698 (39.7)	908 (20.4)	218 (25.0)	41 (5.1)	649 (69.8)
Education n (%)								
Higher	7909 (42.6)	3241 (37.7)	888 (11.1)	3780 (51.2)	2094 (47.2)	414 (18.4)	98 (4.8)	1582 (76.8)
Lower	10667 (57.4)	4617 (44.7)	1217 (11.4)	4833 (43.9)	2345 (52.8)	520 (22.9)	131 (5.9)	1694 (71.3)
Ethnicity n (%)								
Non-White	1228 (6.6)	635 (45.3)	142 (11.4)	451 (43.2)	189 (4.3)	40 (19.5)	11 (6.7)	138 (73.8)
White	17253 (93.4)	7178 (41.3)	1954 (11.4)	8121 (47.3)	4244 (95.7)	893 (20.8)	218 (5.1)	3313 (74.1)
Family history of dementia n (%)								
No	10130 (54.5)	4680 (42.9)	1141 (11.4)	4309 (45.6)	2169 (48.8)	494 (20.8)	129 (5.7)	1546 (73.5)
Not Sure	684 (3.7)	314 (44.6)	74 (10.8)	296 (44.6)	151 (3.4)	43 (26.9)	5 (2.7)	103 (70.5)
Yes	7788 (41.9)	2881 (40.3)	891 (11.4)	4016 (48.3)	2126 (47.8)	398 (20.7)	95 (4.6)	1633 (74.8)
Neurological Condition n (%)								
No	18184 (97.8)	7702 (41.5)	2053 (11.4)	8429 (47.1)	4350 (97.9)	913 (20.6)	221 (5.1)	3216 (74.3)

	Year-1				Year-2			
Characteristics [a]	**Overall [b]**	**No Story Time [c]**	**Dropped During Story Time [d]**	**Continued beyond Story Time [e]**	**Overall**	**No Story Time [c]**	**Dropped During Story Time [d]**	**Continued beyond Story Time [e]**
Yes	417 (2.2)	173 (44.0)	53 (12.3)	191 (43.7)	92 (2.1)	20 (23.6)	8 (9.5)	64 (66.9)
Psychological Condition n (%)								
No	13468 (72.6)	5292 (40.0)	1495 (11.4)	6681 (48.6)	3447 (77.8)	682 (20.1)	154 (4.6)	2611 (75.3)
Yes	5091 (27.4)	2570 (45.6)	606 (11.9)	1915 (42.6)	984 (22.2)	252 (23.0)	74 (7.0)	658 (70.0)
IMD Decile n (%)								
1-2 (Most deprived)	1353 (9.0)	669 (44.5)	157 (11.5)	527 (44.0)	241 (6.4)	70 (25.8)	18 (5.5)	153 (68.6)
3-4	2192 (14.6)	1037 (44.4)	242 (10.8)	913 (44.8)	447 (11.9)	105 (21.4)	28 (6.4)	314 (72.2)
5-6	3177 (21.1)	1349 (40.9)	359 (11.2)	1469 (47.9)	750 (20.0)	168 (20.9)	25 (3.4)	557 (75.7)
7-8	3700 (24.6)	1429 (38.7)	423 (11.7)	1848 (49.7)	965 (25.8)	181 (19.1)	49 (5.0)	735 (75.9)
9-10 (Least deprived)	4623 (30.7)	1680 (37.9)	516 (11.4)	2427 (50.7)	1340 (35.8)	243 (18.9)	79 (6.2)	1018 (74.8)

[a] Adjusted for age (5-year group) where appropriate; [b] This table included a subset of participants as proportions of participants who did not provide socio-demographic information were age (2.9%), sex (0.4%), education (0.7%), ethnicity (1.2%), family history of dementia (0.5%), neurological condition (0.5%), psychological condition (0.8%), and deprivation deciles (19.5%); [c] These participants do not have audio recordings associated with Story Time comic [d] Dropped out before the last story in the Story Time task in the assessment period. [e] These participants attempt Story Time tasks and advance to other tasks (Gallery Game remaining schedule, Mood and Fatigue Questionnaire, Daily Subjective Error Questionnaire) in Mezurio 30 days schedule; Abbreviation: IMD - Index of Multiple Deprivation.

The total number of participants completing at least one of the Story Time sub-tasks in Year-1 was 10742 (497 participants did not provide any audio data), and in Year-2 was 3511 (6 participants did not provide any audio data). **Figure 7.1** shows the proportion of participants completing the learning, immediate and delayed recall subtasks among those who started each story, stratified by sub-tasks and the assessment periods (see **Table A.6** for more detail). There was a wide range in the number of participants who fully completed (learning, immediate recall, and delayed recall) stories (**Figure A.10**). The average number of stories with a full set of subtasks in Year-1 was 3.9 (out of 7), in Year-2 was 3.4 (out of 5), and overall was 4.9 (out of 12). On average, older participants (60-95) fully completed more stories (4.6 stories with a full set of the subtasks learning, immediate recall, and delayed recall) than the younger group (16-39: 3.2 stories; 40-59: 3.6 stories) out of 7 stories in Year-1.

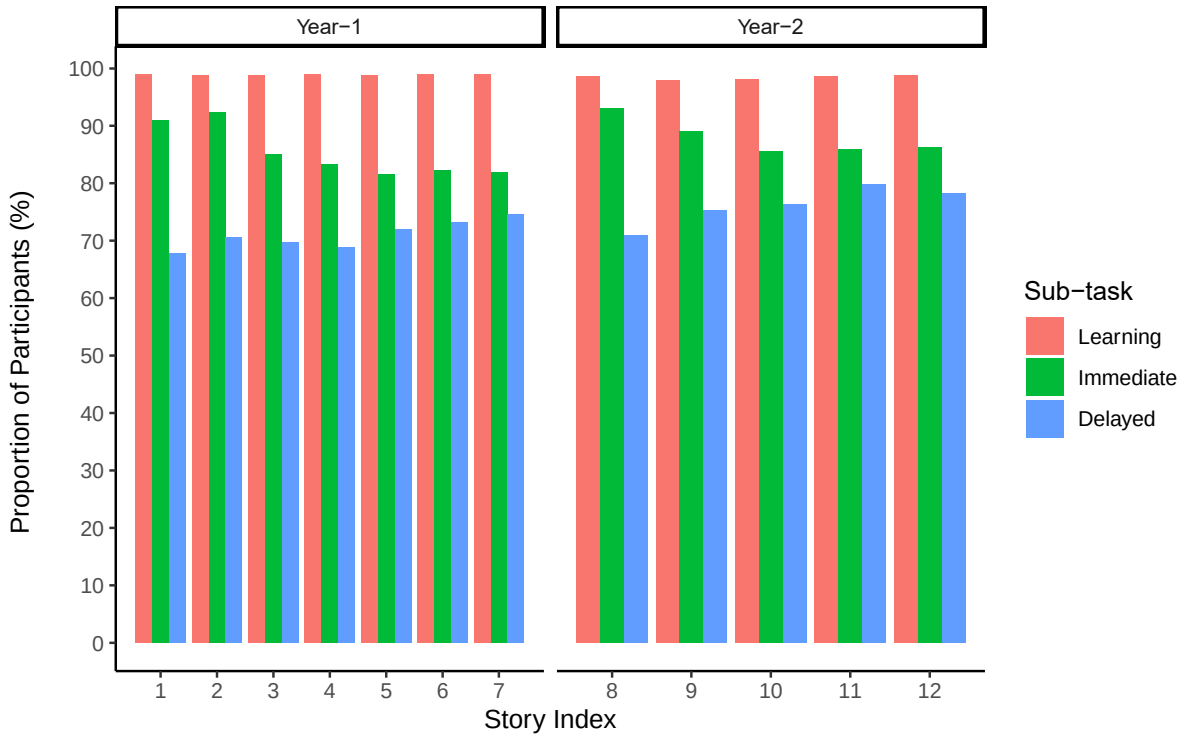


Figure 7.1 Participant's completion of stories by story index stratified by sub-tasks for each story in the GameChanger study. The plot shows the proportion of participants among those who completed the learning task and contributed their speech response to the remaining Story Time sub-tasks (immediate or delayed recall) on Year-1 and Year-2 study schedules. Note that there were seven stories in Year-1 and five stories in Year-2. This figure does not include 4% of the participants in Year-1 and 0.2 % in Year-2 who did not provide audio recordings.

7.3.2 Influence of region on AIU scores

The quality of AIU scores depends on the quality of transcripts. Individuals age, sex, and regional accents are known to produce a large variation in speech [266], which would affect the accuracy of ASR transcripts. To evaluate whether different regional accents influence the AIU scores, the mean AIU scores obtained on the immediate recall of each of England's nine official regions were plotted in the heatmap (**Figure 7.2**).

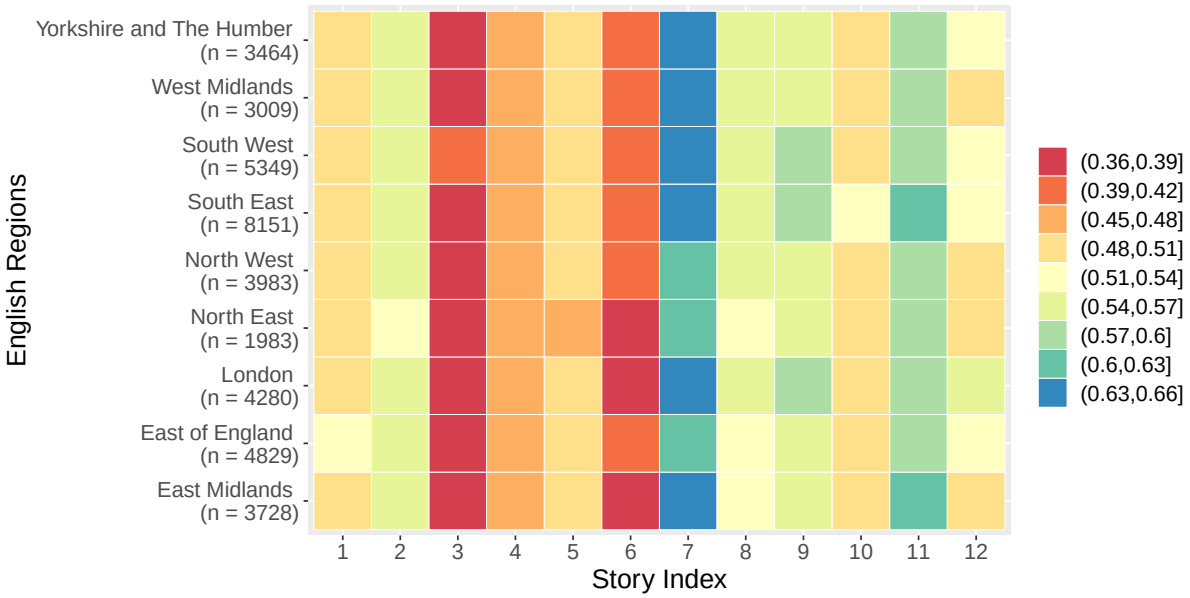


Figure 7.2 The heatmap plot showing AIU immediate recall scores (mean) for twelve Story Time comics by English regions. AIU scores on the immediate recall of Story Time comics were selected because GameChanger participants completed more immediate recalls than delayed ones. The plot represents the mean AIU score rounded to two decimal places. The values in the parenthesis show the number of recordings available for that region.

The visual analyses showed that instead of the influence of region, stories in the Story Time task influenced the AIU score obtained from the immediate recall sub-task. This result suggests that Microsoft Azure's ASR system is robust to the regional accents supporting the result obtained on the manual recall scores in PREVENT study (section 6.3.3), where the influence of the story stimuli was seen in the distribution of immediate and delayed recall scores.

Table 7.4 Influence of English regions on the AIU scores (at immediate recall sub-task). The linear mixed model was used to regress AIU scores on regions with participants and stories as random effects adjusting for age, sex, education and story IUs.

Automated Information Unit (AIU) score [a]	
Predictors (Regions)	**Estimate (SE)**
East of England	0.003 (0.005)
London	-0.002 (0.005)
North East	-0.009 (0.007)
North West	-0.002 (0.005)
South East	0.005 (0.005)
South West	0.005 (0.005)
West Midlands	0.002 (0.006)
Yorkshire and The Humber	-0.000 (0.006)
East Midlands	0[b]
Heterogeneity Test [c]	
$\chi 2$ **Df**	**p**
8.9 8	0.35

[a] AIU scores were obtained from immediate recalls only, excluding participants without regional information, age, sex, and education, which included 8006 participants, 12 stories, and 37644 observations; [b] This parameter is set to zero because it is used as a reference; [c] Heterogeneity test to compare a model without any effect of the region to a model including regions using a likelihood ratio test.

Also, the result from the heterogeneity test between the base model (linear mixed model with participants and story as random effects, adjusting for age, sex, education, and story IUs) and full model, including English regions as a predictor variable on a base model was non-significant, $\chi 2$ (8) = 8.9, p = 0.35. (**Table 7.4**).

7.3.3 Test-retest reliability of AIU scores

Test-retest reliability of AIU scores was evaluated on seven stories presented in the Story Time task during Year-1. Participants with AIU scores for at least two stories were considered. The Cronbach coefficient of reliability (α) was 0.73, 0.74, and 0.70 for learning, immediate recall and delayed recall sub-tasks, respectively, for seven stories in Year-1. In addition to the sub-task-based reliability score, the reliability of average recall AIU scores and delta AIU scores

was evaluated, and the coefficient of reliability obtained was 0.79 and 0.32, respectively. Detailed Cronbach coefficient of reliability (α) estimate considering the first 2, 4, 6 and 7 stories is provided in **Table 7.5**, which showed an increase in the number of stories increased the reliability. With 7 stories, the reliability of the average AIU recall seems to be settling at a satisfactory level. However, considerably more stories would probably be needed for the delta AIU to have a good reliability.

Table 7.5 Cronbach coefficient of reliability (α) estimate on AIU scores considering first 2, 4, 6 and 7 stories in the GameChanger study, Year-1.

	Cronbach coefficient of reliability (α)				
Number of stories	Learning	Immediate	Delayed	Average recall	Delta
2	0.43	0.40	0.37	0.48	0.04
4	0.59	0.59	0.54	0.66	0.14
6	0.70	0.69	0.68	0.77	0.24
7	0.73	0.74	0.70	0.79	0.32

7.3.4 Influence of 24-hr delay and time of the day on AIU recall scores

Similar to the analysis presented in Chapter 6, the association of delay with AIU recall scores was investigated using a linear mixed model with participants and stories as random effects adjusting for covariates, including age, sex, and education (**Figure 7.3**). There was a significant effect of delay on recall scores ($p < 10^{-300}$), where the change in the score between delayed and immediate was -0.08 ± 0.001 (Estimate ± SE). Similarly, there was a significant effect of sex on the AIU score, with male participants producing lower AIU scores compared to female participants.

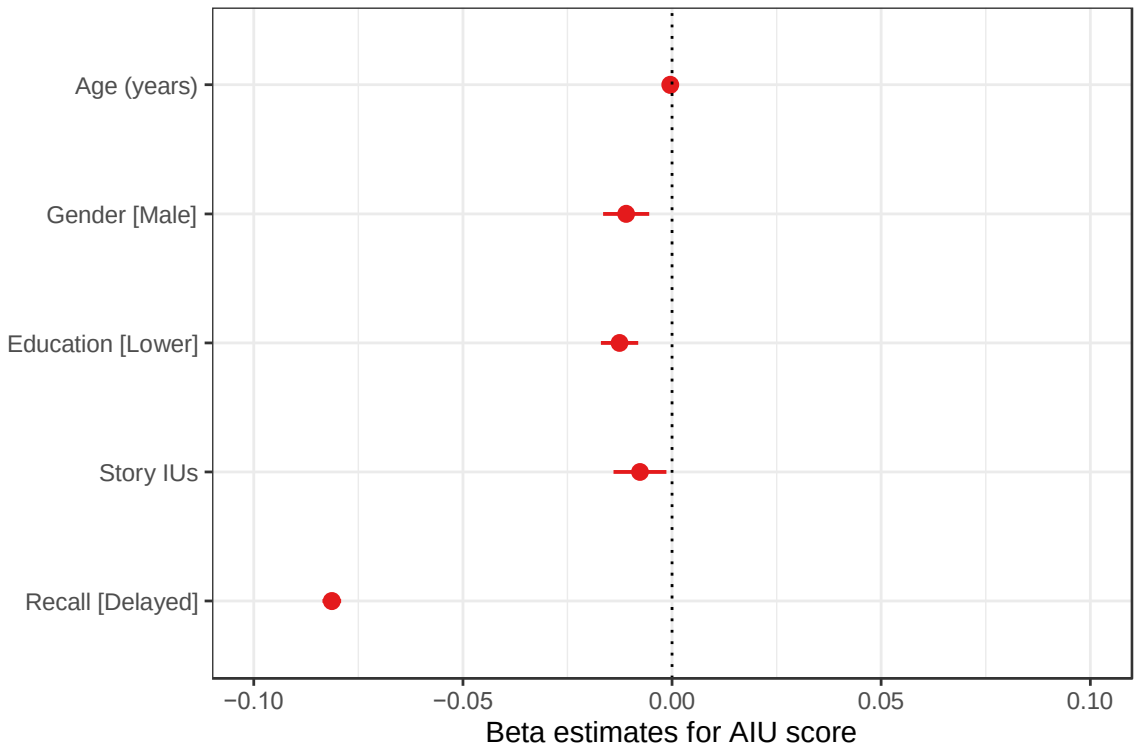


Figure 7.3 Beta estimates obtained by regressing AIU scores at recall on age, gender, education, story IUs and recall type using a linear mixed model with participants and stories as a random effect. The magnitude of the effect is indicated on the x-axis, and whiskers represent 95% CI.

Compared to the in-clinic assessments scheduled for a fixed time and duration, remote digital assessments allowed flexibility for participants to complete the task in their own time. To test the hypothesis that alertness, which is associated with cognition, was higher in the morning and decreased as time progressed [310], the association of AIU score (separately for immediate and delayed) with the time of the day was investigated using linear mixed models with participants and stories as random effects and time of the day as fixed effects adjusted for age, sex, education and story IUs. The model was also adjusted for the variable delay between the recalls for the association of delayed AIU score and the time of the day. (**Figure 7.4**).

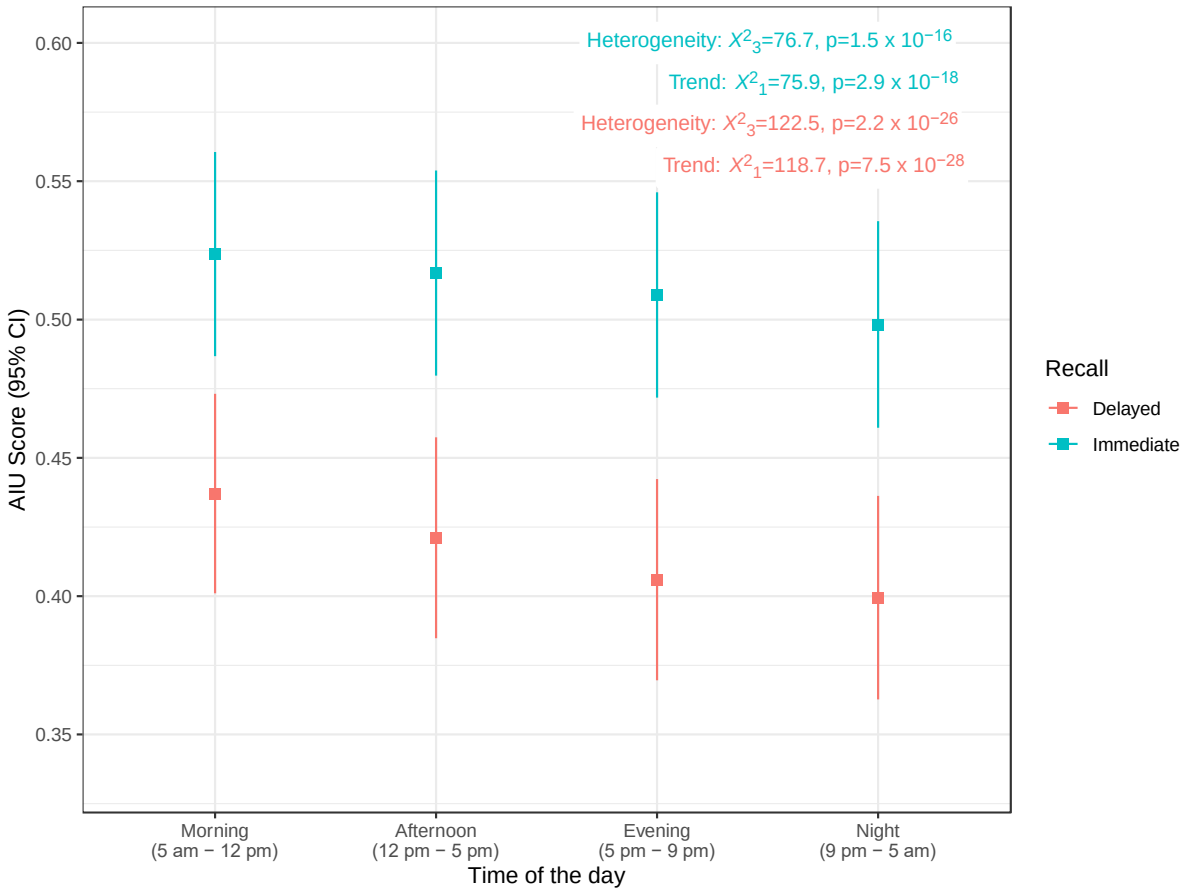


Figure 7.4 The effect of time of the day with AIU scores at immediate and delayed recalls using linear mixed models with participants and stories as random effects and time of the day as a fixed effect adjusted for age, gender, education, and story IU. The model with delayed AIU scores as an outcome variable was also adjusted for the delay interval between immediate and delayed recalls. CI- Confidence Interval; Whiskers represent 95% CI.

The time at which the task was completed showed a significant impact on the participant's recall performances, with steady trends observed in both immediate (p_{trend} = 2.9 x 10^{-18}) and delayed (p_{trend} = 7.5 x 10^{-28}) recalls. Immediate AIU recall scores on tasks completed in the morning were higher than those completed in the afternoon (Estimate = -0.007, SE = 0.002), evening (Estimate = -0.015, SE = 0.002) and night (Estimate = -0.027, SE = 0.003). On delayed recall, there was also a significant decrease in AIU scores as time progressed to the afternoon (Estimate = -0.014, SE = 0.003), evening (Estimate = -0.028, SE = 0.003), and night (Estimate = -0.035, SE = 0.005), compared to tasks completed in the morning.

7.3.5 Exploring associations between AIU recall scores, language, and audio features

The mean (SD) of the pipeline features at each of the three subtasks and the stories completed by the participant is provided in **Table A.7**. The correlation between the pipeline features from the recall task (**Figure A.11**) showed that the AIU score was highly correlated with some of the lexical features (total words, unique words, Brunet's index) and speech fluency features (speech duration, number of syllables, and number of pauses). The correlation between the features within the groups (e.g., lexical, syntactic, speech-fluency, spectral and voice quality features) was typically large (> 0.5 or < -0.5). The correlation of features between the group was mainly small (-0.3 - 0.3). This section evaluates the associations of AIU scores obtained from immediate and delayed recall tasks with remaining speech-based features.

7.3.5.1 Lexical and syntactic features

The joint associations of lexical and syntactic features (please see **Table 6.1** for a brief description and section 2.3.1 for detailed definitions) with AIU recall score are shown in **Table 7.6**. Lexical features, including the number of unique words and Brunet's index, were statistically significant language features associated with AIU scores. A 1 SD unit increase in Brunet's index was associated with a 0.059 unit increase in the AIU recall score. Similarly, a 1 SD unit increase in unique words used was associated with a 0.054 unit increase in AIU recall score. Other lexical and syntactic features also showed statistically significant associations with AIU recall scores.

Table 7.6 Joint association of lexical and syntactic measures with immediate and delayed AIU recall scores.

		Automatic Information Unit (AIU) scores		
Lexical and syntactic features	**SD**	**Estimate (SE)**	**Estimate per SD (SE)**	***p***
Brunets' index	1.03	0.057 (0.001)	0.059 (0.001)	$< 10^{-300}$
Unique words	26.97	0.002 (0.000)	0.054 (0.001)	$< 10^{-300}$
Total repeats	1.10	-0.006 (0.000)	-0.007 (0.001)	1.1×10^{-53}
Pronoun index	0.03	0.229 (0.021)	0.007 (0.001)	1.1×10^{-28}
MATTR	0.02	0.152 (0.019)	0.003 (0.000)	8.8×10^{-15}
Past tense verbs	0.19	0.018 (0.003)	0.003 (0.001)	8.9×10^{-12}
Open class words	0.03	0.059 (0.016)	0.002 (0.000)	0.0001
Proportion of prepositions	0.02	0.109 (0.031)	0.002 (0.000)	0.0003
Verbs per utterances	0.02	0.059 (0.018)	0.001 (0.000)	0.001
Close class words	0.03	0.076 (0.023)	0.002 (0.001)	0.001
Average word length (n)	0.13	0.009 (0.003)	0.001 (0.000)	0.003
Honore's stats	315.93	-0.000 (0.000)	-0.000 (0.000)	0.36

Please see Table 6.1 for definitions of predictors; The model was adjusted for age, sex, and education; The order of features in the table is based on the p-value; Restricted to individuals with complete immediate, delayed recall, age, sex, and education information; $N_{participant}$ = 7291; $N_{stories}$ = 12; $N_{observations}$ = 68708; The likelihood ratio test was used to determine the predictors to be included in the model; MATTR - Moving Average Type Token Ratio; SD - Standard Deviation; SE - Standard Error; SD was calculated from immediate recalls.

7.3.5.2 Psycholinguistic features

The distribution of psycholinguistic features (section 2.3.4) at recall is shown in **Figure A.12**. In **Table 7.7**, specifically, a 1 SD unit increase in imageability was statistically significantly associated with a 0.059 unit higher AIU recall score. Similarly, a 1 SD unit increase in AOA was statistically significantly associated with a 0.012 unit higher AIU recall score. The other psycholinguistic features (concreteness and familiarity) were also statistically significantly associated with AIU scores.

Table 7.7 Joint association of psycholinguistic measures with immediate and delayed AIU recall scores.

	Automated Information Unit (AIU) scores			
Psycholinguistic features	**SD**	**Estimate (SE)**	**Estimate per SD (SE)**	**p**
Imageability	19.68	0.003 (0.0001)	0.059 (0.0019)	3.0×10^{-210}
AOA	0.27	0.045 (0.0017)	0.012 (0.0004)	5.3×10^{-148}
Concreteness	18.79	-0.001 (0.0001)	-0.019 (0.0019)	3.2×10^{-25}
Familiarity	17.30	-0.002 (0.0001)	-0.035 (0.0017)	1.9×10^{-55}

Please see Table 6.1 for definitions of predictors; The model was adjusted for age, sex, and education; The order of features in the table is based on the p-value; Restricted to individuals with complete immediate, delayed recall, age, sex, and education information; $N_{participant}$ = 7291; $N_{stories}$ = 12; $N_{observations}$ = 68708; AOA-Age of Acquisition; SD - Standard Deviation; SE - Standard Error; SD is calculated from immediate recalls.

7.3.5.3 Speech fluency features

The results of joint associations of speech fluency features with AIU recall score showed that fluent speech strongly predicted a higher AIU score (**Table 7.8**). A 1 SD unit increase in speech duration was associated with a 0.041 unit increase in AIU recall score. Similarly, a 1 SD unit increase in average word duration was associated with a 0.032 unit decrease in AIU recall score. Other speech fluency features, including speech rate and average pause duration, also showed statistically significant associations with AIU recall scores.

Table 7.8 Joint association of speech fluency features with immediate and delayed AIU recall scores.

	Automated Information Unit (AIU) score			
Speech fluency features	**SD**	**Estimate (SE)**	**Estimate per SD (SE)**	**p**
Speech duration (sec)	20.66	0.002 (0.000)	0.041 (0.0006)	$< 10^{-300}$
Average word duration (sec)	0.04	-0.793 (0.019)	-0.032 (0.0007)	$< 10^{-300}$
Average pause duration (sec)	0.36	-0.038 (0.002)	-0.014 (0.0007)	2.5×10^{-140}
Speech rate (Syllable/sec)	0.54	0.019 (0.002)	0.010 (0.0011)	2.4×10^{-20}
Hesitation ratio	0.10	-0.022 (0.009)	-0.002 (0.0008)	0.010
Average syllable duration (sec)	0.98	-0.001 (0.000)	-0.001 (0.0005)	0.015
Articulation rate (Syllable/sec)	0.51	-0.002 (0.001)	-0.001 (0.001)	0.174

Please see Table 6.1 for definitions of predictors; The model was adjusted for age, sex, and education; The order of features in the table is based on the p-value; Restricted to individuals with complete immediate, delayed recall, age, sex, and education information; $N_{participant}$ = 7291; $N_{stories}$ = 12; $N_{observations}$ = 68708; The likelihood ratio test was used to determine the predictors to be included in the model; SD - Standard Deviation; SE - Standard Error; SD is calculated from immediate recalls.

7.3.5.4 Spectral and voice quality features

The joint association of spectral and voice quality features with AIU recall scores is shown in **Table 7.9**. A 1 SD unit increase in voice segments was associated with a 0.03 unit increase in AIU recall score. Moreover, many spectral and voice quality features were statistically significantly associated with AIU recall scores.

Table 7.9 Joint association of spectral and voice quality features with immediate and delayed AIU recall scores.

	Automated Information Units (AIU) scores			
Spectral and voice quality features	**SD**	**Estimate (SE)**	**Estimate per SD (SE)**	**p**
Voiced (segments/sec)	0.38	0.0837 (0.0021)	0.0318 (0.0007)	$< 10^{-300}$
Hammarberg index	4.33	-0.0047 (0.0005)	-0.0204 (0.0022)	1.3×10^{-21}
MFCC 3	8.56	-0.0011 (0.0001)	-0.0094 (0.0009)	3.8×10^{-19}
Alpha ratio	3.27	0.0021 (0.0002)	0.0069 (0.0007)	1.4×10^{-18}
MFCC 2	9.40	0.0021 (0.0003)	0.0197 (0.0019)	1.9×10^{-15}
F1 (Hz)	56.48	-0.0003 (0.0000)	-0.0169 (0.0016)	2.0×10^{-15}
F3 (Hz)	86.93	0.0001 (0.0000)	0.0087 (0.0026)	6.7×10^{-9}
F0 (Hz)	3.92	-0.0042 (0.0007)	-0.0165 (0.0027)	2.3×10^{-8}
Energy ratio	0.02	0.0011 (0.0003)	0.0000 (0.0000)	8.1×10^{-6}
RMS energy	0.03	-0.5247 (0.1313)	-0.0157 (0.0039)	6.4×10^{-5}
Loudness (dB)	0.13	0.0871 (0.0255)	0.0113 (0.0033)	6.4×10^{-4}
HNR	1.67	0.0039 (0.0016)	0.0065 (0.0028)	0.013
MFCC 4	7.61	-0.0002 (0.0001)	-0.0015 (0.0008)	0.029
Spectral flux	0.04	0.0688 (0.0399)	0.0028 (0.0016)	0.085
MFCC 1	7.39	0.0005 (0.0003)	0.0037 (0.0022)	0.119
Jitter	0.01	0.1528 (0.1162)	0.0015 (0.0012)	0.188
F2 (Hz)	69.61	0.0000 (0.0000)	0.0028 (0.0023)	0.218
Shimmer	0.15	0.0074 (0.0073)	0.0011 (0.0011)	0.308
Loudness (peaks/sec)	0.54	0.0005 (0.0013)	0.0003 (0.0007)	0.675

Please see Table 6.1 for definitions of predictors; The model was adjusted for age, sex, and education; The order of features in the table is based on the p-value; Restricted to individuals with complete immediate, delayed recall, age, sex, and education information; $N_{participant} = 7291$; $N_{stories} = 12$; $N_{observations} = 68708$; The likelihood ratio test was used to determine the predictors to be included in the model; dB - decibel; HNR - Harmonics to Noise Ratio; Hz - Hertz; MFCC - Mel Frequency Cepstral Coefficient; SD - Standard Deviation; SE - Standard Error; SD is calculated from immediate recalls.

7.3.5.5 Selection of informative pipeline features that explain variation in the AIU score

Stepwise regression analysis was used to select the most relevant pipeline features that explain a significant amount of variation in the AIU score. The features that were selected into the model ($F(15, 7294) = 816.6$, Adjusted $R^2 = 0.63$, $p = 4.9 \times 10^{-58}$) on predicting AIU score are shown in **Table 7.10**.

Table 7.10 Association of the rest of the pipeline features at recalls with AIU Recall score using stepwise regression analysis.

Predictors	Estimate per SD (SE)	p
Brunet's index	0.6130 (0.0156)	$< 10^{-300}$
Close class words	0.2470 (0.0081)	8.2×10^{-191}
Unique words	0.8225 (0.0363)	6.5×10^{-110}
Speech duration	-0.2887 (0.0205)	2.9×10^{-44}
Imageability	-0.3881 (0.0411)	4.8×10^{-21}
Total words	-0.4296 (0.0465)	3.5×10^{-20}
Past tense verb	0.0839 (0.0092)	9.3×10^{-20}
Concreteness	0.3704 (0.0421)	1.7×10^{-18}
Age (years)	-0.0633 (0.0072)	2.1×10^{-18}
MFCC 1	-0.0614 (0.0074)	1.7×10^{-16}
Average word duration (sec)	0.0732 (0.0092)	2.1×10^{-15}
Age of acquisition	-0.0733 (0.0093)	3.8×10^{-15}
Hesitation ratio	-0.0582 (0.0088)	4.4×10^{-11}
Average word length	0.0530 (0.0088)	1.9×10^{-09}
Total repeats	-0.0447 (0.0100)	8.1×10^{-06}

Please see Table 6.1 for definitions of predictors; $N_{participant} = 7291$; The model was adjusted for age, sex, and education; The order of features in the table is based on the p-value; Restricted to individuals with complete immediate, delayed recall, age, sex, and education information; MFCC - Mel Frequency Cepstral Coefficient; SD - Standard Deviation; SE - Standard Error; SD is calculated from immediate recalls.

In addition, when only audio features were included in a stepwise model (**Table 7.11**) to select informative audio features that explain a significant amount of variation in AIU recall score, eighteen audio features were selected in the model ($F(18, 7291) = 160.8$, Adjusted $R^2 = 0.28$, $p = 1.4$ x 10-64). The model explained only 28% of the variation in AIU recall scores compared to 63% when using all the pipeline features, showing the importance of language features.

Table 7.11 Sensitivity analysis on the association of audio features at recalls with AIU recall scores using stepwise regression analysis.

Predictors	**Estimate per SD (SE)**	**p**
Articulation rate (Syllable/sec)	0.3653 (0.0277)	3.6×10^{-39}
Voiced segments	0.1809 (0.0149)	2.2×10^{-33}
Speech rate (Syllable/sec)	-0.4603 (0.0397)	8.7×10^{-31}
Hesitation ratio	-0.3583 (0.0328)	1.4×10^{-27}
Hammarberg index	-0.2431 (0.0267)	1.1×10^{-19}
MFCC 2	0.2243 (0.0278)	9.0×10^{-16}
Speech duration (sec)	0.2072 (0.0276)	6.8×10^{-14}
MFCC 3	-0.0788 (0.0126)	3.7×10^{-10}
Age (years)	-0.0628 (0.0101)	4.9×10^{-10}
Number of pauses	0.1621 (0.0272)	2.6×10^{-09}
Spectral slope	-0.1131 (0.0195)	6.7×10^{-09}
MFCC 1	0.1670 (0.0297)	2.0×10^{-08}
Loudness (peaks/sec)	0.0937 (0.0171)	4.2×10^{-08}
Spectral flux	-0.1114 (0.0209)	4.7×10^{-07}
Alpha ratio	0.0756 (0.0169)	8.9×10^{-06}
Average syllable duration (sec)	0.0424 (0.0110)	1.2×10^{-04}
RMS Energy	0.0867 (0.0239)	2.9×10^{-04}
MFCC 4	-0.0405 (0.0119)	7.0×10^{-04}

Please see Table 6.1 for definitions of predictors; $N_{participant}$ = 7291; The model was adjusted for age, sex, and education; The order of features in the table is based on the p-value; Restricted to individuals with complete immediate, delayed recall, age, sex, and education information; MFCC - Mel Frequency Cepstral Coefficient; RMS - Root Mean Square; SD - Standard Deviation; SE - Standard Error; SD is calculated from immediate recalls.

7.3.6 Associations of pipeline features with other Mezurio scores

This section investigates the relationship between pipeline features and scores from two other tasks within the Mezurio smartphone application: Gallery Game and Subjective Cognitive Decline (SCD) questionnaire.

7.3.6.1 Mezurio Gallery Game scores

The relationship between Story Time AIU recall scores (averaged over stories for a participant) and Gallery Game recall accuracy scores (averaged over games for a participant) (**Figure 7.5**) was in the anticipated direction with a small effect size, r = 0.2, $p = 1.9 \times 10^{-72}$. Similarly, there

was a small negative association between Story Time delta AIU scores and Gallery Game recall accuracy scores, r = -0.17, p = 4.3 x 10^{-48} (**Figure 7.6**).

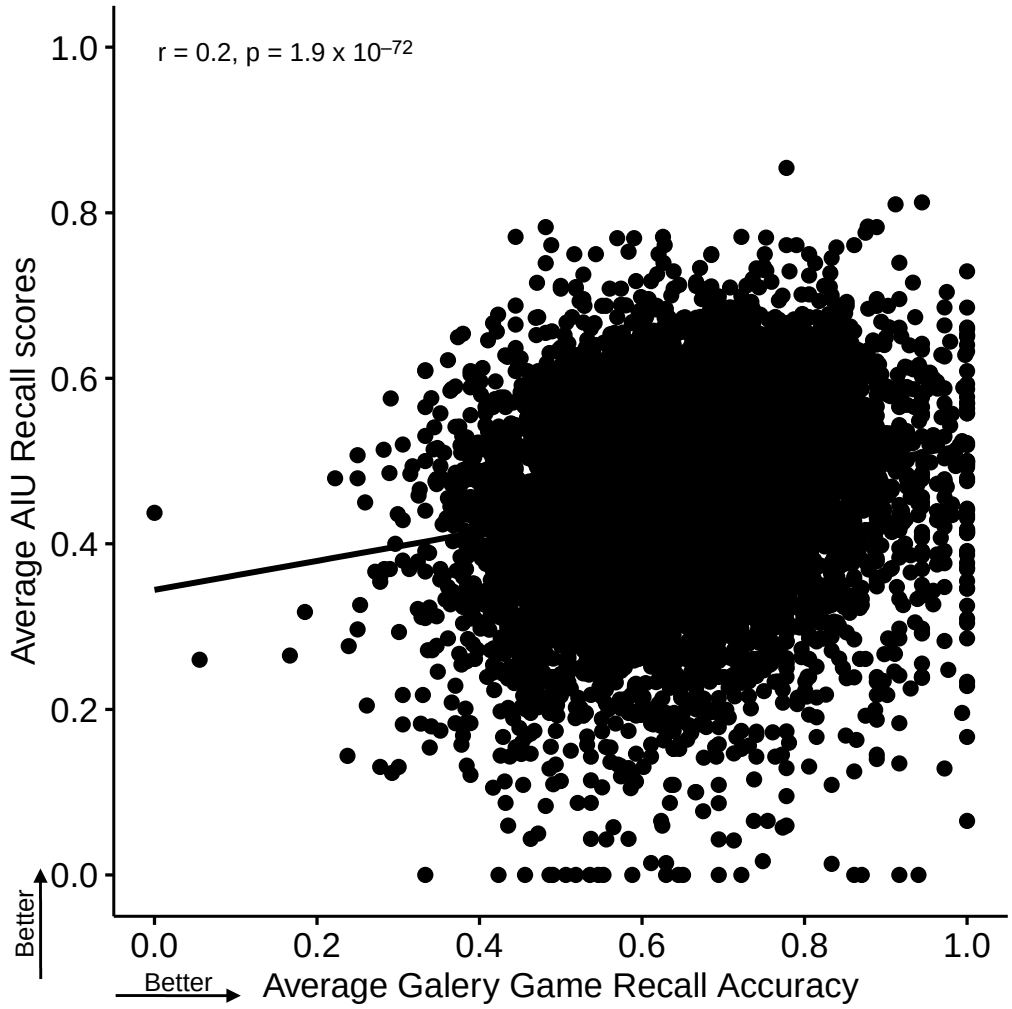


Figure 7.5 Relationship between AIU recall scores from the Story Time task and Mezurio Gallery Game recall accuracy scores and averaged per participant.

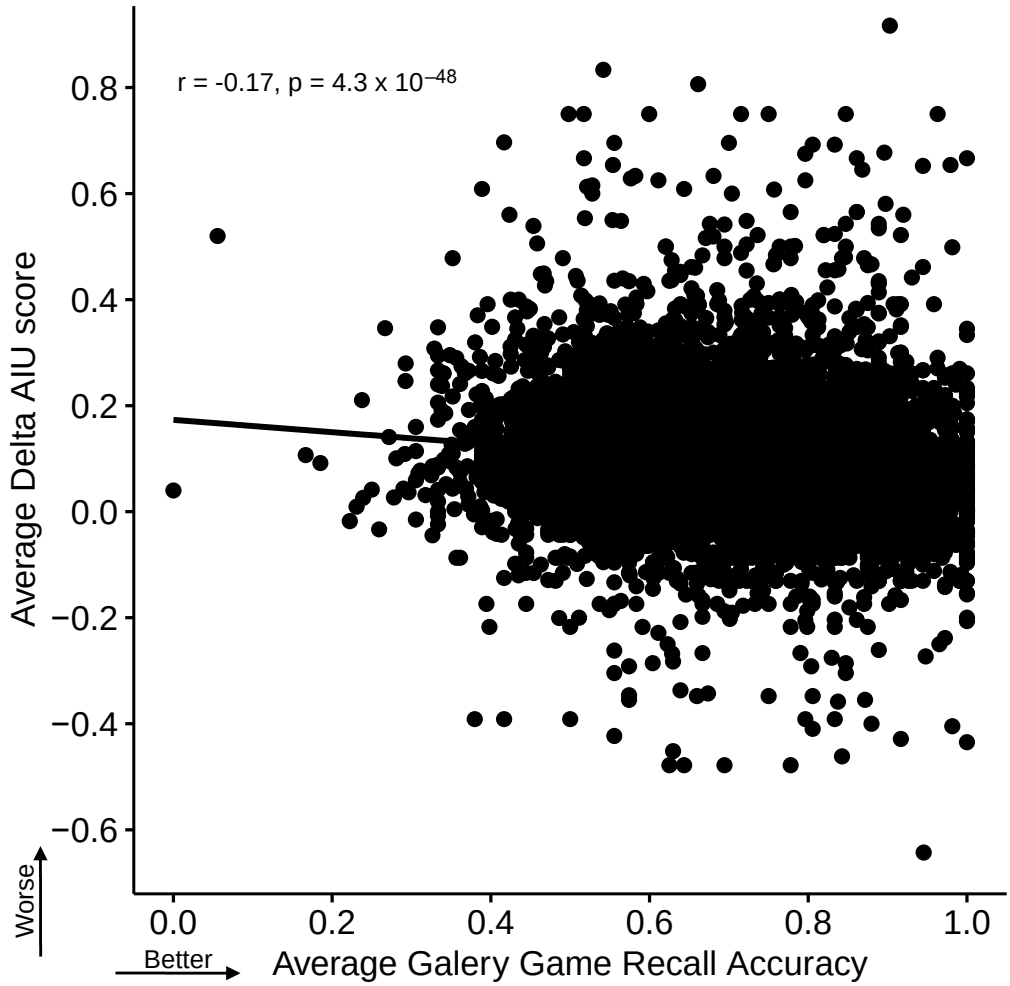


Figure 7.6 Relationship between scores on the Mezurio Gallery Game recall accuracy scores and delta AIU scores from the Story Time task averaged per participant.

The participants in the top quartile of Gallery Game recall accuracy score have higher AIU recall scores than participants in the bottom quartile of Gallery Game recall accuracy score (mean AIU recall scores: top quartile = 0.49, bottom quartile = 0.43) (**Figure 7.7**). Further, the inverse association of the delta AIU score and the Gallery Game recall accuracy score quartile is clearer in **Figure 7.8**.

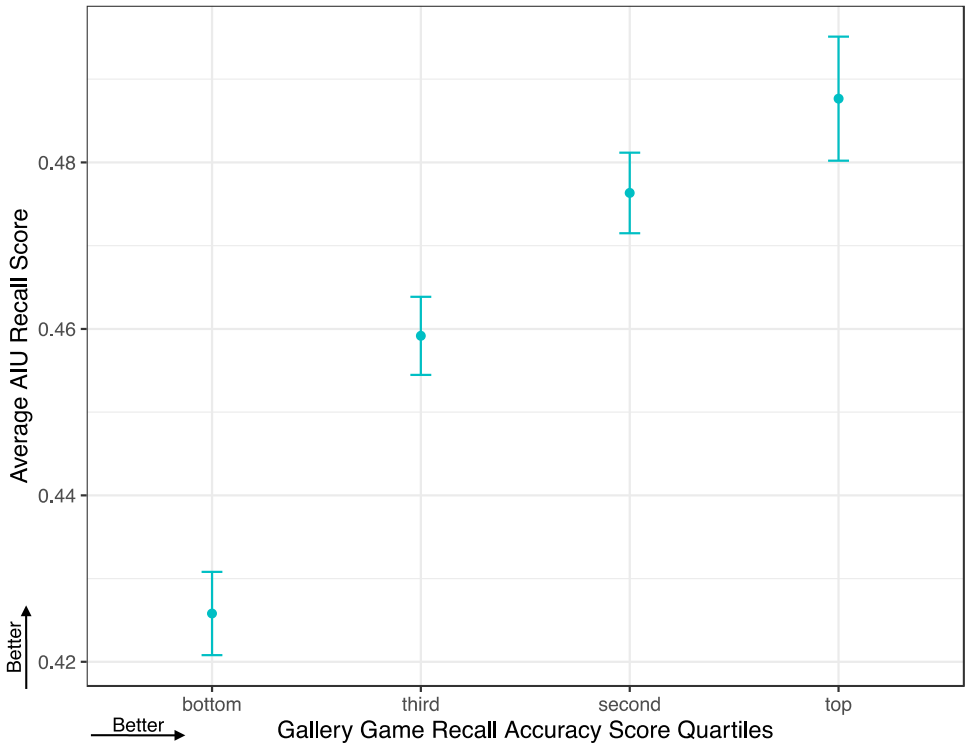


Figure 7.7 Plot showing Mean and 95% CI on average AIU recall scores for Gallery Game recall accuracy score quartiles.

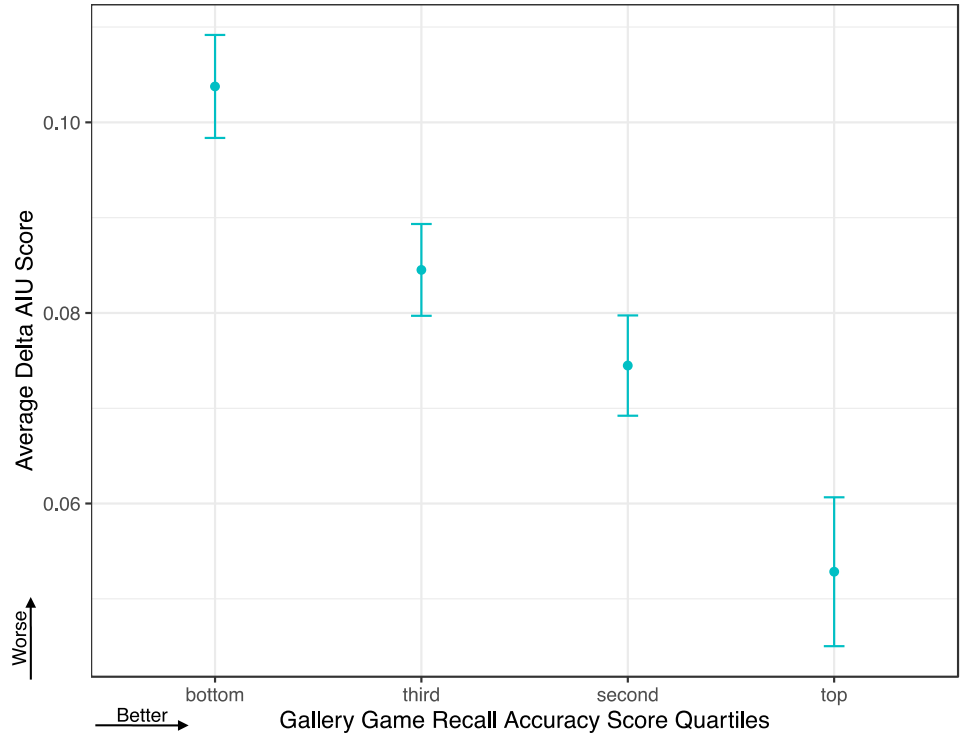


Figure 7.8 Plot showing Mean and 95% CI on average delta AIU scores for Gallery Game recall accuracy score quartiles.

Table 7.12 shows univariate associations of pipeline features and Gallery Game recall accuracy scores. In addition to a higher AIU recall score being associated with a higher Gallery Game recall accuracy score (also seen in **Figure 7.5**), lexical features, including a higher number of total and unique words, were also statistically significantly associated with a higher Gallery Game recall accuracy score with a relationship of similar strength. Other lexical features that were statistically significantly associated with the Gallery Game accuracy score were Brunet's index and total repeats more weakly compared to the AIU recall score. Lexical PC1 was statistically significantly associated with a higher Gallery Game recall accuracy score; however, the strength of the association and the p-value were lower compared to some of the individual lexical pipeline features. Lower use of verbs and determiners and higher use of prepositions were associated with a higher Gallery Game recall accuracy score. The association of syntactic PC1 and Gallery Game recall accuracy score was weaker compared to the use of verbs and other statistically significant predictors among syntactic pipeline features. Speech fluency features showed that the participants who spoke more (higher number of syllables, longer speech duration, and more pauses) and had lower hesitation ratio in their speech response obtained higher Gallery Game recall accuracy scores. Although the speech fluency PC1 was statistically significantly associated with the Gallery Game recall accuracy score, the strength of the association was weaker compared to speech fluency pipeline features (number of syllables, speech duration and number of pauses). Out of nineteen spectral and voice quality features, only MFCC3 showed a statistically significant with the Gallery Game recall accuracy score.

Table 7.12 Association of pipeline features (recalls) with Gallery Game recall accuracy score.

Predictors	Estimate per SD (SE)	p
Semantic features		
AIU score	0.0353 (0.0021)	2.9×10^{-60}
Information density	-0.0132 (0.0019)	3.9×10^{-11}
Lexical features		
Unique words	0.0332 (0.0019)	2.4×10^{-62}
Total words	0.0329 (0.0020)	2.2×10^{-59}
Lexical PC1	0.0186 (0.0011)	2.4×10^{-58}
Brunets' index	0.0300 (0.0019)	2.0×10^{-52}
Total repeats	0.0136 (0.0025)	3.4×10^{-08}
Honore's Stats	-0.0002 (0.0025)	0.922
Average word length	0.0079 (0.0024)	0.001
Moving average type token ratio	0.0069 (0.0023)	0.003
Syntactic features		
Verbs per utterances	-0.0179 (0.0025)	1.8×10^{-12}
Number of prepositions	0.0146 (0.0024)	1.9×10^{-09}
Past tense verb	-0.0105 (0.0023)	3.7×10^{-06}
Syntactic PC1	0.0048 (0.0013)	2.7×10^{-04}
Close class words	0.0065 (0.0024)	0.006
Pronoun index	-0.0049 (0.0022)	0.027
Number of determiners	-0.0011 (0.0023)	0.633
Open class words	-0.0009 (0.0021)	0.661
Psycholinguistic features		
Imageability	0.0045 (0.0022)	0.037
Concreteness	0.0031 (0.0023)	0.171
Psycholinguistic PC1	0.0016 (0.0013)	0.193
Age of acquisition	-0.0008 (0.0023)	0.725
Familiarity	0.0011 (0.0023)	0.624
Speech fluency features		
Number of syllables (n)	0.0317 (0.0019)	2.2×10^{-63}
Speech duration (sec)	0.0308 (0.0018)	1.0×10^{-61}
Speech fluency PC1	0.0173 (0.0011)	1.8×10^{-55}
Number of pauses (n)	0.0293 (0.0019)	2.4×10^{-52}
Hesitation ratio	-0.0076 (0.0018)	3.9×10^{-05}
Average word duration (sec)	0.0051 (0.0017)	0.004
Average pause duration (sec)	-0.0017 (0.0021)	0.440
Average syllable duration (sec)	0.0015 (0.0030)	0.617
Speech rate (syllables/ sec)	0.0006 (0.0019)	0.749
Articulation rate (syllables/ sec)	0.0004 (0.0021)	0.858

Predictors	Estimate per SD (SE)	p
Spectral and voice quality audio features		
MFCC 3	0.0073 (0.0019)	1.7×10^{-04}
Voiced (segments/sec)	0.0052 (0.0019)	0.005
F1 (Hz)	-0.0043 (0.0021)	0.039
MFCC 4	0.0029 (0.0020)	0.142
F0 (Hz)	-0.0034 (0.0025)	0.178
Jitter	0.0021 (0.0019)	0.285
Alpha ratio	0.0020 (0.0019)	0.304
Hammberg index	-0.0018 (0.0019)	0.346
F3 (Hz)	-0.0019 (0.0019)	0.352
Shimmer (dB)	-0.0015 (0.0019)	0.451
HNR (dB)	-0.0018 (0.0024)	0.459
Energy ratio	0.0012 (0.0017)	0.490
Spectral flux	-0.0010 (0.0018)	0.573
F2 (Hz)	-0.0011 (0,0020)	0.578
Spectral and voice quality PC1	-0.0004 (0.0008)	0.627
Loudness (dB)	-0.0007 (0.0017)	0.657
RMS energy	-0.0007 (0.0016)	0.660
MFCC 2	0.0006 (0.0017)	0.729
Loudness (peaks/sec)	0.0004 (0.0019)	0.834
MFCC 1	0.0002 (0.0016)	0.915

Please see Table 6.1 for definitions of predictors; The model was adjusted for age, sex, and education; Restricted to individuals with complete immediate, delayed recall, age, sex, education, and Gallery Game recall accuracy score information; N=7225 individuals; The order of features in the table is based on the p-value for each group; AIU - Automated Information Units; AOA - Age of acquisition; dB - decibel; HNR - Harmonics to Noise Ratio; Hz - Hertz; MFCC - Mel Frequency Cepstral Coefficient; PC1- First dimension of the principal component; SD - Standard Deviation; SE - Standard Error.

Further, to select the most relevant pipeline features that explain a significant amount of variation in Gallery Game recall accuracy score, stepwise regression was used (**Table 7.13**). The final model resulted in four pipeline features (number of syllables, AIU recall score, total repeats, and speech rate) associated with Gallery Game recall accuracy scores, in addition to age and education variables. These features explained 7% of the variation of Gallery Game recall accuracy scores.

Table 7.13 Association of pipeline features at recalls with Gallery Game Recall Accuracy score using stepwise regression analysis.

Predictors	Estimate per SD (SE)	p	Adjusted R^2	F
Model			0.07	86.3
Number of syllables	0.0286 (0.0023)	1.7×10^{-34}		
AIU recall score	0.0252 (0.0024)	3.7×10^{-26}		
Age (years)	-0.0120 (0.0015)	5.4×10^{-15}		
Total repeats	-0.0111 (0.0028)	5.2×10^{-05}		
Speech rate (syllables/sec)	-0.0079 (0.0019)	5.9×10^{-05}		
Education	0.00697 (0.0018)	1.6×10^{-04}		

N = 7225, including participants with complete Gallery Game recall accuracy score, Story Time pipeline features, age, sex, and education; The order of features in the table is based on the p-value; AIU - Automated Information Units.

In addition, only audio features were included in a stepwise model (**Table 7.14**) to investigate informative audio features that explain the variation of Gallery Game recall accuracy scores. Only the number of syllables was selected into the model, and the model explained only 4% of the variation showing the importance of other lexical and semantic features.

Table 7.14 Sensitivity analysis on the association of pipeline features (audio features) at recalls with Gallery Game recall accuracy score using stepwise regression analysis.

Predictors	Estimate per SD (SE)	p	Adjusted R^2	F
Model			0.04	125.4
Number of syllables	0.0325 (0.0019)	6.8×10^{-68}		
Age (years)	-0.0126 (0.0015)	3.6×10^{-16}		
Education	0.0080 (0.0019)	1.7×10^{-5}		

N = 7225, including participants with complete Gallery Game recall accuracy score, Story Time audio pipeline features, age, sex, and education; The order of features in the table is based on the p-value.

7.3.6.2 Mezurio Subjective Cognitive Decline (SCD)

The relationship between Mezurio SCD self-report scores from daily and yearly questionnaires and both AIU recall and delta AIU scores from the Story Time task was evaluated. There was a weak association in the expected direction between higher recall AIU scores and lower SCD scores from daily ($r = -0.07$, $p = 1.7 \times 10^{-9}$) and yearly ($r = -0.08$, $p = 9.9 \times 10^{-12}$). The association between delta AIU scores and SCD scores from daily ($r = 0.08$, $p = 1.5 \times 10^{-11}$) and

yearly ($r = 0.07$, $p = 1.6 \times 10^{-9}$) questionnaires was in the expected direction with weak effect size.

Figure 7.9 showed that the average AIU recall scores in the bottom two SCD score quartiles were higher than the average AIU recall scores from the participants reporting very high cognitive decline (second and top). The average delta AIU score increased steadily with SCD score quartiles (**Figure 7.10**).

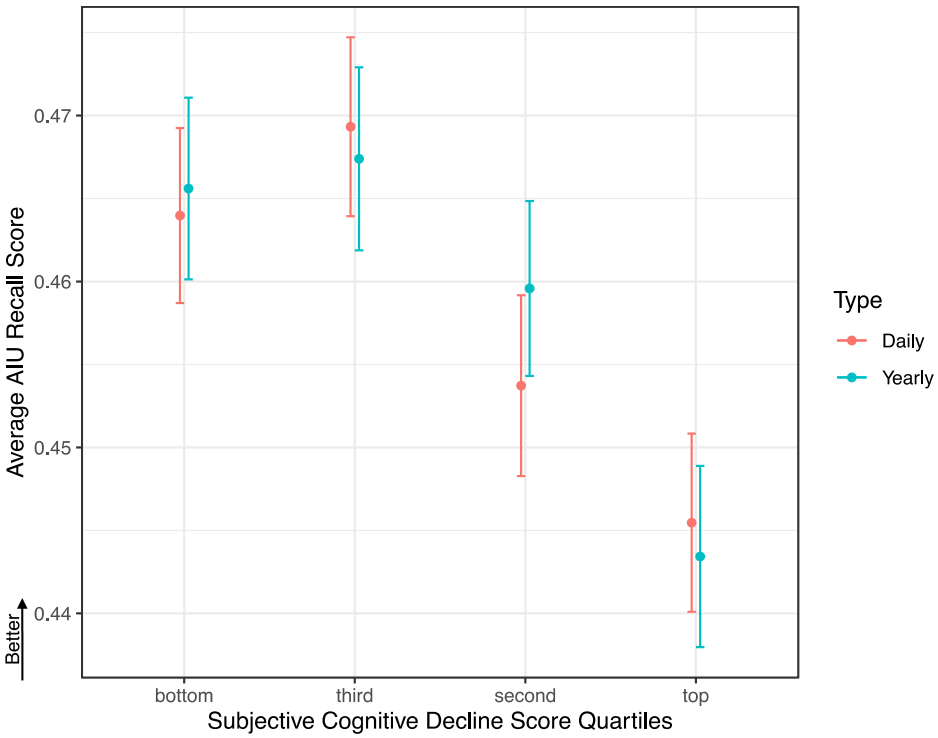


Figure 7.9 Plot showing mean and 95% CI on average AIU recall scores for Subjective Cognitive Decline score quartiles on daily and yearly questionnaires. Participants in the bottom quartile reported fewer cognitive problems.

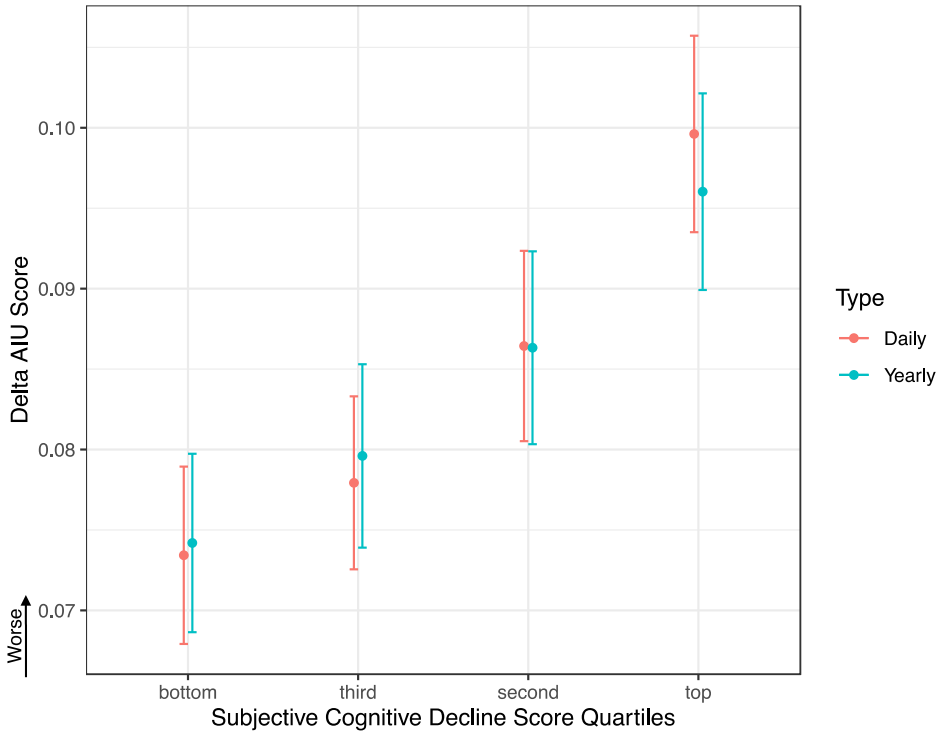


Figure 7.10 Plot showing mean and 95% CI on average delta AIU scores for Subjective Cognitive Decline score quartiles on daily and yearly questionnaires. Participants in the bottom quartile reported fewer cognitive problems.

Table 7.15 shows associations between pipeline features and SCD questionnaire scores. More pipeline features were statistically significantly associated with the SCD-Yearly questionnaire score compared to the SCD-Daily score. In addition to the AIU recall score's negative association with SCD scores, other features (lower total and unique words, lower Brunet's index, and higher Honore's stats) were statistically significant predictors of higher SCD scores, but strengths of the associations were less strong than for AIU recall score. Lexical PC1 was statistically significantly associated with SCD-Yearly score; however, the strength of the association was weaker compared to other lexical pipeline features. Only the lower use of close class words (in SCD-Daily score) and a lower number of determiners (both in daily and SCD-Yearly scores) were statistically significantly associated with higher SCD scores among syntactic features. Syntactic PC1 was statistically significantly associated with SCD-Daily score; however, the strength of the association was weaker compared to other syntactic pipeline

features. Among speech fluency features, higher average pause duration, higher hesitation ratio, lower speech rate, and lower articulation rate was associated with higher SCD scores. Speech fluency PC1 was statistically significantly associated with SCD-Yearly score; however, the strength of the association was weaker compared to other speech fluency pipeline features. Among spectral and voice quality features, a lower variation of loudness and voice segments were statistically significantly associated with higher SCD scores. The association between pipeline features and the spectral and voice quality PC1 did not reach statistical significance.

Table 7.15 Association of pipeline features (from recalls) with SCD questionnaire scores.

	SCD-Daily [a]		SCD-Yearly [b]	
Predictors	**Estimate per SD (SE)**	**p**	**Estimate per SD (SE)**	**p**
Semantic features				
AIU Score	-0.2501 (0.0319)	5.8×10^{-15}	-0.4289 (0.0578)	1.4×10^{-13}
Information Density	-0.0593 (0.0296)	0.045	-0.0503 (0.0533)	0.346
Lexical features				
Total Words	-0.0652 (0.0202)	0.031	-0.1826 (0.0540)	7.3×10^{-04}
Unique Words	-0.0606 (0.0297)	0.041	-0.1908 (0.0532)	3.4×10^{-04}
Moving average type token ratio	-0.0422 (0.0348)	0.225	-0.1552 (0.0624)	0.013
Brunets' Index	-0.1539 (0.0292)	1.4×10^{-07}	-0.2843 (0.0527)	7.1×10^{-08}
Honore's Stats	0.1239 (0.0363)	6.5×10^{-04}	0.1317 (0.0649)	0.043
Average Word Length	-0.0654 (0.0348)	0.060	-0.1942 (0.0628)	0.002
Total Repeats	0.0108 (0.0362)	0.766	0.0948 (0.0646)	0.1431
Lexical PC1	-0.0549 (0.0172)	0.001	-0.1202 (0.0308)	9.5×10^{-05}
Syntactic features				
Open Class Words	0.0707 (0.0318)	0.026	0.0914 (0.0573)	0.111
Close Class Words	-0.1359 (0.0349)	9.7×10^{-05}	-0.1455 (0.0634)	0.022
Pronoun Index	0.0367 (0.0332)	0.269	0.0145 (0.0597)	0.809
Verbs Per Utterances	0.0823 (0.0374)	0.028	0.0635 (0.0680)	0.350
Past Tense Verb	-0.0799 (0.0334)	0.017	-0.1697 (0.0598)	0.005
Number of Prepositions	-0.1335 (0.0359)	2.1×10^{-04}	-0.2336 (0.0653)	3.5×10^{-04}
Number of Determiners	-0.0908 (0.0339)	0.007	-0.0356 (0.0612)	0.561
Syntactic PC1	-0.0697 (0.0196)	3.7×10^{-04}	-0.0693 (0.0356)	0.051
Psycholinguistic features				
Age of Acquisition	-0.0445 (0.0345)	0.198	-0.0382 (0.0626)	0.541

	SCD-Daily [a]		SCD-Yearly [b]	
Predictors	**Estimate per SD (SE)**	**p**	**Estimate per SD (SE)**	**p**
Familiarity	-0.0692 (0.0339)	0.042	-0.1830 (0.0619)	0.003
Concreteness	-0.0442 (0.0336)	0.188	-0.1368 (0.0611)	0.025
Imageability	-0.0693 (0.0322)	0.032	-0.1838 (0.0589)	0.002
Psycholinguistic PC1	-0.0278 (0.0186)	0.136	-0.0836 (0.0339)	0.014
Speech fluency features				
Speech duration (sec)	-0.0152 (0.0276)	0.581	-0.0916 (0.0495)	0.064
Number of syllables (n)	-0.0434 (0.0280)	0.122	-0.1165 (0.0502)	0.020
Number of pauses (n)	0.0465 (0.0286)	0.104	0.0805 (0.0514)	0.117
Average pause duration (sec)	0.1233 (0.0299)	4.0×10^{-5}	0.3278 (0.0565)	6.9×10^{-9}
Average word duration (sec)	0.1195 (0.0262)	5.0×10^{-6}	0.2506 (0.0470)	1.0×10^{-7}
Speech rate (syllables/ sec)	-0.2093 (0.0291)	7.5×10^{-13}	-0.4275 (0.0525)	4.6×10^{-16}
Articulation rate (syllables/ sec)	-0.1525 (0.0318)	1.6×10^{-6}	-0.1905 (0.0576)	0.001
Average syllable duration (sec)	0.0442 (0.0451)	0.327	0.0134 (0.0789)	0.865
Hesitation Ratio	0.1558 (0.0272)	1.0×10^{-8}	0.3827 (0.0489)	5.8×10^{-15}
Speech fluency PC1	-0.0402 (0.0163)	0.013	-0.1100 (0.0293)	1.8×10^{-4}
Spectral and voice quality audio features				
RMS Energy	-0.0379 (0.0244)	0.120	-0.0023 (0.0437)	0.958
Loudness (dB)	-0.0616 (0.0245)	0.012	-0.0539 (0.0438)	0.218
MFCC 1	-0.0045 (0.0247)	0.855	0.0783 (0.0442)	0.076
MFCC 2	0.0550 (0.0252)	0.029	0.0427 (0.0453)	0.345
MFCC 3	0.0159 (0.0287)	0.581	0.0069 (0.0516)	0.893
MFCC 4	-0.0007 (0.0297)	0.982	-0.0189 (0.0533)	0.722
Jitter	0.0803 (0.0287)	0.005	0.0867 (0.0516)	0.093

	SCD-Daily [a]		SCD-Yearly [b]	
Predictors	**Estimate per SD (SE)**	**p**	**Estimate per SD (SE)**	**p**
Shimmer (dB)	-0.0667 (0.0288)	0.021	-0.1237 (0.0517)	0.017
HNR (dB)	0.1014 (0.0359)	0.005	0.1476 (0.0642)	0.022
F0 (Hz)	0.0221 (0.0373)	0.554	-0.0034 (0.0665)	0.959
F1 (Hz)	0.0261 (0.0311)	0.401	0.0376 (0.0556)	0.498
F2 (Hz)	0.0298 (0.0306)	0.329	0.0433 (0.0536)	0.419
F3 (Hz)	0.0398 (0.0300)	0.186	0.0593 (0.0522)	0.256
Hammberg Index	0.0733 (0.0281)	0.009	0.1694 (0.0505)	0.001
Alpha Ratio	-0.02268 (0.0289)	0.433	-0.0548 (0.0519)	0.291
Energy Ratio	-0.0280 (0.0248)	0.257	0.0158 (0.0443)	0.722
Spectral Flux	-0.0418 (0.0263)	0.112	-0.0559 (0.0473)	0.237
Voiced (segments/sec)	-0.2041 (0.0275)	1.3×10^{-13}	-0.3768 (0.0492)	2.3×10^{-14}
Loudness (peaks/sec)	-0.1687 (0.0294)	1.0×10^{-8}	-0.4037 (0.0527)	2.1×10^{-14}
Spectral and voice quality PC1	-0.0125 (0.0128)	0.328	-0.0024 (0.0228)	0.914

Please see Table 6.1 for definitions of predictors; The model was adjusted for age, sex, and education; Restricted to individuals with complete immediate, delayed recall, age, sex, education, and SCD-Daily or SCD-Yearly information; [a] N = 7133 and [b] N = 6855 participants; Reference group is least deprived group; AIU - Automated Information Units; AOA - Age of acquisition; dB - decibel; HNR - Harmonics to Noise Ratio; Hz - Hertz; MFCC - Mel Frequency Cepstral Coefficient; PC1- First dimension of the principal component; SCD - Subjective Cognitive Decline; SD - Standard Deviation; SE - Standard Error.

Stepwise regression analysis was conducted on SCD scores separately for daily and yearly scores to obtain the most relevant pipeline features in explaining SCD scores (**Table 7.16**). In the final model, where SCD-Daily score was the dependent variable, three pipeline features (AIU recall score, voice segments (per second), and unique words) age were selected. On the other stepwise model explaining SCD-Yearly score, only speech rate and AIU recall scores were selected as the most important predictors, along with age. These selected variables explained 2% of the variation of SCD scores on both models.

Table 7.16 Association of pipeline features at recalls with SCD scores using stepwise regression analysis.

Predictors	**Estimate per SD (SE)**	**p**	**Adjusted R^2**	**F**
SCD-Daily score [a]			0.02	42.06
AIU recall score	-0.3123 (0.0402)	9.6×10^{-15}		
Voice segments (per second)	-0.1849 (0.0281)	5.3×10^{-11}		
Age (years)	0.1433 (0.0229)	4.3×10^{-10}		
Unique words	0.1640 (0.0374)	1.2×10^{-05}		
SCD-Yearly score [b]			0.02	57.89
Age (years)	0.3279 (0.0579)	3.1×10^{-16}		
Speech rate (syllables/sec)	-0.3624 (0.0525)	4.5×10^{-12}		
AIU recall score	-0.3713 (0.0579)	1.5×10^{-10}		

[a] N = 7133 and [b] N = 6855 including participants with complete SCD-Daily or SCD-Yearly score, Story Time pipeline features, age, sex, and education; The order of features in the table is based on the p-value; SCD - Subjective Cognitive Decline

A sensitivity analysis that included only audio features (**Table 7.17**), voiced segments and speech rate were selected as the most relevant audio features in the model explaining a significant amount of variation in the SCD-Daily score. Similarly, voiced segments and hesitation ratio were selected as the most important predictors for predicting SCD-Yearly score. These selected variables explained 2% of the variation of SCD scores on both models (similar to **Table 7.16**).

Table 7.17 Sensitivity analysis on the association of audio features at recalls with SCD scores using stepwise regression analysis.

Predictors	Estimate per SD (SE)	p	Adjusted R^2	F
SCD-Daily score [a]			0.02	39.89
Voice segments (per second)	-0.1354 (0.0334)	5.1×10^{-5}		
Age (years)	0.1572 (0.0225)	2.9×10^{-12}		
Speech rate (syllables/sec)	-0.1251 (0.0355)	4.2×10^{-4}		
SCD-Yearly score [b]			0.02	49.78
Age (years)	0.3230 (0.0401)	8.8×10^{-16}		
Voice segments (per second)	-0.2435 (0.0555)	1.2×10^{-05}		
Hesitation ratio	0.2652 (0.0553)	1.7×10^{-06}		

[a] N = 7133 and [b] N = 6855 including participants with complete SCD-Daily or SCD-Yearly score, Story Time audio pipeline features; The order of features in the table is based on the p-value; SCD - Subjective Cognitive Decline

7.3.7 The association of participant's sociodemographic factors with the Story Time pipeline features

This section investigates the association of Story Time pipeline features from recalls (immediate and delayed recall) and their delta features with age, education, family history of dementia (either parent diagnosed with dementia) and deprivation.

7.3.7.1 Participant's age

In the GameChanger study population, the mean (SD) age of the participant was 54.9 (12.2) years. As evidenced from the literature, it is expected that cognitive performance decreases with increasing age in cross-sectional analyses [303], [304]. The association of each pipeline feature with age was univariately evaluated using linear mixed models with participant and story as random effects adjusted for story IUs (**Table 7.18**).

Among language features, AIU and information density scores at recall showed a statistically significant association with increasing age cross-sectionally. Delta AIU score and delta information density showed a positive association with age. Other delta features had weaker

associations with age compared to the delta AIU score. Many lexical features at recall were statistically significantly associated with age, with the strongest associations (most extreme p-values) found for total and unique word count. Among the syntactic features found to be statistically significantly associated with age, the use of past tense verbs demonstrated the strongest association. PC1 features showed a similar strength p-value to the strongest individual features for lexical but not for syntactic or psycholinguistic features.

Speech-fluency features at recall, including higher speech duration, a higher number of syllables, and a higher number of pauses, were strongly associated with age, while lower speech rate and higher hesitation ratio had a weaker but statistically significant association. Speech fluency PC1 showed a statistically significant association with age with a lower strength p-value to the strongest individual feature in that group. Most of the spectral and voice quality features were also statistically significantly associated with age, with HNR being the strongest predictor of age cross-sectionally. PC1 feature for spectral and voice quality was statistically significantly associated with age; however, the strength of the p-value was lower than the strongest individual feature in that group. Delta features were generally much less extreme p-values.

Table 7.18 The difference in Story Time pipeline features per year of age cross-sectionally.

	Recalls		Delta	
Outcome variables	**Estimate (SE) in SD units**	**p**	**Estimate (SE) in SD units**	**p**
Semantic features				
AIU score	-0.0046 (0.0005)	4.8×10^{-22}	0.0061 (0.0006)	1.9×10^{-22}
Information density	-0.0115 (0.0006)	2.9×10^{-66}	0.0045 (0.0006)	1.0×10^{-13}
Lexical features				
Total words	0.0099 (0.0007)	2.8×10^{-42}	0.0011 (0.0004)	0.015
Unique words	0.0098 (0.0007)	9.9×10^{-42}	0.0002 (0.0004)	0.619
Moving average type token ratio	-0.0004 (0.0005)	0.440	-0.0023 (0.0006)	8.1×10^{-5}
Brunets' index	0.0082 (0.0007)	7.9×10^{-30}	0.0017 (0.0006)	0.003
Honore's stats	-0.0003 (0.0005)	0.565	-0.0021 (0.0007)	0.003
Average word length	0.0029 (0.0005)	8.9×10^{-8}	-0.0051 (0.0005)	4.9×10^{-23}
Total repeats	0.0052 (0.0006)	7.0×10^{-18}	0.0014 (0.0005)	0.007
Lexical PC1	0.0188 (0.0013)	1.5×10^{-48}	-0.0043 (0.0011)	0.0001
Syntactic features				
Open class words	-0.0049 (0.0006)	2.6×10^{-15}	-0.0034 (0.0006)	2.3×10^{-9}
Close class words	0.0045 (0.0004)	4.3×10^{-28}	-0.0005 (0.0004)	0.235
Pronoun index	-0.0002 (0.0005)	0.715	0.0024 (0.0005)	9.7×10^{-7}
Verbs per utterances	-0.0035 (0.0005)	3.1×10^{-14}	0.0005 (0.0006)	0.331
Past tense verb	-0.0089 (0.0006)	6.6×10^{-45}	-0.0007 (0.0001)	1.3×10^{-8}
Number of prepositions	0.0048 (0.0005)	2.9×10^{-23}	-0.0004 (0.0005)	0.512
Number of determiners	0.0027 (0.0004)	3.1×10^{-12}	-0.0005 (0.0004)	0.277
Syntactic PC1	0.0012 (0.0011)	0.292	-0.0004 (0.0011)	0.711
Psycholinguistic features				
AOA	0.0040 (0.0005)	6.8×10^{-16}	-0.0014 (0.0005)	0.0006
Familiarity	-0.0037 (0.0005)	2.1×10^{-14}	0.0024 (0.0007)	0.0004
Concreteness	-0.0053 (0.0005)	1.2×10^{-28}	0.0021 (0.0007)	0.002
Imageability	-0.0047 (0.0005)	2.4×10^{-20}	0.0026 (0.0007)	0.00043
Psycholinguistic PC1	-0.0039 (0.0012)	0.001	0.0052 (0.0016)	0.001
Speech fluency features				

	Recalls		Delta	
Outcome variables	**Estimate (SE) in SD units**	**p**	**Estimate (SE) in SD units**	**p**
Speech duration (sec)	0.0125 (0.0008)	3.2×10^{-55}	0.0008 (0.0004)	0.032
Number of syllables (n)	0.0124 (0.0008)	5.3×10^{-56}	0.0001 (0.0004)	0.916
Average word duration (sec)	0.0107 (0.0008)	1.1×10^{-38}	0.0004 (0.0004)	0.366
Number of pauses (n)	0.0130 (0.0007)	6.1×10^{-64}	0.0017 (0.0004)	0.0001
Average pause duration (sec)	0.0027 (0.0007)	2.6×10^{-4}	-0.0006 (0.0007)	0.361
Speech rate (syllables/ sec)	-0.0029 (0.0008)	9.1×10^{-5}	-0.0035 (0.0005)	2.5×10^{-12}
Articulation rate (syllables/ sec)	0.0012 (0.0007)	0.079	-0.0038 (0.0005)	2.1×10^{-12}
Average syllable duration (sec)	0.0001 (0.0004)	0.837	0.0003 (0.0007)	0.692
Hesitation ratio	0.0067 (0.0007)	2.7×10^{-19}	0.0003 (0.0003)	0.378
Speech fluency PC1	0.0211 (0.0014)	6.6×10^{-54}	0.0001 (0.0012)	0.963
Spectral and voice quality audio features				
RMS energy	-0.0018 (0.0009)	0.045	-0.0005 (0.0002)	0.022
Loudness (dB)	-0.0029 (0.0009)	0.001	-0.0006 (0.0002)	0.017
MFCC 1	-0.0068 (0.0009)	1.8×10^{-14}	-0.0006 (0.0002)	0.013
MFCC 2	0.0050 (0.0008)	7.4×10^{-9}	0.0009 (0.0003)	0.002
MFCC 3	-0.0041 (0.0008)	5.1×10^{-8}	-0.0003 (0.0003)	0.508
MFCC 4	0.0029 (0.0007)	8.6×10^{-5}	-0.0002 (0.0004)	0.717
Jitter	0.0022 (0.0008)	0.003	0.0018 (0.0004)	3.8×10^{-5}
Shimmer (dB)	0.0096 (0.0008)	1.1×10^{-32}	-0.0014 (0.0004)	0.0004
HNR (dB)	-0.0103 (0.0006)	2.3×10^{-62}	0.0007 (0.0003)	0.009
F0 (Hz)	-0.0042 (0.0006)	9.4×10^{-13}	0.0007 (0.0002)	0.005
F1 (Hz)	-0.0059 (0.0007)	2.5×10^{-17}	0.0013 (0.0004)	0.001
F2 (Hz)	-0.0060 (0.0007)	5.5×10^{-17}	0.0009 (0.0004)	0.020
F3 (Hz)	-0.0041 (0.0007)	2.1×10^{-8}	0.0013 (0.0004)	0.002
Hammarberg index	-0.0024 (0.0008)	0.002	0.0010 (0.0004)	0.150
Alpha ratio	0.0056 (0.0007)	6.5×10^{-14}	-0.0005 (0.0005)	0.267
Energy ratio	0.0041 (0.0009)	4.4×10^{-6}	-0.0007 (0.0003)	0.021
Spectral flux	-0.0039 (0.0008)	2.3×10^{-9}	0.0002 (0.0004)	0.593
Voiced (segments/sec)	-0.0028 (0.0008)	4.7×10^{-4}	-0.0009 (0.0005)	0.049
Loudness (peaks/sec)	-0.0034 (0.0007)	3.7×10^{-6}	-0.0017 (0.0005)	0.00078

	Recalls		Delta	
Outcome variables	**Estimate (SE) in SD units**	**p**	**Estimate (SE) in SD units**	**p**
Spectral and voice quality PC1	-0.0116 (0.0018)	1.4×10^{-11}	0.0028 (0.0011)	0.011

Please see Table 6.1 for definitions of predictors; $N_{Participants} = 7346$; $N_{Stories} = 12$; $N_{Observations} = 34354$ recall pairs; The models with spectral and voice quality audio features were adjusted for sex as these features are strongly associated with sex; Restricted to individuals with complete learning, immediate, delayed recall, and age information; SE - Standard Error; MFCC - Mel Frequency Cepstral Coefficient, HNR - Harmonics to Noise Ratio; Hz - Hertz; dB - decibel; PC1- First dimension of the principal component.

7.3.7.2 Participants Education

The distribution of GameChanger participants in various age groups stratified by education is shown in **Figure A.13**. The number of participants with higher education in each age group was higher compared to lower education group. The association of Story Time pipeline features extracted from recalls and delta with education was evaluated cross-sectionally using linear mixed effect models with participant and story as random effects adjusted for age and story IUs and education group being the predictor variable. **Table 7.19** shows the cross-sectional association of lower education with Story Time pipeline features.

Semantic pipeline features, AIU recall score and information density were statistically significantly associated with education. Participants in the lower education group (participants with no university degree) produced lower AIU recall scores compared to the higher education group, while the information density was higher in the lower education group, likely due to their shorter speech duration compared to the more highly educated participants. Lexical features, including total words, unique words, Brunet's index, average word length, and total repeats, demonstrated statistically significant associations with education at recall. Specifically, less educated participants produced a lower number of total and unique words, lower average word length and higher number of repetitions compared to higher education groups. Among the syntactic features that were found to be statistically significantly associated with lower education, the use of verbs demonstrated the strongest association. PC1 features showed a similar strength p-value to the strongest individual features for lexical but not for syntactic or psycholinguistic features.

Speech fluency measures, including speech duration, number of syllables, and number of pauses, also showed statistically significant differences between higher and lower education

groups. Participants with lower education produced shorter speech, with fewer syllables and pauses than those in the higher education groups. Speech fluency PC1 showed the strongest association with lower education and the higher strength p-value among the individual speech fluency features. Among spectral and voice quality features that were found to be statistically significantly associated with lower education, jitter demonstrated the strongest association. Delta features were generally less strongly associated with education than the features from the recalls.

Table 7.19 The difference in Story Time pipeline features per lower education cross-sectionally.

Outcome variables	Recalls		Delta	
	Estimate (SE) in SD units	p	Estimate (SE) in SD units	p
Semantic features				
AIU score	-0.1436 (0.0148)	4.6×10^{-22}	0.0613 (0.0148)	3.5×10^{-5}
Information density	0.1261 (0.0160)	3.7×10^{-15}	0.0611 (0.0145)	2.4×10^{-5}
Lexical features				
Total words	-0.1817 (0.0175)	7.3×10^{-25}	-0.0019 (0.0106)	0.860
Unique words	-0.1974 (0.0176)	6.7×10^{-29}	0.0054 (0.0109)	0.619
Moving average type token ratio	-0.0205 (0.0115)	0.075	-0.0109 (0.0135)	0.419
Brunets' index	-0.1565 (0.0174)	2.4×10^{-19}	0.0542 (0.0138)	8.7×10^{-5}
Honore’s stats	-0.0162 (0.0125)	0.193	-0.0442 (0.0165)	0.007
Average word length	-0.0671 (0.1329)	4.5×10^{-7}	-0.0103 (0.0122)	0.397
Total repeats	-0.0711 (0.146)	1.1×10^{-6}	-0.0067 (0.0121)	0.580
Lexical PC1	-0.3421 (0.0313)	1.2×10^{-27}	-0.0621 (0.0282)	0.027
Syntactic features				
Open class words	0.0311 (0.0151)	0.040	-0.0331 (0.0135)	0.014
Close class words	-0.0364 (0.0099)	2.3×10^{-4}	-0.0019 (0.0106)	0.859
Pronoun index	0.0363 (0.0115)	0.002	0.0411 (0.0117)	4.3×10^{-4}
Verbs per utterances	0.0788 (0.0112)	2.6×10^{-12}	-0.0194 (0.0133)	0.145
Past tense verb	0.0629 (0.0154)	4.7×10^{-5}	0.0044 (0.0030)	0.145
Number of prepositions	-0.0631 (0.0116)	5.5×10^{-8}	0.0131 (0.0129)	0.311
Number of determiners	-0.0036 (0.0094)	0.702	-0.0126 (0.0099)	0.201
Syntactic PC1	-0.0156 (0.0276)	0.572	-0.0099 (0.0276)	0.717
Psycholinguistic features				
AOA	-0.0205 (0.0120)	0.088	-0.0103 (0.0121)	0.395
Familiarity	0.0327 (0.0117)	0.005	0.0275 (0.0161)	0.088

Outcome variables	Recalls		Delta	
	Estimate (SE) in SD units	p	Estimate (SE) in SD units	p
Concreteness	0.0288 (0.0116)	0.013	0.0343 (0.0157)	0.028
Imageability	0.2132 (0.0121)	0.078	0.0409 (0.0174)	0.019
Psycholinguistic PC1	0.0061 (0.0289)	0.833	0.0635 (0.0385)	0.099
Speech fluency features				
Speech duration (sec)	-0.2102 (0.0192)	1.6×10^{-27}	-0.0004 (0.0093)	0.966
Number of syllables (n)	-0.1993 (0.0190)	1.5×10^{-25}	0.0059 (0.0096)	0.541
Number of pauses (n)	-0.1982 (0.0186)	2.9×10^{-26}	-0.0113 (0.0091)	0.214
Average pause duration (sec)	0.0513 (0.0162)	0.002	-0.0418 (0.0168)	0.013
Average word duration (sec)	-0.0133 (0.0199)	0.501	-0.0117 (0.0102)	0.254
Speech rate (syllables/ sec)	-0.0084 (0.0183)	0.644	0.0415 (0.0119)	5.3×10^{-4}
Articulation rate (syllables/ sec)	0.0109 (0.0166)	0.513	0.0137 (0.0128)	0.281
Average syllable duration (sec)	-0.0127 (0.0103)	0.219	0.0065 (0.0164)	0.689
Hesitation ratio	0.0352 (0.0182)	0.053	-0.0120 (0.0083)	0.146
Speech fluency PC1	-0.4047 (0.0330)	3.3×10^{-34}	0.0871 (0.0314)	0.006
Spectral and voice quality audio features				
RMS energy	0.0284 (0.0219)	0.196	0.0104 (0.0051)	0.042
Loudness (dB)	0.0257 (0.0219)	0.240	0.0170 (0.0057)	0.003
MFCC 1	-0.0043 (0.0217)	0.843	-0.0067 (0.0053)	0.214
MFCC 2	-0.0781 (0.0212)	2.2×10^{-4}	0.0002 (0.0007)	0.979
MFCC 3	-0.0239 (0.0184)	0.194	0.0166 (0.0089)	0.064
MFCC 4	-0.0642 (0.0179)	3.4×10^{-4}	-0.0087 (0.0098)	0.376
Jitter	-0.1299 (0.0183)	1.5×10^{-12}	-0.0088 (0.0103)	0.392
Shimmer (dB)	-0.0403 (0.0184)	0.028	0.0114 (0.0093)	0.220
HNR (dB)	0.0347 (0.0149)	0.019	-0.0079 (0.0065)	0.221
F0 (Hz)	0.0379 (0.0143)	0.008	-0.0025 (0.0056)	0.648
F1 (Hz)	-0.0023 (0.0172)	0.891	-0.0012 (0.0091)	0.896

Outcome variables	Recalls		Delta	
	Estimate (SE) in SD units	p	Estimate (SE) in SD units	p
F2 (Hz)	-0.0108 (0.0175)	0.536	0.0141 (0.0095)	0.139
F3 (Hz)	-0.0228 (0.0178)	0.201	-0.0020 (0.0102)	0.842
Hammarberg index	-0.0766 (0.0189)	5.4×10^{-5}	-0.0119 (0.0097)	0.219
Alpha ratio	-0.0189 (0.0181)	0.296	0.0123 (0.0109)	0.259
Energy ratio	-0.0215 (0.0216)	0.318	0.0047 (0.0073)	0.517
Spectral flux	-0.0117 (0.0203)	0.562	0.0076 (0.0092)	0.407
Voiced (segments/sec)	-0.0489 (0.0193)	0.114	0.0604 (0.0107)	1.7×10^{-8}
Loudness (peaks/sec)	-0.0577 (0.0179)	0.877	0.0593 (0.0123)	7.0×10^{-5}
Spectral and voice quality PC1	-0.0243 (0.5150)	0.538	-0.0129 (0.0268)	0.628

Please see Table 6.1 for definitions of predictors; $N_{Participants} = 7291$; $N_{Stories} = 12$; $N_{Observations} = 34304$ recall pairs; The models with spectral and voice quality audio features were adjusted for sex as these features are strongly associated with sex; Restricted to individuals with complete learning, immediate, delayed recall, age, and education information; Reference group: higher education; AIU - Automated Information Units; AOA - Age of acquisition; dB - decibel; HNR - Harmonics to Noise Ratio; Hz - Hertz; MFCC - Mel Frequency Cepstral Coefficient; PC1- First dimension of the principal component; SD - Standard Deviation; SE - Standard Error.

7.3.7.3 Family history of dementia

The baseline characteristics of participants with and without a family history of dementia are presented in **Table A.8**. A higher proportion of older participants, females, and those in the most deprived group had a family history of dementia. The association of family history of dementia with pipeline features was evaluated univariately using a linear mixed model with participants and stories as random effects, family history of dementia as fixed effect adjusted for story IUs, age, sex, education and neurological condition (**Table 7.20**). None of the pipeline features made significant contributions in predicting a positive family history of dementia.

Table 7.20 The difference in Story Time pipeline features associated with having any family history of dementia as the predictor variable cross-sectionally.

	Recalls		Delta	
Outcome variables	**Estimate (SE) in SD units**	**p**	**Estimate (SE) in SD units**	**p**
Semantic features				
AIU score	-0.0117 (0.0116)	0.312	-0.0210 (0.0149)	0.159
Information density	-0.0269 (0.0161)	0.094	0.0027 (0.0146)	0.851
Lexical features				
Total words	0.0269 (0.0178)	0.129	-0.0071 (0.0107)	0.508
Unique words	0.0150 (0.0178)	0.399	-0.0091 (0.0109)	0.409
Moving average type token ratio	-0.0196 (0.0114)	0.086	0.0239 (0.0135)	0.076
Brunets' index	0.0351 (0.0174)	0.044	-0.0214 (0.0139)	0.124
Honore’s stats	-0.0077 (0.0126)	0.535	0.0209 (0.0167)	0.207
Average word length	-0.0284 (0.0133)	0.033	0.0076 (0.0122)	0.537
Total repeats	0.0079 (0.0148)	0.589	-0.0161 (0.0122)	0.188
Lexical PC1	0.0541 (0.0315)	0.087	0.3552 (0.0282)	0.208
Syntactic features				
Open class words	-0.0170 (0.0152)	0.263	0.0224 (0.0136)	0.099
Close class words	0.0321 (0.0098)	0.002	0.0076 (0.0107)	0.476
Pronoun index	0.0023 (0.0116)	0.841	-0.0070 (0.0117)	0.550
Verbs per utterances	-0.0194 (0.0113)	0.086	-0.0009 (0.0134)	0.943
Past tense verb	0.0178 (0.0156)	0.255	-0.0025 (0.0031)	0.410
Number of prepositions	0.0295 (0.0116)	0.011	0.0103 (0.0131)	0.432
Number of determiners	0.0228 (0.0093)	0.015	0.0028 (0.0099)	0.778
Syntactic PC1	0.0479 (0.0277)	0.084	-0.0084 (0.0279)	0.764
Psycholinguistic features				
AOA	-0.0171 (0.0121)	0.110	-0.0174 (0.0122)	0.156
Familiarity	0.0123 (0.0118)	0.298	-0.0139 (0.0163)	0.392

	Recalls		Delta	
Outcome variables	**Estimate (SE) in SD units**	**p**	**Estimate (SE) in SD units**	**p**
Concreteness	0.0100 (0.0118)	0.394	-0.0107 (0.0158)	0.499
Imageability	0.0102 (0.0122)	0.403	-0.0123 (0.0175)	0.479
Psycholinguistic PC1	0.0418 (0.0290)	0.149	0.0225 (0.0384)	0.559
Semantic Content				
AIU score	-0.0117 (0.0116)	0.312	-0.0210 (0.0149)	0.159
Information density	-0.0269 (0.0161)	0.094	0.0027 (0.0146)	0.851
Speech fluency features				
Speech duration (sec)	0.0358 (0.0194)	0.065	0.0044 (0.0095)	0.643
Number of syllables (n)	0.0264 (0.0191)	0.168	0.0002 (0.0098)	0.982
Number of pauses (n)	0.0243 (0.0188)	0.195	-0.0005 (0.0093)	0.960
Average pause duration (sec)	0.0172 (0.0163)	0.289	0.0192 (0.0171)	0.261
Average word duration (sec)	-0.0299 (0.0200)	0.136	0.0171 (0.0103)	0.968
Speech rate (syllables/ sec)	-0.0199 (0.0182)	0.276	-0.0139 (0.0121)	0.249
Articulation rate (syllables/ sec)	-0.0247 (0.0167)	0.139	-0.0174 (0.0129)	0.178
Average syllable duration (sec)	0.0015 (0.0099)	0.875	0.0054 (0.0166)	0.744
Hesitation ratio	0.0116 (0.0183)	0.527	0.0067 (0.0084)	0.429
Speech fluency PC1	0.0467 (0.0331)	0.158	-0.0262 (0.0315)	0.404
Spectral and voice quality audio features				
RMS energy	-0.0263 (0.0222)	0.236	0.0009 (0.0052)	0.851
Loudness (dB)	-0.0237 (0.0222)	0.284	-0.0035 (0.0058)	0.543
MFCC 1	-0.0432 (0.0219)	0.049	-0.0093 (0.0054)	0.008
MFCC 2	0.0183 (0.0215)	0.394	-0.0003 (0.0074)	0.962
MFCC 3	-0.0185 (0.0187)	0.322	-0.0086 (0.0091)	0.347
MFCC 4	0.0136 (0.0181)	0.451	-0.0108 (0.0099)	0.275
Jitter	-0.0096 (0.0185)	0.603	-0.0243 (0.0104)	0.019
Shimmer (dB)	-0.0299 (0.0186)	0.107	-0.0102 (0.0093)	0.279
HNR (dB)	-0.0206 (0.0150)	0.171	0.0096 (0.0065)	0.141

	Recalls		Delta	
Outcome variables	**Estimate (SE) in SD units**	**p**	**Estimate (SE) in SD units**	**p**
F0 (Hz)	-0.0239 (0.0145)	0.098	0.0088 (0.0057)	0.120
F1 (Hz)	0.0045 (0.0173)	0.796	0.0020 (0.0092)	0.825
F2 (Hz)	-0.0105 (0.0177)	0.553	-0.0044 (0.0097)	0.647
F3 (Hz)	-0.0155 (0.0077)	0.045	-0.0135 (0.0104)	0.194
Hammarberg index	-0.0246 (0.0192)	0.198	-0.0103 (0.0098)	0.292
Alpha ratio	-0.002 (0.0184)	0.971	0.0007 (0.0111)	0.949
Energy ratio	-0.0493 (0.0218)	0.023	0.0006 (0.0074)	0.932
Spectral flux	-0.0341 (0.0206)	0.098	-0.0019 (0.0094)	0.838
Voiced (segments/sec)	-0.0199 (0.0193)	0.301	-0.0112 (0.0108)	0.297
Loudness (peaks/sec)	0.0038 (0.0180)	0.834	-0.0182 (0.0124)	0.143
Spectral and voice quality PC1	-0.0864 (0.0424)	0.042	-0.0261 (0.0272)	0.338

Please see Table 6.1 for definitions of predictors; $N_{participants}$ = 7076; $N_{stories}$ = 12; $N_{Observation}$ = 33209 recall pairs; The models were adjusted for age, sex, education, family history of dementia and neurological condition; Restricted to individuals with complete learning, immediate, delayed recall, age, sex, education, family history of dementia (yes or no), and neurological condition information; Participants who reported "not sure" for the family history of dementia were excluded for this analysis; AIU - Automated Information Units; AOA - Age of acquisition; dB - decibel; HNR - Harmonics to Noise Ratio; Hz - Hertz; MFCC - Mel Frequency Cepstral Coefficient; PC1- First dimension of the principal component; SD - Standard Deviation; SE - Standard Error; Reference group is no family history of dementia.

7.3.7.4 Socioeconomic deprivation deciles

The baseline characteristics of the participants in the most and least deprived groups are presented in **Table A.9**. A higher proportion of participants from the least deprived group were older, had higher educational attainment, and with either parent with a family history of dementia.

Univariate analyses of the pipeline features at recall (**Table 7.21**) showed a statistically significant difference between the most and least deprived groups in terms of AIU recall score and lexical features such as total words, unique words, and Brunet's index. Specifically, these features were significantly lower in the most deprived group compared to the least deprived group. The association of deprivation with syntactic and psycholinguistic features did not reach statistical significance.

Among speech fluency features, the most deprived group produced shorter speech duration, fewer syllables, and a higher hesitation ratio, all were statistically significant predictors of socioeconomic deprivation. Speech fluency PC1 showed a slightly stronger association with lower among the individual speech fluency features. Among spectral and voiced quality features that were statistically significantly associated with deprivation, MFCC1 demonstrated the strongest association. PC1 feature for spectral and voice quality was statistically significantly associated with the most deprived group; however, the strength of the p-value was lower than the strongest individual feature in that group. None of the delta features was found to be informative in differentiating deprivation decile groups.

Table 7.21 The difference in Story Time pipeline features per most deprived group as the predictor variable cross-sectionally.

Outcome variables	Recalls		Delta	
	Estimate (SE) in SD units	p	Estimate (SE) in SD units	p
Semantic features				
AIU score	-0.0486 (0.0140)	5.2×10^{-4}	0.0472 (0.0179)	0.008
Information density	0.0162 (0.01925)	0.403	0.0441 (0.0174)	0.012
Lexical features				
Total words	-0.0853 (0.0214)	6.9×10^{-5}	0.0129 (0.0128)	0.314
Unique words	-0.0814 (0.0215)	1.5×10^{-4}	0.0213 (0.0131)	0.106
Moving average type token ratio	-0.0055 (0.0139)	0.689	0.0041 (0.0166)	0.804
Brunets' index	-0.1021 (0.0208)	8.9×10^{-7}	0.0461 (0.0166)	0.005
Honore's stats	0.0242 (0.0150)	0.107	-0.0273 (0.0197)	0.167
Average word length	-0.0424 (0.0159)	0.008	0.0004 (0.0149)	0.981
Total repeats	-0.0298 (0.0182)	0.101	0.0036 (0.0148)	0.808
Lexical PC1	-0.1573 (0.0373)	2.5×10^{-5}	-0.0578 (0.0341)	0.091
Syntactic features				
Open class words	0.0276 (0.0181)	0.128	-0.0524 (0.0161)	0.001
Close class words	-0.0219 (0.0119)	0.066	0.0301 (0.0128)	0.019
Pronoun index	0.0003 (0.0139)	0.977	0.0207 (0.0139)	0.139
Verbs per utterances	0.0097 (0.0137)	0.481	-0.0306 (0.0161)	0.057
Past tense verb	-0.0136 (0.0189)	0.471	0.0029 (0.0037)	0.426
Number of prepositions	-0.0363 (0.0141)	0.009	-0.0005 (0.0006)	0.390
Number of determiners	-0.0033 (0.0114)	0.749	-0.0092 (0.0108)	0.392
Syntactic PC1	-0.0401 (0.0326)	0.219	-0.0763 (0.0331)	0.021
Psycholinguistic features				
AOA	0.0271 (0.0146)	0.064	0.0284 (0.0148)	0.055
Familiarity	-0.0176 (0.0143)	0.218	0.0498 (0.0192)	0.009

Outcome variables	Recalls		Delta	
	Estimate (SE) in SD units	p	Estimate (SE) in SD units	p
Concreteness	-0.0261 (0.0142)	0.066	0.0493 (0.0188)	0.009
Imageability	-0.0255 (0.0147)	0.083	0.0582 (0.0206)	0.005
Psycholinguistic PC1	-0.0879 (0.0346)	0.111	0.1205 (0.0451)	0.008
Speech fluency features				
Speech duration (sec)	-0.1098 (0.0232)	2.2×10^{-6}	0.0245 (0.0115)	0.033
Number of syllables (n)	-0.1074 (0.0229)	2.7×10^{-6}	0.0108 (0.0118)	0.357
Number of pauses (n)	-0.0527 (0.0227)	0.020	0.0090 (0.0111)	0.416
Average pause duration (sec)	0.0529 (0.0199)	0.008	-0.0317 (0.0211)	0.133
Average word duration (sec)	-0.0092 (0.0239)	0.703	0.0070 (0.0124)	0.573
Speech rate (syllables/ sec)	-0.0704 (0.0221)	0.001	0.0045 (0.0145)	0.758
Articulation rate (syllables/ sec)	-0.0089 (0.0203)	0.659	-0.0299 (0.0153)	0.051
Average syllable duration (sec)	0.0135 (0.0127)	0.289	0.0182 (0.0204)	0.372
Hesitation ratio	0.0827 (0.0221)	1.9×10^{-4}	-0.0148 (0.0103)	0.151
Speech fluency PC1	-0.1974 (0.0386)	3.2×10^{-7}	0.0633 (0.0385)	0.100
Spectral and voice quality audio features				
RMS energy	0.1613 (0.0265)	1.3×10^{-9}	0.0106 (0.0063)	0.091
Loudness (dB)	0.1403 (0.0264)	1.1×10^{-7}	0.0174 (0.0069)	0.013
MFCC 1	0.1717 (0.0265)	1.1×10^{-10}	-0.0007 (0.0065)	0.919
MFCC 2	-0.1363 (0.0259)	1.5×10^{-7}	-0.0048 (0.0089)	0.594
MFCC 3	-0.0220 (0.0225)	0.328	0.0084 (0.0110)	0.446
MFCC 4	-0.0255 (0.0219)	0.244	0.0133 (0.0119)	0.266
Jitter	-0.0033 (0.0224)	0.881	-0.0225 (0.0124)	0.071
Shimmer (dB)	0.0839 (0.0222)	1.6×10^{-4}	-0.0038 (0.0112)	0.738
HNR (dB)	-0.0069 (0.0182)	0.701	0.0037 (0.0079)	0.638
F0 (Hz)	0.0132 (0.0177)	0.455	-0.0036 (0.0068)	0.598
F1 (Hz)	0.0167 (0.0209)	0.426	-0.0094 (0.0111)	0.397
F2 (Hz)	-0.0033 (0.0214)	0.876	-0.0113 (0.0119)	0.340

Outcome variables	Recalls		Delta	
	Estimate (SE) in SD units	p	Estimate (SE) in SD units	p
F3 (Hz)	0.0301 (0.0094)	0.001	-0.0082 (0.0129)	0.522
Hammarberg index	0.0627 (0.0232)	0.006	-0.0087 (0.0119)	0.464
Alpha ratio	0.0386 (0.0222)	0.083	0.0083 (0.0133)	0.534
Energy ratio	0.1272 (0.0262)	1.3×10^{-6}	0.0009 (0.0089)	0.923
Spectral flux	0.1012 (0.0248)	4.5×10^{-5}	0.0126 (0.0116)	0.281
Voiced (segments/sec)	-0.0095 (0.0233)	0.682	0.0203 (0.0130)	0.119
Loudness (peaks/sec)	-0.0946 (0.0218)	1.5×10^{-4}	0.0196 (0.0148)	0.186
Spectral and voice quality PC1	0.2405 (0.0501)	1.6×10^{-6}	-0.0226 (0.0333)	0.498

Please see Table 6.1 for definitions of predictors; $N_{Participants}$ = 6093; $N_{stories}$ = 12; $N_{Observations}$ = 28694 recall pairs; The models were adjusted for age, gender, education, ethnicity, family history of dementia, psychological condition; Restricted to individuals with complete learning, immediate, delayed recall, age, gender, education, ethnicity, family history of dementia, psychological condition, and deprivation decile information; N=27819 observations; AIU - Automated Information Units; AOA - Age of acquisition; dB - decibel; HNR - Harmonics to Noise Ratio; Hz - Hertz; MFCC - Mel Frequency Cepstral Coefficient; PC1- First dimension of the principal component; SD - Standard Deviation; SE - Standard Error; Reference group is the least deprived group.

7.4 Discussion

7.4.1 Overall findings

7.4.1.1 Feasibility and acceptability

In this large-scale study, the feasibility of the Story Time task was assessed based on the participant's commitment and completeness to the task. In the most relevant age group (60-95), the proportion of participants who did not do the Story Time task was lower (26%) among those older participants. Among those participants who attempted at least one Story Time task, their mean compliance was 67.1% in Year-1 and 71.2% in Year-2. Although these participants did not complete all the tasks, on average, they completed enough to get a meaningful assessment (3.9 stories with a full set of learning, immediate, and delayed recall in Year-1). Older participants (60-95 years) were more committed to the Mezurio schedule. They were more likely to continue beyond the Story Time task than younger participants, agreeing with the results published in other studies [311], [312], where participants older than the age of 50 years contributed longer to the study. The older participant's higher commitment to the study compared to younger participants may be because they have a greater chance of being close to someone with Alzheimer's disease and are interested in digital gameplay apps perceived as "brain training" to improve cognition [313] or have more time on their hands due to being retired. Younger participant's lower commitment to the study may be because the Story Time task requires participants to be in a quiet place to speak to the microphone, unlike other tasks that could be completed while commuting to work or at the office.

7.4.1.2 Characteristics of AIU score: impact of the region, delay, time of the day, and reproducibility

The AIU scoring method developed in this thesis used the word embeddings that calculated semantic similarity between the words used in the participant responses to the Story Time tasks. In PREVENT (Chapter 6), MIU and AIU were compared, showing a high correlation and agreement between scores generated from these two methods. In this GameChanger study, a large sample gave an opportunity to explore the effect of region on AIU scores. The literature on automatic speech recognition has reported discrepancies with regional and sociolinguistic speech variants due to their differences in phonological, phonetic, and prosodic characteristics [262], [314]. In this study, recall scores varied by different stories rather than by region, suggesting that using the current ASR system in transcribing speech responses in the English language from various English regions and using NLP technologies are not a problem for scalable analysis. However, people might not live where they grew up, and their accents cannot be determined solely by the region where they live, suggesting that an extra question in the demographic questionnaire asking about the region of upbringing might be helpful in future studies.

The test-retest reliabilities of the Story Time sub-tasks and average recall scores measured using AIU score were between 0.70 - 0.79, showing moderate reliability, which is in the range (0.61-0.73) of California Verbal Learning task (CVLT=II), the established neuropsychological memory assessment task [315]. The reliability of the loss in AIU score (delta) was poor (Cronbach alpha = 0.32). The lack of reliability of the difference score has been well documented in the literature, suggesting that the lower reliability may be because this score has measurement errors from both immediate and delayed recall readings [316], [317].

GameChanger participants produced fewer information units (AIU score) in delayed compared to immediate recall. This effect of delay on memory was also seen in PREVENT (section 6.3.6) and other studies [239], justifying AIU as a valid measure of memory function. The literature on sex differences in episodic memory has revealed a female advantage in tasks where participants were asked to remember words or faces and a male advantage in spatial memory tasks [318]; however, such difference between the sexes has not been reported for story retelling tasks. In this study, there was a significant effect of sex on the AIU score, with female participants producing higher AIU recall compared to male participants. Several factors, such as mood, stress level, and sleep duration, are known to affect cognitive and memory performance [309], [319], [320], varying the performance throughout the day. In the GameChanger study, although participants were asked to complete Mezurio tasks in the morning hours, they were provided with the flexibility to complete Mezurio tasks on their own time. Participants completed Story Time tasks around the clock. The result suggested a significant influence of the time of the day on recalls, with better performance in the morning hours and decreased performance as the time increased. Similar results of reduced performance on memory in evening hours compared to morning hours were also found in another study [310]. This result shows that allowing participants to complete tasks in their leisure time adds to the variability of the measures compared to in-clinic tests, often done in the morning or early afternoon. However, despite this, the within-person variability is acceptable for the recalls.

7.4.1.3 Associations of the AIU score with socio-demographic variables and Mezurio cognitive scores

Higher age and more education were associated with higher recall AIU scores from the recalls cross-sectionally. Previous research has shown that older age and lesser education are associated with worst memory performance [303], [304], [321]. Additionally, participants in the most deprived group obtained lower AIU recall scores than those in the least deprived

group, suggesting that deprivation may impact an individual's cognitive ability. None of the studies in Chapter 3 directly assessed the association of deprivation with cognitive function assessed from speech recordings. However, deprivation has been found to be associated with dementia risk, with the influence of lower socioeconomic status on the diagnosis of dementia and mortality from dementia at an earlier age [322]. The result showed no evidence to suggest that a family history of dementia was associated with AIU scores, which was not surprising, given the cognitively healthy participants in the GameChanger study.

Although the correlation between the AIU recall score and Gallery Game recall accuracy score was in the expected direction, the association was weak. This may be because Story Time is a free recall task compared to object-pair associative memory task - Gallery Game, and these two measures may be capturing different constructs. The literature provides evidence of a weak or nonsignificant association between subjective cognitive decline and objective memory tests, supporting the weak correlation between SCD scores and AIU scores in the GameChanger participants [323]. This result suggested that SCD scores themselves may not clearly indicate objective episodic memory scores; however, these scores may provide information on the perceived decline in cognition at earlier stages and should be used in the evaluation of the relationship between cognitive status and memory performance [324], [325].

7.4.1.4 Associations of other language features with AIU score, socio-demographic variables and Mezurio cognitive scores

Among language features, there was a statistically significant relationship between the higher AIU score and the lexical diversity measured by higher unique words, higher type-token ratio, lower Brunet's index, and fewer repetitions. Similar findings have been reported where the associations between MMSE score (or clinical group based on MMSE score) and language features were evaluated [21], [52], [60], [326]. Earlier literature reported that the syntactic

features were well preserved in the later stage of AD [58], [151], [171], [327], with some studies contradicting these findings, showed that participants with AD had deterioration in close class words and propositional contents [111], [187]. In this study, syntactic features such as the higher proportion of nouns, verbs, adverbs, and adjectives to the total words, higher verbs per utterance and proportion of prepositions also showed a significant positive relationship with higher AIU score. The length of speech sample produced from the Story Time comics was longer (measured with average total words in learning: 199, immediate recall: 158, and delayed recall: 142, see **Table A.7**) than the standard picture description tasks (average of 100-150 words) [42], [49], which may have influenced such effects of syntactic features. Lower Brunet's index value is known to be associated with lexically rich responses, and Bucks et al. (2000) reported a negative association between MMSE score and Brunet's index obtained from picture description tasks containing a single picture [78]. The contrasting result was found in the Story Time task, where Brunet's index and the AIU recall score were positively associated. This might be because each story in the Story Time task contains nine comic strips, with story IUs that should be mentioned several times to create a flow in the story, unlike a single scene with fixed IUs as in the picture description task (e. g., Cookie theft picture). This result suggests that Brunet's index is not informative for measuring vocabulary/lexical richness in the story recall task. All the psycholinguistic features were significantly associated with the AIU scores, including AOA and frequency measures. These measures are mostly studied using a semantic fluency and picture naming task [328], [329]; however, the association of cognition (using MMSE score) and the association of AOA and frequency with AIU scores were in the same direction as observed in the GameChanger study [128]. The association of language features with the Gallery Game recall accuracy score were similar to the AIU score. The association of language features with the subjective cognitive decline score were in the opposite direction compared to their relationship with the AIU score, which was as expected.

The literature provides evidence of the association of lexical features with age, showing the speech response of the elderly participant contained more words and more repetitions compared to the younger group [330], [331]. In this study, lexical features, including total words, unique words, and Brunet's index, were associated with age, education, deprivation, Gallery Game recall accuracy score and SCD yearly scores. Among syntactic features, the use of verbs and past tense verbs showed negative associations with age cross-sectionally, evidenced by previous research that found that with increasing age, the use of past tense verbs declined [332].

7.4.1.5 Associations of audio features with AIU score, socio-demographic variables and Mezurio cognitive scores

Speech fluency measure has been found to decline more rapidly in various stages of AD, with increasing pause or hesitations [54], [132], [184], [216]. In this study, a higher speaking rate was associated with a higher AIU recall score showing that fluency in recall performance is sensitive to the memory retrieval [333]. Spectral and voice quality features represent voice quality and articulatory characteristics in speech, and some of these features were found to have significant associations in predicting the higher AIU recall score. Although interpreting these features in the context of cognition is difficult, they have been used in classifying healthy controls from the AD group [42], [105], [106]. These features capture spectral and articulatory information on the speech responses and indicate pathological speech [42], [102], [214], [334].

With the relationship with socio-demographic variables, speech fluency features, including speech duration and the number of syllables, were associated with age and education. Other speech fluency features, including the number of pauses, average pause duration, speech rate, and hesitation ratio, were associated with age, Gallery Game recall accuracy and SCD scores.

Overall, the longer speech duration and higher number of words were produced by older GameChanger participants; however, their AIU recall scores in story retelling significantly decreased with an increase in age cross-sectionally, supporting the finding by Saling et al. (2012) [330]. Spectral and voice quality features, including MFCC features, shimmer, HNR, and frequency features (F0, F1, F2, F3), also showed significant association with age, as expected because they are known to indicate vocal instability that has been found with an increase in age [335], [336]. Some of the features, including loudness, RMSE, MFCC, and shimmer, were associated with deprivation; however, associations of these features with deprivation are unknown.

7.4.1.6 Sensitivity analyses and data reduction approach

In the stepwise regression, pipeline features including lexical (average word length, unique words, total repeats), syntactic features (use of past tense verbs, proportion of close class words), speech fluency features (speech duration, average word duration, hesitation ratio), some of the spectral and voice quality features were included in the model predicting AIU recall score. Also, the model predicting Gallery Game recall accuracy score included lexical, speech fluency and AIU recall score as the most informative pipeline features.

In sensitivity analyses which included only the audio features at recall, the model explained less variability of AIU recall and the Gallery Game recall accuracy score compared to including all the pipeline features in the model, demonstrating the importance of language features. Earlier studies have reported that using both audio and language features provided better accuracy in classifying patient groups from healthy controls [199], [223]. The result from this study also found that using both these features explained more variability of cognition compared to using audio features alone.

The simple PCA approach was used as a method for dimensionality reduction; however, the first dimension of PCs was not strong compared to the individual features. There are several other methods for data reduction, including factor analysis (FA), linear discriminant analysis (LDA), singular value decomposition (SVD), Kernel PCA and multidimensional scaling (MDS) [337]. Using a more complicated method and obtaining PCs from sub-groups within the spectral and voice quality features should be considered in the future. Similarly, within the spectral and voice quality features, there were features related to frequency (F0, F1, F2, F3), voice intensity and quality and spectral energy and should be further sub-grouped for granular data reduction in future.

7.4.2 Strengths and limitations

This study has several strengths, including a diverse population with a wide range of ages and regions, other assessments including the Gallery Game and Subjective Cognitive Decline, repeated measures on each assessment wave and availability of follow-up data in Year-2. This analysis used multiple speech and language features at the individual level (repeated measures), allowing evaluation of the relationship between the well-established factors (age and education) associated with the risk of Alzheimer's disease and age-related cognitive changes. This analysis added value to the evaluations performed in Chapter 6 with a large sample size with the availability of additional socio-demographic factors. The results of this chapter explored various socio-demographic influences in the story-retelling task providing new insights into audio and language research using cognitive outcomes.

The findings must be considered with the certain limitations of the study. Firstly, the population in this study are all healthy participants, which prevents the comparison between various disease stages of AD. Female participation was higher (> 70%) in this GameChanger cohort,

and such gender bias has been reported in other citizen science studies [63], [145], [147], [163], [190], [219], [239], [312], suggesting future studies need to be careful in recruiting participants to reach a more diverse sample without gender bias. Another limitation is that some of the participants (n = 497) opened the Story Time task but did not provide any audio data, possibly because people felt uncomfortable providing their speech recordings. Although the participant's family history of dementia was recorded, the pre-clinical markers, such as amyloid and CSF-Tau, were not available in this study.

7.5 Summary

In this large prospective study of healthy adults without the diagnosis of any stages of dementia or Alzheimer's disease, the self-administered tasks, such as Story Time, showed moderate compliance and completeness to the task schedule. There was no apparent influence of region on the information unit-based score, indicating that the method to extract the AIU score developed in this thesis can be used for participants from any English region. The well-established influence of age and education on cognitive tests was also seen in the Story Time task indicating that it behaves as expected in a cognitive test. Informative features from among the approximately 50 pipeline features based on their associations with age, cognitive function score and other dementia risk factors were identified to suggest that these features should be used in future studies. These results should be of value to future research, showing the feasibility and utility of novel methods for use in a large-scale study for early AD detection to assess the decline in speech, language, and memory function more precisely.

Chapter 8. Discussion

8.1 Introduction

The burden of AD is significant, with the number of people living with dementia increasing at an alarming rate [1]. There is currently no cure for this disease, and the drugs developed to help the symptoms have shown limited benefits and were associated with adverse effects [7]. The heterogeneity of the AD population, with variation in disease progression, and the fact that the disease is diagnosed in its later stage when significant neuronal damage and cognitive impairments have already occurred pose a significant challenge in developing interventions that can achieve consistent effects. There may be more potential to prevent or slow disease at earlier stages. However, in such populations, the event rate of dementia-related outcomes will be low. The modest treatment effect size, in combination with a low event rate, will mean studies need to be very large and long in duration, and previous studies have often been underpowered due to small sample sizes [338], [339]. With the increasing interest in smartphone-based remote cognitive assessment deployed on a population level, a large amount of data can be collected. In this era of big data, scalable processing is essential in ensuring efficiency, effectiveness, and feasibility in handling and processing large input sizes. Feature extraction and analysis of large amounts of data manually is laborious, time-consuming, and costly. Therefore, automation and scalability are critical to improve efficiency, expand studies to include more participants, reduce human errors, and facilitate research collaboration and data sharing. In this thesis, a literature review on audio and language processing for cognitive assessment was undertaken (October 2020), which showed that the sample size of most studies was small (range 5-521; 75% of the studies with less than 100 participants) and either focused on audio or language features alone. The literature was in its infancy regarding the research on

the pre-clinical population. Although, with the more recent ADDRESo (Alzheimer's Dementia Recognition through Spontaneous Speech only) challenge dataset [340], containing 78 AD and 78 non-AD participants, recent studies applied the latest AI/ML techniques to enhance the classification of AD and non-AD participants [341]–[352], however, the sample size was still small, and the number of studies with the inclusion of the pre-clinical AD stage is very limited [86], [353], [354].

This doctoral research aimed to address the gaps in the sample size of the study, the use of a standardised feature set with both audio and language features, and the use of computational methods for scalability and automation to investigate the feasibility and usability of speech-based measures for early AD detection using a smartphone application. A cohesive feature extraction pipeline was developed by employing NLP, ASP, and ML/AI to extract audio and language features (Chapter 5). The accuracy of the computational method compared to the manual approach and its consistency with the clinical parameters were investigated in the pilot study (Chapter 6). Finally, the methods developed in Chapters 5 and 6 were evaluated in the large-scale study - Gamechanger (Chapter 7).

8.2 Scalability of computational methods in a large dataset

An automated feature extraction pipeline was developed in Python (Chapter 5) to extract audio and language features from speech recordings collected using the Story Time task within the Mezurio app and investigated its efficiency and scalability. PREVENT dataset was used to validate the computational methods developed in Chapter 5, which were evaluated in the GameChanger study. The quality of language features depends on the quality of transcripts. A systematic comparison of six ASR systems (including commercial and non-commercial) on the accuracy of English transcription was conducted to select the best available system. The

comparison of the ASR systems showed that Microsoft Azure provided the transcription outputs with substantially better accuracy (WER = 0.16) in comparison to KALDI (WER = 0.33), which is most commonly used in previous studies [43], [117], [355].

The ability to automate the process of scoring is critical to scalability. Many previous and recent studies have used manual scoring of information units [44], [172], [250], following the rule published by Nicholas and Brookshire (1993) [77]. The scoring method developed in this thesis employs a combination of ASR, NLP, ASP, and AI/ML for automation, which helped to reduce human biases and improve consistency in scoring. These methods consider the semantic neighbourhood, utilise part-of-speech tags, and automatically eliminate surrogate phrases to extract the information unit-based scores. AIU scores from this scoring method were compared with the manual approach of scoring (MIU). AIU scores showed a very high correlation (r = 0.98) and agreement with MIU scores demonstrating that automation has the potential to replace the manual approach to extract features related to informativeness. Information unit, also (known as a semantic content unit or information content unit in the literature) is one of the most widely researched semantic measures in AD research (see **Table 3.2**), with evidence that with the AD disease progression, the speech response becomes less informative, empty, and vague [114], [172], [186].

8.3 Evidence of feasibility

In both PREVENT and GameChanger studies, participants contributed to the research from their home environment by engaging with the Story Time task, making this task a feasible data collection method. In the GameChanger study, the Story Time task commencement was lower (Year 1: 59.8%, Year 2: 78.8%) compared to other tasks within the Mezurio smartphone app. The factors contributing to this lower commencement might be because the Story Time task

was administered towards the end of the Mezurio schedule (some participants dropped before reaching the Story Time task), the task required participants to be in a quiet place to speak to the microphone, unlike other tasks that could be completed during a commute to work or at the office. Older participants showed more commitment to the task suggesting that those in the study may have more leisure time (possibly due to retirement). Although participants with a family history of dementia might be expected to have better compliance as they may be more aware of the disease and have the risk of developing AD, the compliance result showed no such difference between participants with or without a family history of dementia (Table 7.3). The design of the Story Time task with parallel forms of stories introduced minimal practice effects (section 5.3.4), providing reliable cognitive estimates compared to other studies that found significant learning effects when applying the same cognitive battery for repeated testing [315], [356]. In this thesis, the test-retest reliability measured using Cronbach's alpha of learning, immediate, delayed, and average AIU recall scores were between 0.70-0.79, showing moderate to good reliability. In addition, the reliability of the loss in AIU score was poor (Cronbach's alpha = 0.32), demonstrating that the number of stories (n = 7) was insufficient to capture the difference in score. Such lack of reliability of the difference score has been well documented in the literature, suggesting the difference score has measurement error from both immediate and delayed recall readings [316], [317]. In both PREVENT and GameChanger studies, participants showed reduced recall performance on 24-hour delay compared to immediate recall on AIU score, suggesting this task can be used to measure episodic memory, a hallmark feature of AD [319].

8.4 Face validity

8.4.1 Association of AIU recall scores with in-clinic tests and other cognitive tests within the Mezurio app

There was a moderate correlation in the anticipated direction between recall AIU and in-clinic COGNITO narrative recall task in the PREVENT study, $r = 0.47$, $p = 0.02$. Although not statistically significant, there was a suggestion of an association between AIU and ACE-R ($r = 0.26$, $p = 0.22$). The statistical power to detect associations in this small study was limited.

In the GameChanger study, the association between AIU recall scores and Gallery Game was in the anticipated direction with weak effect size, $r = 0.2$, $p = 1.9 \times 10^{-72}$. The weak correlation between the two tasks may be because these two measures capture different constructs, the Story Time being a free recall task and Gallery Game being an object-pair associative memory task. The association between AIU recall scores and SCD scores was weak but in the expected direction, $r = -0.07$, $p = 1.7 \times 10^{-9}$. Sensitivity analysis showed that only a few features were informative of cognition, which included the number of syllables, AIU recall scores, use of unique words, total repetition, speech rate, and voice segments, with AIU recall score statistically significantly contributing to the model. This result agrees with the literature on IU being the most widely researched semantic feature in AD research [114], [172], [241].

8.4.2 Association of AIU recall scores with the pipeline features

There was a strong association between AIU recall scores and speech-based features. Participants with speech responses with higher unique words, higher lexical richness, and fewer repetitions were associated with higher AIU recall scores. Similarly, a higher speech rate, lower average pause duration, and lower average word duration were associated with a higher AIU

recall score showing that fluency in recall tasks is a marker of memory. In line with this finding, Hoffman et al. (2010) found that speech rate (measured as phoneme/sec) was positively correlated with cognition score (MMSE), indicating that the participant's speech becomes slower with lower cognition [184]. A recent study showed that average pause duration and average word duration were positively associated, whereas speech rate and repetition were negatively associated with clinician-rated word-finding difficulty rating [302].

8.4.3 Associations of pipeline features with the sociodemographic variables and other Mezurio tests

Statistically significant associations of pipeline features (audio and language) at recall with age and education cross-sectionally were found. At higher ages, participant's speech responses contained more words and more word repetitions, decreased use of verbs, and reduced speech rate with increased hesitation (**Table 7.18**). Such association of lexical and speech fluency features with age is consistent with previous studies that show the speech narration of the elderly group contains more words, more repetitions and a slower speech rate (measured with words/min) compared to the younger group [330], [331]. Despite producing more words and longer duration in the speech response at higher ages, the proportion of information units (the AIU recall score) significantly decreased with age cross-sectionally, supporting the finding by Saling et al. (2012), where the narratives of older adults appeared to be detailed but essential information of the story were missing from their story retellings [330]. Speech response with verbosity and empty content (non-informative words) is also one of the characteristics seen in participants with AD [46], [49]. Spectral and voice quality features, including MFCC features, Shimmer, HNR, and frequency features (F0, F1), also showed significant association with age, as expected because they are known to indicate vocal instability that is seen with an increase in age [335]. Participants with lower education produced shorter speech duration with fewer

unique words and obtained lower AIU recall scores than the participants with higher education. Similar to this finding, a previous study evaluating the effect of education level on the immediate recall of stories showed a positive association of story recall scores with education [203].

The participants in the GameChanger study were cognitively healthy, with 4940 (45%) participants having one of their parents with dementia. However, none of the audio and language features showed any statistically significant association with directly observable family history variable (**Table 7.20**). Socioeconomic deprivation is known to be associated with dementia risk [322]. In this study, participants in the most deprived group showed lower AIU recall scores indicating the influence of socioeconomic deprivation on their cognitive ability.

8.4.4 Applicability of the feature extraction pipeline

The Mezurio app has been deployed in several other studies, including the Remote Assessment of Disease And Relapse (RADAR-AD) [86], Deep and Frequent Phenotyping (DFP) [85], and Early Detection of Neurodegenerative diseases (EDoN) [87]. Out of these studies, RADAR-AD includes participants from various countries, and their speech response is collected in different languages. Although the feature extraction pipeline developed in this thesis supports the extraction of language features from speech responses in the English language only, the extraction of audio features does not have this limitation. Therefore, the language features of the pipeline, including the AIU score, would only be directly applicable with further work. In the sensitivity analysis using only the audio features that would be directly applicable to studies in other languages, the model explained only 4% of the variation of Gallery Game recall

accuracy scores compared to 7%, using all the pipeline features, highlighting the importance of language features, but still obtaining half of the value by using audio features alone.

Although ASR systems have made significant progress in recent years, ASRs trained in a larger dataset with more linguistic resources provide more accurate transcripts compared to those trained in languages with low-resource languages [357]. Current NLP research focusing on developing multilingual ASR shows some promise for future research in automatically extracting language features from speech recordings in different languages [358].

8.4.5 Preliminary analysis of audio features in the RADAR-AD cohort

Recently, a subset of the data from the RADAR-AD (the Remote Assessment of Disease And Relapse (recruitment completed at the end of January 2023; final n = 229) study [86] was made available, containing participants including healthy control (MMSE > 27, CDR = 0; negative Aβ biomarker if available), pre-clinical AD (MMSE > 26, CDR = 0, positive Aβ biomarker), prodromal AD (MMSE > 23, CDR = 0.5, positive Aβ biomarker), and Mild-to-moderate AD (MMSE > 17, CDR > 0.5, positive Aβ biomarker) recruited from fourteen different European countries. The study deployed the Mezurio app, where the Story Time task contained a try-it-now story (practice test during the initial setup of the app) and six test stories with learning, immediate and 30-minute delayed recall. Six Story Time tasks were scheduled on the 29th, 30th, 31st, 41st, 42nd, and 43rd days during the 8-week study duration. Further information on study design, methods, and baseline participant characteristics are provided in Appendix C. RADAR-AD cohort. Extraction of language features, including AIU score, required the responses in the English language, and there were only a few participants (n=15) recruited from the UK (London and Oxford sites) and (n=2) from Spain (Barcelona site) who provided story narration in English, restricting to study audio pipeline features only. The feature extraction pipeline was applied to 121 participants with Story Time data in the RADAR-AD study as in PREVENT

and GameChanger. The difference in speech fluency, spectral and voice quality features between clinical groups was examined univariately using linear mixed models with participants, story and the language in which the recording was made as random effects, adjusted for age, sex, and years of education to investigate whether these audio features can distinguish early disease groups from the healthy controls (**Table 8.1**).

Table 8.1 The difference in audio pipeline features on RADAR-AD clinical groups cross-sectionally.

	Pre-clinical AD (n=20)		Prodromal AD (n=33)		Mild-to-moderate AD (n=15)	
Outcomes	**Estimate per SD (SE)**	**p**	**Estimate per SD (SE)**	**p**	**Estimate per SD (SE)**	**p**
Speech fluency features						
Number of syllables (n)	0.4303 (0.2196)	0.052	-0.0219 (0.1864)	0.907	-0.6100 (0.2386)	0.012
Number of pauses (n)	0.2291 (0.2126)	0.284	-0.0631 (0.1792)	0.725	-0.2390 (0.2335)	0.308
Average pause duration (sec)	-0.0226 (0.1327)	0.865	0.1321 (0.1144)	0.252	0.8569 (0.1459)	9.8×10^{-08}
Speech rate (syllables/ sec)	-0.0813 (0.1956)	0.678	-0.3236 (0.1649)	0.052	-1.1316 (0.2143)	6.2×10^{-07}
Articulation rate (syllables/ sec)	-0.2777 (0.1849)	0.136	-0.3681 (0.1564)	0.020	-0.9122 (0.2023)	1.7×10^{-05}
Average syllable duration (sec)	0.0242 (0.0948)	0.799	0.1364 (0.0815)	0.103	0.1940 (0.1056)	0.074
Hesitation ratio	-0.1113 (0.1931)	0.565	0.1931 (0.1630)	0.239	0.7469 (0.2116)	0.001
Spectral and voice quality audio features						
MFCC 1	-0.4845 (0.2578)	0.063	0.0578 (0.2169)	0.790	-0.3146 (0.2839)	0.270
MFCC 2	0.0739 (0.2438)	0.762	0.2333 (0.2103)	0.269	0.0629 (0.2654)	0.813
MFCC 3	-0.3879 (0.1966)	0.051	-0.1099 (0.1658)	0.508	-0.1965 (0.2158)	0.364
MFCC 4	-0.2812 (0.1978)	0.158	0.0571 (0.1669)	0.733	-0.2736 (0.2172)	0.211
Jitter	0.0478 (0.1994)	0.811	0.2972 (0.1685)	0.081	-0.0559 (0.2182)	0.798
HNR (dB)	-0.1036 (0.1804)	0.567	0.1648 (0.1522)	0.282	0.5852 (0.1964)	0.004
F0 (Hz)	-0.1448 (0.1592)	0.365	0.1379 (0.1349)	0.309	0.3222 (0.1740)	0.067
F1 (Hz)	-0.2342 (0.1836)	0.205	0.1476 (0.1556)	0.345	0.2323 (0.2008)	0.250
F2 (Hz)	-0.3401 (0.1774)	0.059	0.1010 (0.1508)	0.505	-0.0557 (0.1939)	0.774
F3 (Hz)	-0.3456 (0.1829)	0.063	0.1099 (0.1554)	0.481	-0.1513 (0.2000)	0.452
Hammarberg index	-0.1407 (0.2255)	0.534	0.3841 (0.1914)	0.048	0.0991 (0.2449)	0.687
Alpha ratio	-0.1651 (0.2293)	0.473	-0.1616 (0.1931)	0.404	0.0853 (0.2519)	0.736
Energy ratio	-0.1590 (0.2803)	0.572	0.0943 (0.2380)	0.693	-0.0853 (0.3065)	0.781
Spectral flux	0.0339 (0.2352)	0.886	0.0941 (0.2015)	0.642	-0.4097 (0.2554)	0.112
Voiced (segments/sec)	-0.0909 (0.1873)	0.628	-0.3182 (0.1592)	0.048	-1.2233 (0.2026)	2.4×10^{-08}
Loudness (peaks/sec)	-0.0039 (0.1921)	0.984	-0.3292 (0.1635)	0.047	-1.1803 (0.2080)	1.1×10^{-07}
In-clinic test						
MMSE score	-0.0785 (0.1566)	0.617	-0.8168 (0.1315)	8.8×10^{-09}	-2.4232 (0.1733)	1.8×10^{-26}
Animal fluency score	0.2794 (0.2179)	0.202	-0.6649 (0.1877)	0.001	-1.4649 (0.2413)	1.8×10^{-08}

The reference group is healthy controls (n= 53); Please see Table 6.1 for definitions of variables; restricted to individuals with complete immediate, delayed recall, age, sex, and education information; 121 participants; N = 820 observations; languages = 11; dB - decibel; HNR - Harmonics to Noise Ratio; Hz - Hertz; MFCC - Mel Frequency Cepstral Coefficient; SD - Standard Deviation; SE - Standard Error; This approach involves a large number of statistical tests; therefore, the significance of results is interpreted using an approximate Bonferroni corrected p-value ($\alpha < .002$) to allow for multiple testing.

The preliminary result showed a significant difference between healthy controls and the mild-to-moderate AD group, with lower speech and articulation rates and higher average pause duration and hesitation ratio in the mild-to-moderate AD group. Mild-to-moderate AD participants had lower loudness and voice segments compared to healthy controls.

In-clinic MMSE test scores showed significant differences between healthy controls, prodromal, and mild-to-moderate AD. This result was expected because these test scores were used in grouping these clinical groups. Animal fluency scores between prodromal AD and healthy controls were also significant.

Audio features were not informative in the pre-clinical or prodromal AD group participants, which may be due to the small study size, highlighting the need for larger studies and a balanced dataset (roughly equal number of participants in each AD disease stage). RADAR-AD study has completed recruitment in late January 2023, and the full dataset should be available in the near future. Due to multilingual speech responses, language features, including AIU scores, were not extracted. This preliminary result highlighted the need to study language features along with the audio feature in these clinical groups.

As discussed in the previous section (8.4.4), current NLP research focuses on developing multilingual ASR with improved accuracy by utilising deep learning models [357], [358]. Also, word embedding models are publicly available only for a few European languages, and to train the Word2Vec model from scratch, a large corpus of language text is required. Also, substantially more work, including domain adaptation and validation, is necessary to improve the accuracy of the ASR system and generic Word2Vec models, as implemented in developing the feature extraction pipeline (Chapter 5).

8.5 Strengths and limitations of the study

There were several strengths and limitations of the study. The feature extraction pipeline developed in this thesis provided an extensive set of lexical, syntactic, paralinguistic, semantic, speech fluency, spectral, and voice quality features, which is beneficial for audio and language analysis as the analysts do not have to develop scripts to obtain specific features using separate libraries. The pipeline was also optimised for processing larger datasets and could be deployed in other large studies.

The small-sized PREVENT study allowed manual extraction of the AIU scores from the speech samples automatically, which is essential in accessing the quality of automation. Moreover, the manual and automated scores were highly correlated, showing the feasibility of extracting these measures automatically from the larger population. The in-clinic measures in PREVENT allowed investigation of the construct validity of the Story Time task using recall AIU score, suggesting a moderate correlation in the anticipated direction, although limited conclusions on these associations can be drawn from this small study. The GameChange study had several strengths, including the large study population, completeness, availability of Year-2 data, and repeated measures on each assessment wave (Year-1 and Year-2). This analysis used multiple speech and language features at the individual level (repeated measures), allowing evaluation of the relationship between the well-established factors (age and education) associated with the risk of Alzheimer's disease and age-related cognitive changes. This analysis added value to the evaluations performed in PREVENT with a much larger sample size, additional socio-demographic factors and Story Time pipeline features. The GameChanger chapter explored various socio-demographic influences in the story-retelling task providing new insights into using audio and language features in AD research.

The findings must be considered with the certain limitations of the study. Firstly, participants in this study were all healthy, and only cross-sectional associations could be studied. Due to the lack of disease status in both PREVENT and GameChanger studies, investigating whether the audio and language features distinguish early disease groups from healthy controls could not be done. While the results show promise, it is worth noting that the small sample size in the PREVENT study limited the power of the analyses for comparison against in-clinic tests. The other limitation of this study is that the Story Time task was optional. Some participants only started Story Time during the six-month assessment period due to operational reasons in the trial administration. Participants were aware of research with a high commitment to the dementia study, which might have affected the compliance. The participants were highly educated, which may have influenced the generalizability of the current findings. It is also necessary to assess the feasibility of using Story Time in a larger more diverse dataset.

In the GameChanger study, the participation bias was present, with a higher number of females in the study, similar to citizen science research and findings from the literature [241]. Although the participant's family history of dementia was available, the pre-clinical markers of AD, such as amyloid and CSF-Tau, were unavailable in this study. RADAR-AD study contained the disease status, including the participants in the pre-clinical AD group. However, only a subset of the data was available. As the feature extraction pipeline is applicable to the speech recorded in the English language, language features which included the AIU score could not be extracted.

8.6 Implications for future research

Following on from my thesis, some recommendations for future research are discussed in this section.

8.6.1 Validation of Story Time outcomes in cohorts with established biomarkers and cognitive measures

Mezurio has been deployed in other studies, including RADAR-AD [86], Deep and Frequent Phenotyping (DFP) (ongoing with the target to recruit 250 participants from across the UK) [85], and the Early Detection of Neurodegenerative diseases (EDoN) (currently ongoing; n = 216) [87], where biomarkers such as positron emission tomography (PET) imaging, cerebrospinal fluid (CSF), structural MRI markers of neurodegeneration, APOE, and tau protein (p-tau) have been measured along with the validated in-clinic paper-pencil tests including MMSE, CDR, ADCS-ADL, and AIADL. These studies use biomarkers, including p-tau and amyloid beta, to further stratify participants into pre-clinical AD groups [25]. Incorporating this information into the analysis would benefit the investigation of the distribution of audio and language characteristics in these populations.

Additionally, there are several studies such as PREVENT-Elicitation of Dialogues (PREVENT-ED) [73], PROSPECT-AD [354], The Speech on the Phone Assessment (SPeAK) [359], and others that capture speech using various data collection methods (for example, map description, picture description, and conversational speech). The feature extraction pipeline developed in this thesis (Chapter 5) could be applied in other studies that do not deploy the Story Time task, as the features, except for the AIU score, are independent of the specific task or protocol. To adapt the pipeline for studies that deploy other audio data collection protocols, minimal code writing is required, which includes adding a Python function to accept audio file inputs instead of parquet files (please refer to section 5.6.2).

8.6.2 Assessing the feasibility of the Story Time data extraction pipelines and outcome measures in other languages

The current Story Time feature extraction pipeline was developed targeting the UK-based English-speaking population. The linguistic features extracted in this study use the NLP techniques developed for English. The ASR system and word-embedding model for another language can be used in the existing feature extraction pipeline discussed in Chapter 5. However, an extensive comparison with all the available ASR systems in various languages must be performed, and the pipeline should use the best service with the fewest transcription errors. In the future, the analysis should not be restricted to just a single language, and the pipeline should be refined to be robust to other languages. This will be beneficial for analysing speech recordings collected from various European countries in the RADAR-AD and other future studies.

8.6.3 Assessing the value of Story Time reading variation

There is evidence of reduced reading ability in AD patients, marked by reduced speech and articulation rates and an increase in the number and proportion of pauses [95], [360], [361]. Although this thesis has explored the ability of participants to form their own stories (free recall) by looking at the comic strips, the reading variation is yet to be studied. GameChanger study Year-3 assessment wave contains Story Time reading variation where participants are shown comic strips and text to read in the learning sub-task. However, the start of this wave was delayed due to the COVID-19 outbreak, and due to the unavailability of data, the participant's reading ability could not be studied. Further investigation on Story Time reading variation could provide more insight into the value of reading ability in AD research.

8.7 Summary

In this thesis, a novel automated pipeline was developed using various emerging technologies to extract speech-based features of potential relevance to dementia using speech data collected via a smartphone application. This pipeline has the potential to facilitate much larger studies and trials through its scalability and wide deployment. Story Time task was a feasible method for the speech data collection, and the AIU recall score was found to be a potentially useful clinical assessment method for examining participants performance in a story-retelling task. The well-established influence of age and education on cognitive tests was also observed in the Story Time task, indicating its expected behaviour in a cognitive assessment. Furthermore, the associations between Story Time scores and comparable in-clinic assessments showed the potential utility of the task in measuring episodic memory. Future work can build on this research by refining the feature extraction pipeline to support multiple languages and validating speech-based features extracted from speech recordings on cohorts with pre-clinical AD biomarkers or a larger number of disease cases occurring.

Appendix A. Supplementary Tables and Figures

A.1 Supplementary Tables

Table A.1 PRISMA checklist for the systematic review of studies on audio and language-based cognitive assessment for Alzheimer's Disease (AD).

Section/topic	#	Checklist item	Reported
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	Yes
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	Yes
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	Yes
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	Yes
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	No
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	Yes
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	Yes

Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Yes
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	Yes
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	Yes
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Yes
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	No
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	Yes
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	No
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	No
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	Yes
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Yes
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Yes
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	No
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	No

Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	No
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	No
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	No
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	Yes
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	Yes
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	Yes
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	No

Table A.2 Word Error Rate (WER) for each of the Cloud-based ASR and open-source ASR system's outputs when compared to manual transcriptions. The lowest WER for a given recording is highlighted in bold.

Participant	Story	Word count	AWS BASE	AWS CLM	AWS CV	AZURE BASE	AZURE CLM	AZURE PHRASE	DEEP SPEECH BASE	DEEP SPEECH CLM	DEEP SPEECH WORD	GOOGLE BASE	GOOGLE ADAPT	IBM WATSON BASE	IBM WATSON CLM	IBM WATSON KW	KALDI BASE	KALDI CLM
P1	story1	80	0.29	0.24	0.28	**0.18**	**0.18**	**0.18**	0.49	0.50	0.49	0.24	0.28	0.30	0.19	0.23	0.38	0.23
P1	story2	157	0.23	0.20	0.24	**0.09**	**0.09**	0.10	0.28	0.27	0.28	0.19	0.19	0.22	0.13	0.22	0.22	0.16
P1	story3	171	0.34	0.29	0.33	0.23	**0.18**	0.23	0.38	0.35	0.38	0.32	0.31	0.33	0.29	0.37	0.39	0.35
P1	story4	151	0.31	0.27	0.31	**0.13**	**0.13**	0.17	0.52	0.43	0.52	0.33	0.37	0.29	0.24	0.33	0.36	0.29
P1	story5	153	0.27	0.24	0.27	0.18	**0.15**	0.18	0.48	0.40	0.48	0.29	0.26	0.30	0.24	0.34	0.41	0.31
P2	story1	79	0.49	0.35	0.48	0.29	**0.16**	0.43	0.78	0.68	0.78	0.72	0.71	0.66	0.49	0.68	0.70	0.38
P2	story2	215	0.27	0.20	0.25	0.14	**0.12**	0.16	0.60	0.42	0.60	0.60	0.59	0.37	0.20	0.43	0.33	0.15
P2	story3	119	0.73	0.67	0.68	0.47	**0.39**	0.49	0.84	0.74	0.84	0.72	0.67	0.68	0.69	0.76	0.76	0.63
P2	story4	199	0.42	0.37	0.41	0.20	**0.19**	0.21	0.60	0.56	0.60	0.46	0.45	0.37	0.38	0.45	0.40	0.34
P2	story5	162	0.36	0.34	0.36	**0.15**	0.17	0.21	0.61	0.51	0.61	0.71	0.67	0.44	0.49	0.58	0.58	0.54
P3	story1	143	0.23	0.17	0.23	**0.06**	0.07	**0.06**	0.43	0.39	0.43	0.35	0.36	0.15	0.16	0.28	0.24	0.20
P3	story2	171	0.22	0.20	0.22	0.07	**0.06**	0.11	0.43	0.36	0.43	0.26	0.26	0.22	0.23	0.27	0.29	0.20
P3	story3	167	0.21	0.18	0.20	0.13	**0.10**	0.16	0.40	0.28	0.40	0.29	0.31	0.21	0.19	0.27	0.24	0.24
P3	story4	163	0.11	0.13	0.11	**0.03**	0.04	0.04	0.28	0.23	0.28	0.25	0.23	0.20	0.14	0.21	0.14	0.15
P3	story5	52	0.33	0.29	0.33	**0.06**	0.08	0.08	0.40	0.42	0.40	0.44	0.38	0.21	0.15	0.29	0.17	0.19
P4	story1	97	0.28	0.30	0.24	**0.18**	0.19	0.18	0.80	0.71	0.80	0.43	0.43	0.67	0.45	0.49	0.62	0.38
P4	story2	91	0.35	0.29	0.34	0.15	**0.13**	0.18	0.79	0.75	0.79	0.37	0.38	0.53	0.36	0.43	0.45	0.35
P4	story3	76	0.16	0.12	0.14	**0.07**	0.09	0.08	0.55	0.47	0.55	0.29	0.24	0.32	0.20	0.36	0.29	0.21
P4	story4	50	0.22	0.22	0.22	0.10	**0.06**	**0.06**	0.68	0.68	0.68	0.24	0.20	0.50	0.20	0.20	0.38	0.34
P4	story5	151	0.17	0.15	0.18	0.16	**0.09**	0.13	0.60	0.50	0.60	0.31	0.30	0.38	0.17	0.38	0.33	0.20
P5	story1	357	0.28	0.26	0.28	**0.19**	**0.19**	**0.19**	0.57	0.66	0.57	0.46	0.44	0.41	0.38	0.44	0.44	0.43
P5	story2	313	0.29	0.27	0.28	**0.19**	0.23	0.21	0.62	0.67	0.62	0.58	0.54	0.45	0.41	0.43	0.44	0.44
P5	story3	377	0.45	0.40	0.41	0.42	**0.30**	0.38	0.73	0.76	0.73	0.72	0.69	0.56	0.52	0.59	0.56	0.53

Participant	Story	Word count	AWS BASE	AWS CLM	AWS CV	AZURE BASE	AZURE CLM	AZURE PHRASE	DEEP SPEECH BASE	DEEP SPEECH CLM	DEEP SPEECH WORD	GOOGLE BASE	GOOGLE ADAPT	IBM WATSON BASE	IBM WATSON CLM	IBM WATSON KW	KALDI BASE	KALDI CLM
P5	story4	353	0.33	0.32	0.34	**0.20**	0.22	**0.20**	0.61	0.65	0.61	0.44	0.44	0.47	0.42	0.48	0.45	0.41
P5	story5	365	0.37	0.32	0.36	**0.26**	0.28	0.28	0.65	0.71	0.65	0.56	0.57	0.55	0.42	0.53	0.56	0.52
Mean		176.5	0.31	0.27	0.29	0.17	**0.16**	0.18	0.56	0.52	0.56	0.42	0.41	0.39	0.31	0.40	0.41	0.33

Abbreviations - AWS: Amazon Web Server, CLM: Custom Language Model, CV: Custom Vocabulary, KW: Keywords

Table A.3 Some examples of surrogate phrases contained in the real story retelling data and were detected by the fine-tuned BERT model. The ones highlighted in bold with underlined text were not used as labels to fine-tune the model.

Examples	Detected surrogate phrases
1.	this lady she she she left she dropped her wallet she didn't notice and the child **my god** look what a **thing** she dropped her wallet she didn't notice and the child told her that it had happened **i just didn't say** that she went to the market
2.	a lady is gone shopping. and when she finished shopping she noticed that she had forgotten her wallet she was without her wallet and then she was very worried because how could she pay for the purchases then she came home and she had to leave shopping naturally **i don't know** if you said that but **i made up** the story then she had to leave she went home when she got home she rang the phone and it was a little girl who had found her wallet so she was very relieved because she found the wallet that's it
3.	a lady was shopping at the supermarket when she went absentmindedly absentmindedly, she did not notice that her purse wallet fell **i think** she picked **something** up so she dropped it she didn't notice and when she went to the cashier to pay she saw that she had no way to pay she left everything there and she left home upset is not then when she got home she rang the phone she was a person who had found the wallet a girl who naturally had her phone in her wallet she called to say that she had found her wallet then she was relieved **i imagine**
4.	**i'm confused** here as to which story you want me to recall so **i'll go back** to the one that i started on and told you yesterday but **i don't know** whether that's the right one once upon a time a boy called corky was asked by his father if he would go in

Table A.4 Definition and description of audio and language features extracted from text.

Feature Name	Feature Description
Lexical	
Total words (n)	Total words the participant stated in the speech response
Unique words (n)	The total number of unique or not repeated words the participants stated in the speech response
Average word length (n)	The average of word length in number of characters
Type Token Ratio (TTR)	Type (unique words) to token (total words)
Moving Average TTR	Average TTR in a moving window size (10, 20, 30)
Brunet's index	A measure of vocabulary richness
Honore statistics	Another measure of vocabulary richness
Syntactic	
Proportion of open-class words	Proportion of number of nouns, verbs, and adverbs the participants used
Proportion of close-class words	Proportion of number of function words i.e., prepositions, determiners, and conjunction the participants used
Pronoun Index	The number of pronouns divided by the total number of words
Verb per utterances	Proportion of verbs in speech response
Past tense verbs	Proportion of past tense verbs
Proportion of propositions	Proportion of propositions by total words
Proportion of determiners	Proportion of determiners by total words
Audio (from transcript)	
Words per minute (n)	The total number of words uttered per minute
Average word duration (sec)	Average word duration in second
Number of long pauses (n)	The total number of long pauses
Total long pauses (sec)	The duration of long pauses in the audio sample
Number of short pauses (n)	The total number of short pauses
Total short pauses (sec)	The duration of short pauses in the audio sample
Average pause duration (sec)	Average pause duration in the sample
Hesitation ratio	Ratio of pause duration and total audio duration
Semantic	
Automated Information Units	The number of IUs elicited divided by total IUs in the story
Information Density	The total number of IU's per total words
Psycholinguistic	
Age of Acquisition	The age at which a word was learnt.

Familiarity	A measure of the frequency of exposure to a word, e.g.: *'vertex'* would be low in familiarity, while *'zebra'* would be high)
Imageability	A semantic variable that measures how easy it is for a word to arouse mental images, e.g.: *'hope'* is low in imageability, while *'house'* is high
Concreteness	The degree to which a concept refers to a perceptible entity, e.g.: 'couch' is a concrete word, while 'ideal' is an abstract word

Table A.5 Description and definition of audio features.

Variables	Description	Definition
audio_number_of_syllables	Number of syllables in the recording	Number of syllables are extracted from the voiced segment using audio intensity (dB). Intensity finds peaks in the energy contour, as vowel within a syllable has higher energy than surrounding sounds. Intensity contour is used to make sure that the intensity between current peak and the preceding peak is sufficiently low. This helps us to remove multiple peaks within one syllable.
audio_number_of_pauses	Number of pauses in the recording	Number of silent pauses in the utterances
audio_average_pause_duration	Average pause duration in the recording (seconds)	Sum of pause durations divided by total number of pauses
audio_file_duration_in_second	Duration of recording (seconds)	Total time taken to complete a task
audio_total_speech_duration	Total speech duration (seconds)	Sum of speech duration excluding pauses
audio_speaking_rate	Audio speaking rate from number of syllables	Number of syllables per second calculated from the total duration
audio_articulation_rate	Articulation rate from number of syllables	Number of syllables per second calculated from the speech duration
audio_average_syllable_duration	Average syllable duration (seconds)	Sum of syllable duration divided by total number of syllables
audio_rms_energy	RMS Energy	Root mean square of the total magnitude of the audio signal
audio_hesitation_ratio	Hesitation ratio (pause duration/audio duration)	Ratio of total duration of pauses to total speech duration
ost_F0semitoneFrom27_5Hz_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 1	Pitch - logarithmic F0 on a semitone frequency scale (mean)
ost_F0semitoneFrom27_5Hz_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 2	Pitch - logarithmic F0 on a semitone frequency scale (sd)
ost_F0semitoneFrom27_5Hz_sma3nz_percentile20_0	Paralinguistic OpenSmile GeMaps Feature 3	Pitch - logarithmic F0 on a semitone frequency scale (20 percentile)
ost_F0semitoneFrom27_5Hz_sma3nz_percentile50_0	Paralinguistic OpenSmile GeMaps Feature 4	Pitch - logarithmic F0 on a semitone frequency scale (50 percentile)
ost_F0semitoneFrom27_5Hz_sma3nz_percentile80_0	Paralinguistic OpenSmile GeMaps Feature 5	Pitch - logarithmic F0 on a semitone frequency scale (80 percentile)

ost_F0semitoneFrom27_5Hz_sma3nz_pctlrange0_2	Paralinguistic OpenSmile GeMaps Feature 6	Pitch - logarithmic F0 on a semitone frequency scale (range)
ost_F0semitoneFrom27_5Hz_sma3nz_meanRisingSlope	Paralinguistic OpenSmile GeMaps Feature 7	Pitch - logarithmic F0 on a semitone frequency scale from rising slope (mean)
ost_F0semitoneFrom27_5Hz_sma3nz_stddevRisingSlope	Paralinguistic OpenSmile GeMaps Feature 8	Pitch - logarithmic F0 on a semitone frequency scale from rising slope (sd)
ost_F0semitoneFrom27_5Hz_sma3nz_meanFallingSlope	Paralinguistic OpenSmile GeMaps Feature 9	Pitch - logarithmic F0 on a semitone frequency scale from falling slope (mean)
ost_F0semitoneFrom27_5Hz_sma3nz_stddevFallingSlope	Paralinguistic OpenSmile GeMaps Feature 10	Pitch - logarithmic F0 on a semitone frequency scale from falling slope (sd)
ost_loudness_sma3_amean	Paralinguistic OpenSmile GeMaps Feature 11	Mean value of approximation of perceived signal intensity
ost_loudness_sma3_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 12	Approximation of perceived signal intensity (sd)
ost_loudness_sma3_percentile20_0	Paralinguistic OpenSmile GeMaps Feature 13	Approximation of perceived signal intensity (percentile 20)
ost_loudness_sma3_percentile50_0	Paralinguistic OpenSmile GeMaps Feature 14	Approximation of perceived signal intensity (percentile 50)
ost_loudness_sma3_percentile80_0	Paralinguistic OpenSmile GeMaps Feature 15	Approximation of perceived signal intensity (percentile 80)
ost_loudness_sma3_pctlrange0_2	Paralinguistic OpenSmile GeMaps Feature 16	Approximation of perceived signal intensity (range)
ost_loudness_sma3_meanRisingSlope	Paralinguistic OpenSmile GeMaps Feature 17	Approximation of perceived signal intensity on rising slope (mean)
ost_loudness_sma3_stddevRisingSlope	Paralinguistic OpenSmile GeMaps Feature 18	Approximation of perceived signal intensity on rising slope (sd)
ost_loudness_sma3_meanFallingSlope	Paralinguistic OpenSmile GeMaps Feature 19	Approximation of perceived signal intensity on falling slope (mean)
ost_loudness_sma3_stddevFallingSlope	Paralinguistic OpenSmile GeMaps Feature 20	Approximation of perceived signal intensity on falling slope (sd)
ost_spectralFlux_sma3_amean	Paralinguistic OpenSmile GeMaps Feature 21	Difference of the spectra of two consecutive frames - also known as voice timbre (mean)
ost_spectralFlux_sma3_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 22	Difference of the spectra of two consecutive frames - also known as voice timbre (sd)
ost_mfcc1_sma3_amean	Paralinguistic OpenSmile GeMaps Feature 23	Mel-Frequency Cepstral Coefficients 1 (mean)
ost_mfcc1_sma3_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 24	Mel-Frequency Cepstral Coefficients 1 (sd)
ost_mfcc2_sma3_amean	Paralinguistic OpenSmile GeMaps Feature 25	Mel-Frequency Cepstral Coefficients 2 (mean)
ost_mfcc2_sma3_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 26	Mel-Frequency Cepstral Coefficients 2 (sd)
ost_mfcc3_sma3_amean	Paralinguistic OpenSmile GeMaps Feature 27	Mel-Frequency Cepstral Coefficients 3 (mean)

ost_mfcc3_sma3_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 28	Mel-Frequency Cepstral Coefficients 3 (sd)
ost_mfcc4_sma3_amean	Paralinguistic OpenSmile GeMaps Feature 29	Mel-Frequency Cepstral Coefficients 4 (mean)
ost_mfcc4_sma3_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 30	Mel-Frequency Cepstral Coefficients 4 (sd)
ost_jitterLocal_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 31	Variation of frequency from one period to the next period (mean)
ost_jitterLocal_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 32	Variation of frequency from one period to the next period (sd)
ost_shimmerLocaldB_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 33	Variation of amplitude from one period to the next period (mean)
ost_shimmerLocaldB_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 34	Variation of amplitude from one period to the next period (sd)
ost_HNRdBACF_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 35	Relation of energy in harmonic components to energy in noise-like components (mean)
ost_HNRdBACF_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 36	Relation of energy in harmonic components to energy in noise-like components (sd)
ost_logRelF0_H1_H2_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 37	Mean ratio of energy of the first F0 harmonic (H1) to the energy of the second F0 harmonic (H2)
ost_logRelF0_H1_H2_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 38	Ratio of energy of the first F0 harmonic (H1) to the energy of the second F0 harmonic (H2) (sd)
ost_logRelF0_H1_A3_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 39	Ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3) (mean)
ost_logRelF0_H1_A3_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 40	Ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3) (sd)
ost_F1frequency_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 41	Mean of Centre frequency of first formant
ost_F1frequency_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 42	Standard Deviation of Centre frequency of first formant
ost_F1bandwidth_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 43	Mean of Bandwidth of first formant
ost_F1bandwidth_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 44	Standard deviation of Bandwidth of first formant
ost_F1amplitudeLogRelF0_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 45	Mean Amplitude of first formant
ost_F1amplitudeLogRelF0_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 46	Standard deviation of amplitude of first formant
ost_F2frequency_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 47	Mean of Centre frequency of second formant
ost_F2frequency_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 48	Standard Deviation of Centre frequency of second formant
ost_F2bandwidth_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 49	Mean of Bandwidth of second formant

ost_F2bandwidth_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 50	Standard deviation of Bandwidth of second formant
ost_F2amplitudeLogRelF0_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 51	Mean Amplitude of second formant
ost_F2amplitudeLogRelF0_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 52	Standard deviation of amplitude of second formant
ost_F3frequency_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 53	Mean of Centre frequency of third formant
ost_F3frequency_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 54	Standard Deviation of Centre frequency of third formant
ost_F3bandwidth_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 55	Mean of Bandwidth of third formant
ost_F3bandwidth_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 56	Standard deviation of Bandwidth of third formant
ost_F3amplitudeLogRelF0_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 57	Mean Amplitude of third formant
ost_F3amplitudeLogRelF0_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 58	Standard deviation of amplitude of third formant
ost_alphaRatioV_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 59	Mean Ratio of summed energy in voiced region from 50-1000 Hz and 1-5 KHz
ost_alphaRatioV_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 60	Standard deviation of Ratio of summed energy in voiced region from 50-1000 Hz and 1-5 KHz
ost_hammarbergIndexV_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 61	Mean Ratio of the strongest voiced energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region
ost_hammarbergIndexV_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 62	Standard deviation of ratio of the strongest voiced energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region
ost_slopeV0_500_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 63	Linear regression slope of the logarithmic power spectrum within 0-500 Hz (mean)
ost_slopeV0_500_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 64	Linear regression slope of the logarithmic power spectrum within 0-500 Hz (sd)
ost_slopeV500_1500_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 65	Linear regression slope of the logarithmic power spectrum within 500-1500Hz (mean)
ost_slopeV500_1500_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 66	Linear regression slope of the logarithmic power spectrum within 500-1500Hz (Std. Dev.)
ost_spectralFluxV_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 67	Difference of the spectra of two consecutive frames (mean)
ost_spectralFluxV_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 68	Difference of the spectra of two consecutive frames (sd)
ost_mfcc1V_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 69	Mel-Frequency Cepstral Coefficients 1 in voiced regions (mean)

ost_mfcc1V_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 70	Mel-Frequency Cepstral Coefficients 1 in voiced regions (sd)
ost_mfcc2V_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 71	Mel-Frequency Cepstral Coefficients 2 in voiced regions (mean)
ost_mfcc2V_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 72	Mel-Frequency Cepstral Coefficients 2 in voiced regions (sd)
ost_mfcc3V_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 73	Mel-Frequency Cepstral Coefficients 3 in voiced regions (mean)
ost_mfcc3V_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 74	Mel-Frequency Cepstral Coefficients 3 in voiced regions (sd)
ost_mfcc4V_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 75	Mel-Frequency Cepstral Coefficients 4 in voiced regions (mean)
ost_mfcc4V_sma3nz_stddevNorm	Paralinguistic OpenSmile GeMaps Feature 76	Mel-Frequency Cepstral Coefficients 4 in voiced regions (sd)
ost_alphaRatioUV_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 77	Mean Ratio of summed energy in voiced and unvoiced region from 50-1000 Hz and 1-5 KHz
ost_hammarbergIndexUV_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 78	Mean Ratio of the strongest voiced and unvoiced energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region
ost_slopeUV0_500_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 79	Linear regression slope of the logarithmic power spectrum within 0-500 Hz in voiced and unvoiced region (mean)
ost_slopeUV500_1500_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 80	Linear regression slope of the logarithmic power spectrum within 500-1500Hz in voiced and unvoiced region (mean)
ost_spectralFluxUV_sma3nz_amean	Paralinguistic OpenSmile GeMaps Feature 81	Difference of the spectra of two consecutive voiced and unvoiced frames (mean)
ost_loudnessPeaksPerSec	Paralinguistic OpenSmile GeMaps Feature 82	Loudness measures as number of peaks per second
ost_VoicedSegmentsPerSec	Paralinguistic OpenSmile GeMaps Feature 83	Voiced segments per second
ost_MeanVoicedSegmentLengthSec	Paralinguistic OpenSmile GeMaps Feature 84	Mean value of voiced segment length in seconds
ost_StddevVoicedSegmentLengthSec	Paralinguistic OpenSmile GeMaps Feature 85	Standard deviation of voiced segments
ost_MeanUnvoicedSegmentLength	Paralinguistic OpenSmile GeMaps Feature 86	Mean unvoiced (silent) segments
ost_StddevUnvoicedSegmentLength	Paralinguistic OpenSmile GeMaps Feature 87	Standard deviation of unvoiced segments
ost_equivalentSoundLevel_dBp	Paralinguistic OpenSmile GeMaps Feature 88	Sound level measured in decibels - higher the decibel level, the louder the noise

Table A.6 The proportions of participants completing each Story Time learning, immediate, and delayed recall distributed across the GameChanger assessment periods (Year 1 and Year 2).

Assessment period	Story Index [a]	(No. of participants [c])	Completion [b] (%)			
			Learning	Immediate Recall	Delayed Recall	Recall Pairs [d]
Year1	1	7647	98.9	91.0	67.8	63.7
	2	6747	98.7	92.4	70.6	67.0
	3	6422	98.7	85.1	69.8	62.2
	4	5964	98.9	83.3	68.8	60.5
	5	5660	98.8	81.6	71.9	62.2
	6	5373	98.9	82.3	73.2	63.5
	7	5289	98.9	81.9	74.6	64.6
Year2	8	2556	98.6	93.0	70.9	67.5
	9	2439	97.9	89.1	75.3	68.5
	10	2477	98.1	85.6	76.4	67.1
	11	2352	98.6	85.9	79.9	70.8
	12	2383	98.7	86.2	78.3	70.2

Values represent the proportion of participants completing each Story Time learning, immediate recall, delayed recall, and proportion of recall pairs. [a] Stories in order as they appear in the Mezurio Story Time task. [b] Completion is calculated for each participant as the proportion of completed task schedules. Successful completion involved an audio recording with at least five words in transcription output. [c] These are all the participants who attempted the Story Time task, but they might not have completed it successfully, as stated in [b]. [d] Immediate and delayed recall pairs.

Table A.7 Mean (SD) of each pipeline feature at each of the three subtasks among participants and stories they completed.

Features	Learning	Immediate Recall	Delayed Recall
	Mean (SD)	Mean (SD)	Mean (SD)
Semantic features			
AIU score	0.56 (0.14)	0.50 (0.12)	0.42 (0.14)
Information density	0.09 (0.03)	0.10 (0.03)	0.09 (0.03)
Lexical features			
Total words	198.88 (101.87)	158.14 (67.89)	141.91 (68.67)
Unique words	102.12 (36.46)	83.24 (26.97)	75.80 (27.91)
Moving average type token ratio	0.94 (0.02)	0.94 (0.02)	0.93 (0.02)
Brunets' index	11.17 (1.17)	11.04 (1.03)	10.74 (1.18)
Honore’s stats	1736.62 (334.10)	1595.59 (315.93)	1613.25 (350.74)
Average word length	3.77 (0.13)	3.77 (0.13)	3.77 (0.15)
Total repeats	1.22 (1.60)	0.92 (1.10)	0.81 (1.00)
Syntactic features			
Open class words	0.51 (0.03)	0.49 (0.03)	0.49 (0.04)
Close class words	0.21 (0.03)	0.21 (0.03)	0.21 (0.04)
Pronoun index	0.10 (0.02)	0.11 (0.03)	0.11 (0.03)
Verbs per utterances	0.22 (0.02)	0.22 (0.02)	0.22 (0.02)
Past tense verb	0.44 (0.18)	0.46 (0.19)	0.52 (0.17)
Number of prepositions	0.10 (0.02)	0.10 (0.02)	0.10 (0.02)
Number of determiners	0.11 (0.02)	0.11 (0.03)	0.11 (0.03)
Psycholinguistic features			
AOA	4.97 (0.24)	4.95 (0.27)	4.94 (0.30)
Familiarity	215.40 (14.92)	221.95 (17.30)	220.16 (22.12)
Concreteness	222.71 (16.57)	228.74 (18.79)	226.54 (23.48)
Imageability	315.57 (17.28)	321.81 (19.68)	318.96 (26.89)
Speech fluency features			
Recording duration (sec)	60.40 (30.07)	45.07 (20.66)	41.68 (21.43)
Number of syllables (n)	243.91 (123.22)	186.58 (83.98)	169.56 (87.67)
Number of pauses (n)	37.16 (20.17)	26.40 (13.78)	24.77 (13.84)
Average pause duration (sec)	1.48 (0.74)	0.89 (0.36)	0.94 (0.51)
Average word duration (sec)	0.30 (0.04)	0.29 (0.04)	0.30 (0.04)
Speech rate (syllables/ sec)	2.20 (0.49)	2.79 (0.54)	2.68 (0.55)
Articulation rate (syllables/ sec)	4.07 (0.56)	4.17 (0.51)	4.08 (0.54)
Average syllable duration (sec)	0.34 (1.06)	0.31 (0.98)	0.34 (1.44)
Hesitation ratio	0.48 (0.10)	0.38 (0.10)	0.40 (0.11)
Spectral and voice quality audio features			
RMS energy	0.02 (0.02)	0.02 (0.03)	0.02 (0.02)
Loudness (dB)	0.15 (0.11)	0.18 (0.13)	0.17 (0.13)
MFCC 1	37.85 (7.37)	38.05 (7.39)	38.31 (7.22)
MFCC 2	4.19 (9.40)	4.88 (9.40)	5.79 (9.33)

MFCC 3	11.21 (8.64)	11.85 (8.56)	12.65 (8.41)
MFCC 4	-6.40 (7.58)	-6.28 (7.61)	-5.51 (7.65)
Jitter	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)
Shimmer (dB)	1.15 (0.14)	1.15 (0.15)	1.12 (0.15)
HNR (dB)	7.20 (1.67)	7.12 (1.67)	7.26 (1.72)
F0 (Hz)	31.22 (3.95)	30.83 (3.92)	30.60 (3.94)
F1 (Hz)	656.36 (55.60)	647.96 (56.48)	647.24 (57.90)
F2 (Hz)	1729.31 (65.12)	1724.17 (69.61)	1724.49 (72.40)
F3 (Hz)	2802.69 (78.64)	2797.93 (86.93)	2799.39 (90.62)
Hammarberg index	28.18 (4.36)	28.27 (4.33)	28.60 (4.24)
Alpha ratio	-9.12 (3.49)	-8.84 (3.27)	-8.79 (3.19)
Energy ratio	0.02 (0.02)	0.03 (0.02)	0.03 (0.02)
Spectral flux	0.03 (0.03)	0.04 (0.04)	0.04 (0.05)
Sound level (dB)	-39.93 (9.73)	-38.63 (9.75)	-38.91 (9.75)
Voiced (segments/sec)	1.66 (0.36)	2.02 (0.38)	1.91 (0.39)
Loudness (peaks/sec)	1.99 (0.47)	2.56 (0.54)	2.44 (0.56)

Please see Table 6.1 for definitions of features; AIU - Automated Information Units; AOA - Age of acquisition; dB - decibel; HNR - Harmonics to Noise Ratio; Hz - Hertz; MFCC - Mel Frequency Cepstral Coefficient; SD - Standard Deviation.

Table A.8 Baseline characteristics of individuals with and without a family history of dementia.

Characteristics	**FH (-)** [a]	**FH (+)** [b]	**Overall**	**p**
n	5470	4940	10865	
Age (years) (mean (SD))	52.1 (14.1)	55.7 (10.6)	53.8 (12.7)	1.1×10^{-50}
Age Group, %				5.7×10^{-82}
16-39	21.0	7.6	14.6	
40-59	45.6	54.6	49.9	
60-95	33.4	37.8	35.5	
Sex, %				1.2×10^{-6}
Female	78.5	82.3	80.3	
Male	21.5	17.7	19.7	
Education, %				8.6×10^{-9}
Higher	57.9	52.2	55.2	
Lower	42.1	47.8	44.8	
Ethnicity, %				0.240
Non-White	5.4	4.9	5.2	
White	94.6	95.1	94.8	
Neurological Condition, %				0.005
No	97.4	98.2	97.8	
Yes	2.6	1.8	2.2	
Psychological Condition, %				0.624
No	76.9	76.5	76.7	
Yes	23.1	23.5	23.3	
Multiple Deprivation Decile, %				0.014
Decile 1-2 (Lowest Deprivation)	8.0	6.9	7.5	
Decile 3-4	13.7	12.4	13.1	
Decile 5-6	20.6	20.4	20.5	
Decile 7-8	25.3	25.2	25.3	
Decile 9-10 (Highest Deprivation)	32.3	35.2	33.6	

Note: Table shows family history data from 10865 participants, excluding 371 participants who reported being "unsure" on their family history of dementia; [a] No first-degree of family history of dementia; [b] First degree of family history of dementia; p-values are obtained from t-test for continuous and χ^2 test for categorical variables.

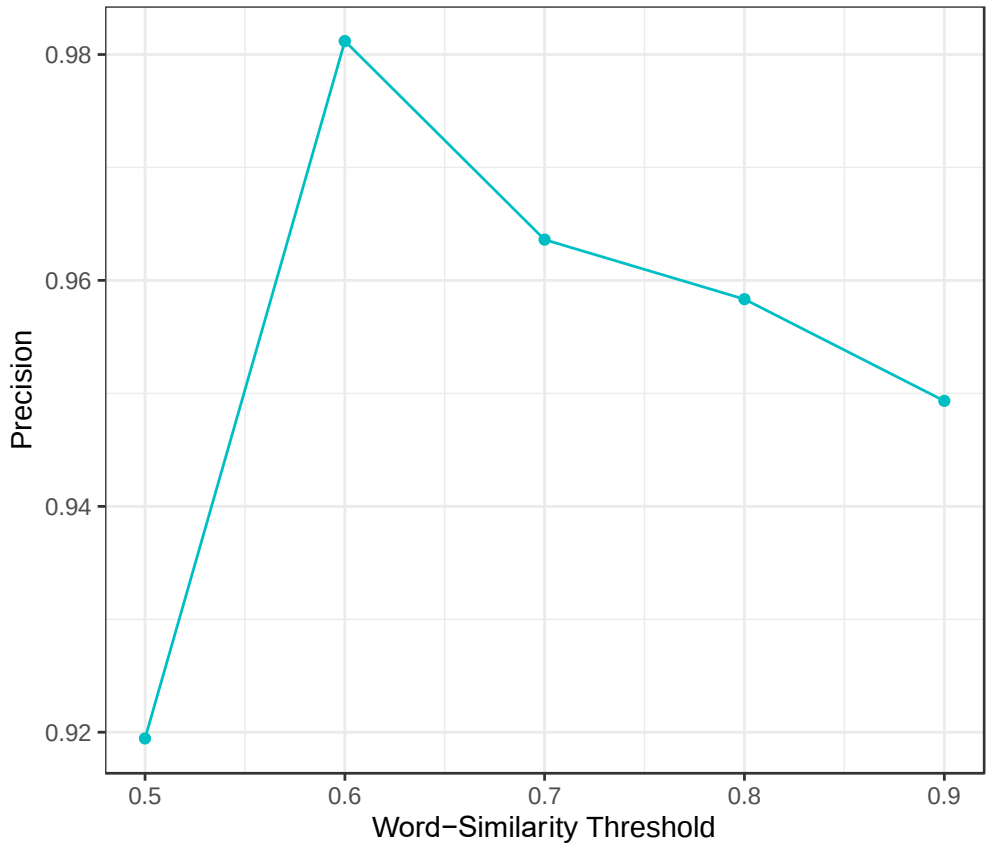
Table A.9 Baseline characteristics of individuals in various socioeconomic deprivation deciles.

Characteristics	**English Multiple Deprivation Deciles [a]**			
	Most deprived	**Least deprived**	**Overall**	**p**
n	2711	6248	8959	
Age (years) (mean (SD))	51.85 (13.30)	54.63 (12.21)	53.79 (12.62)	1.9×10^{-20}
Age Group, %				4.3×10^{-22}
16-39	19.9	12.3	14.6	
40-59	49.1	50.1	49.8	
60-95	31.0	37.6	35.6	
Sex, %				1.2×10^{-6}
Female	83.4	78.5	80.0	
Male	16.6	21.5	20.0	
Education, %				8.6×10^{-9}
Higher	50.7	56.7	54.9	
Lower	49.3	43.3	45.1	
Ethnicity, %				0.011
Non-White	6.2	4.9	5.3	
White	93.8	95.1	94.7	
Family history of dementia, %				0.009
No	53.1	49.6	50.7	
Not Sure	3.3	3.3	3.3	
Yes	43.6	47.1	46.0	
Neurological Condition, %				0.669
No	97.6	97.8	97.7	
Yes	2.4	2.2	2.3	
Psychological Condition, %				6.6×10^{-13}
No	71.9	78.9	76.8	
Yes	28.1	21.1	23.2	

Most deprived are coded for the first five deciles, and the remaining deciles are coded as least deprived; [a] English deprivation deciles obtained from postcode lookup from the Ministry of Housing, Communities and Local Government, data available for 9252 participants; p-values are obtained from t-test for continuous and χ^2 test for categorical variables.

A.2 Supplementary Figures

Figure A.1 Word similarity threshold result based on precision. Precision was calculated as the proportion of word pairs that were identified by the Word2Vec model that are indeed words for the same story IU.



Note: This empirical evaluation involved cosine similarity of 3986 word-pairs.

Figure A.2 Association of in-clinic test scores and age.

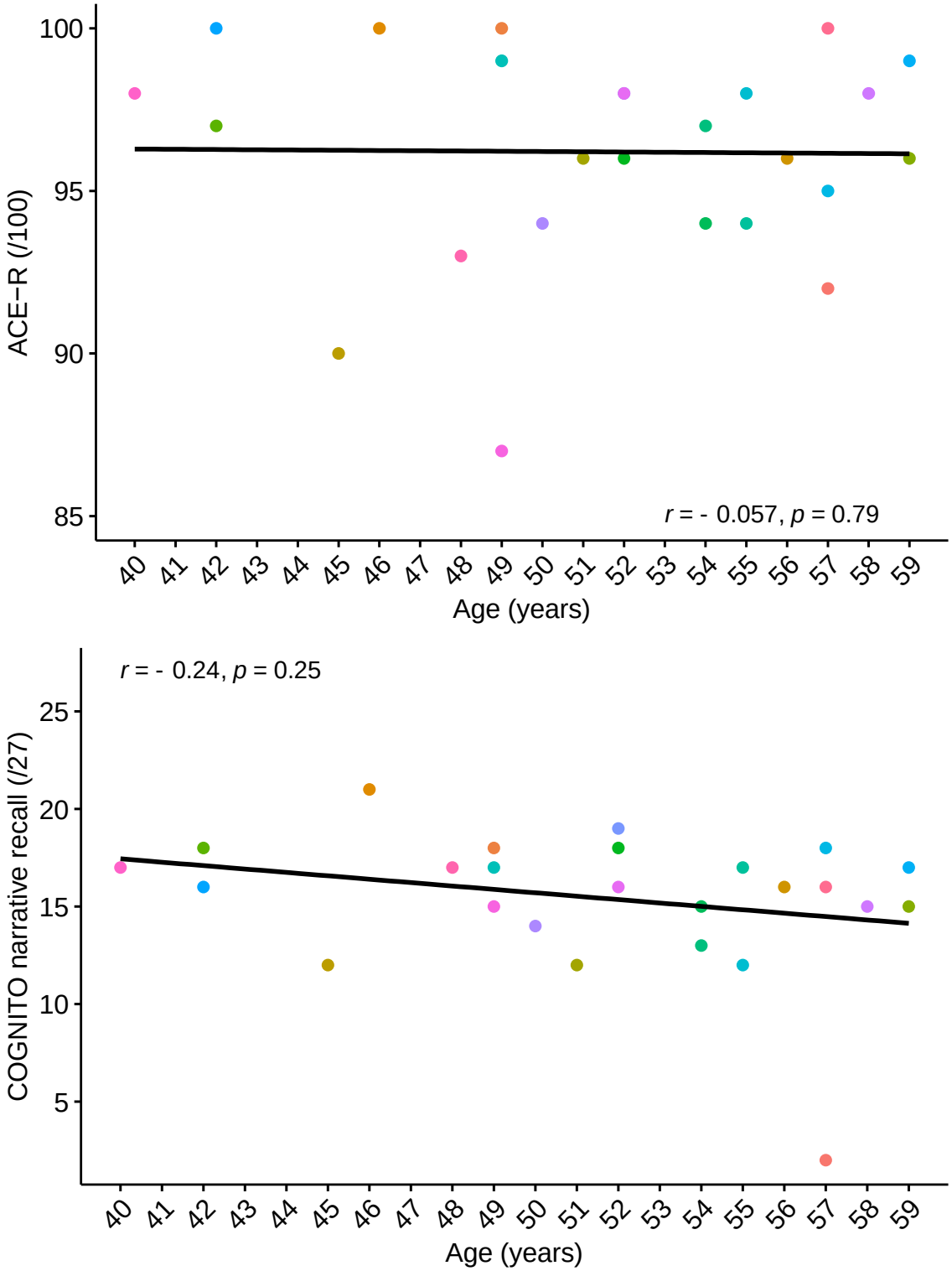


Figure A.3 Association of in-clinic test scores and years of education.

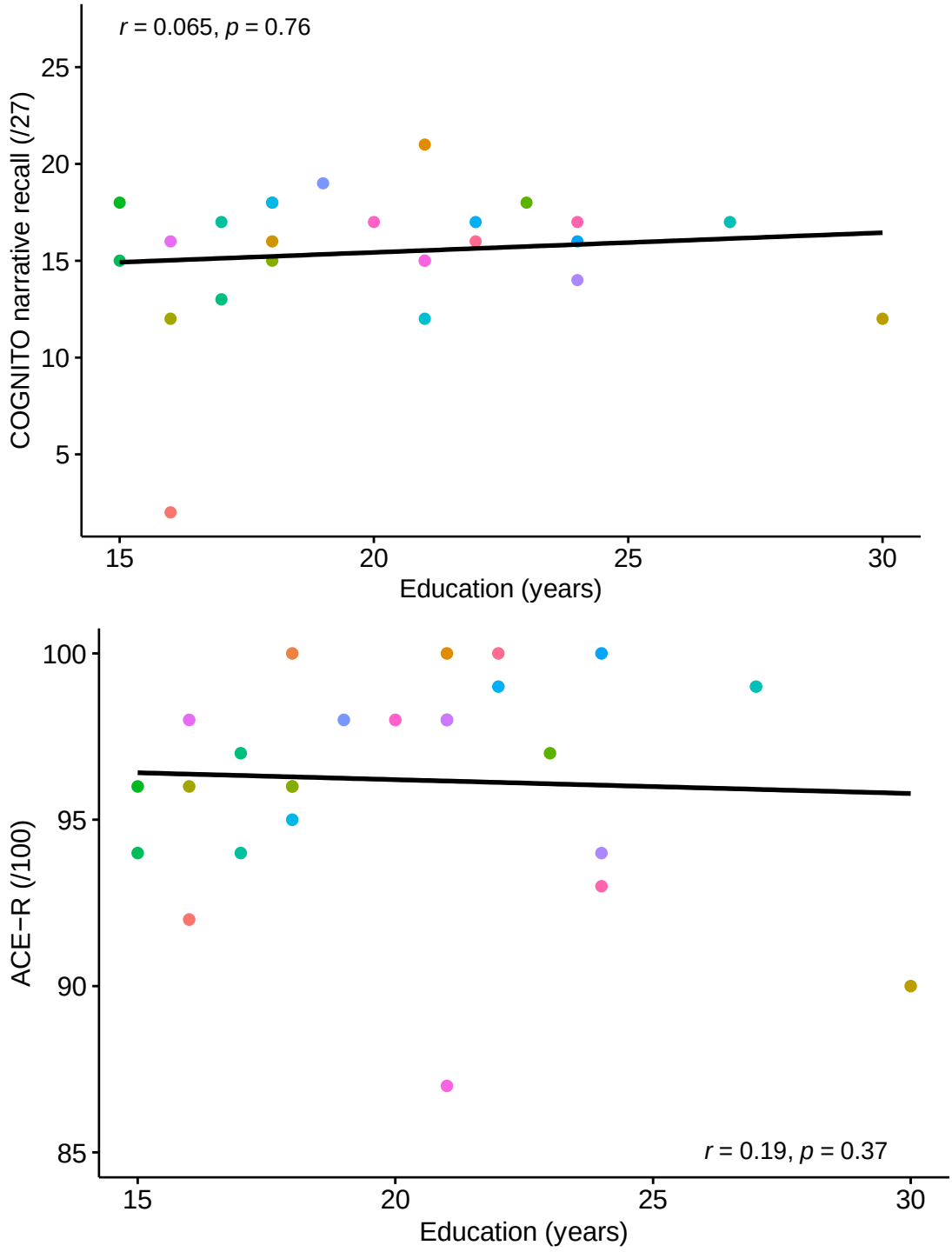


Figure A.4 Density plot showing the frequency distribution of various audio features of each PREVENT participant.

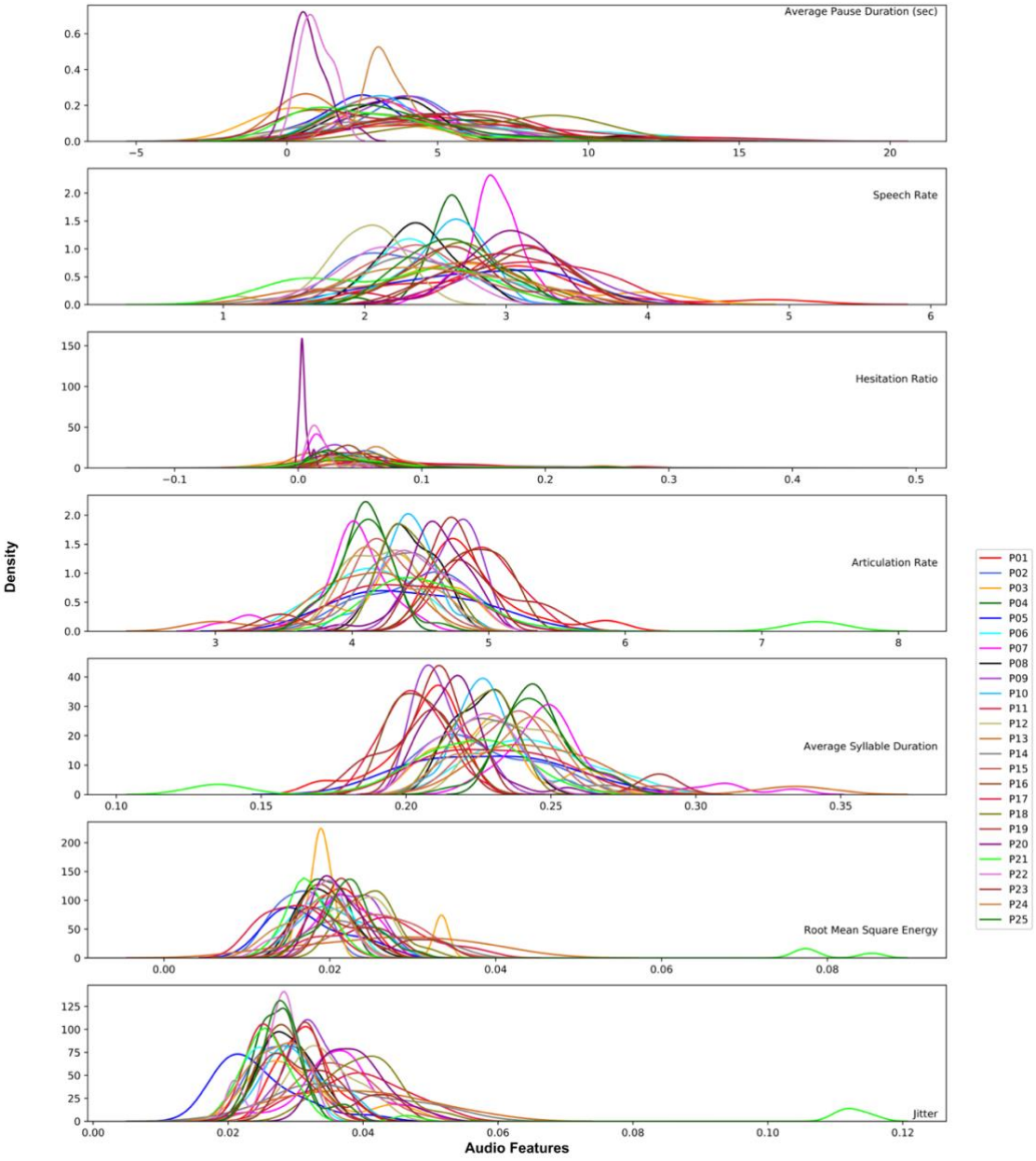
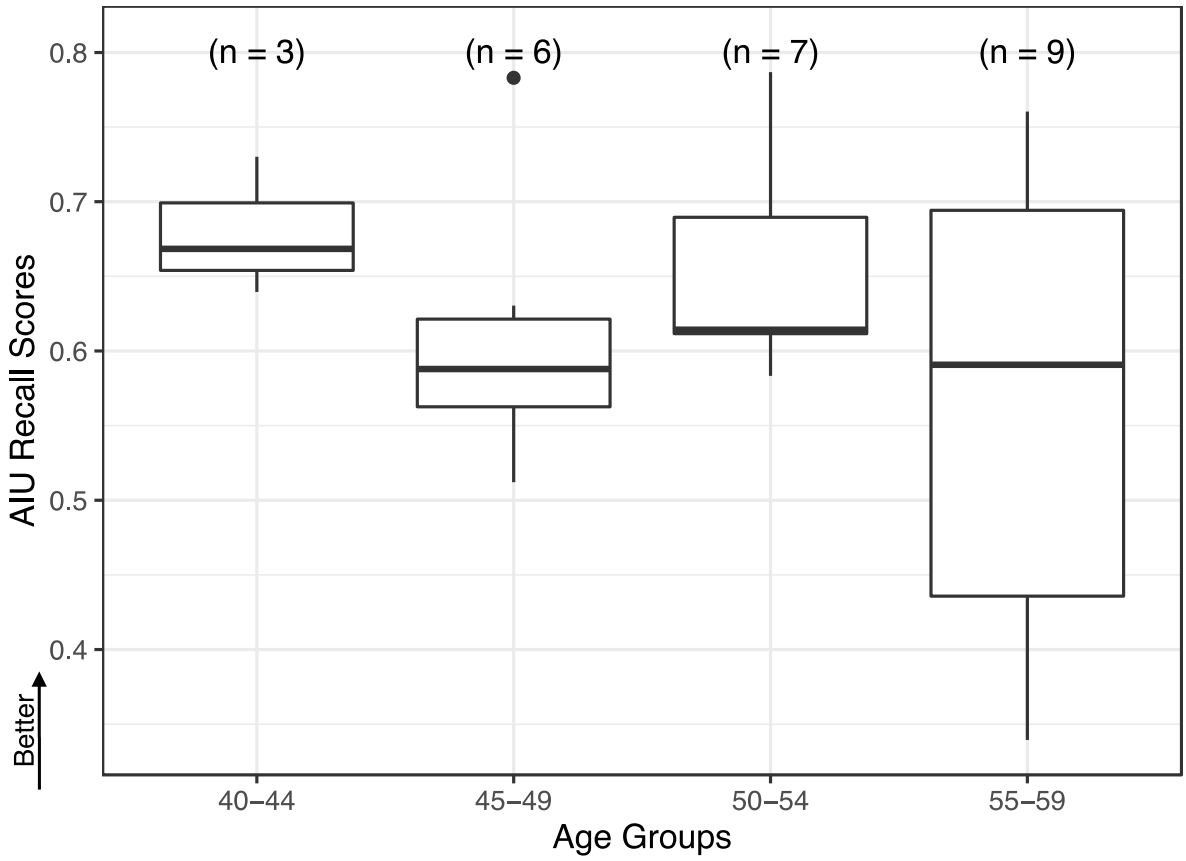
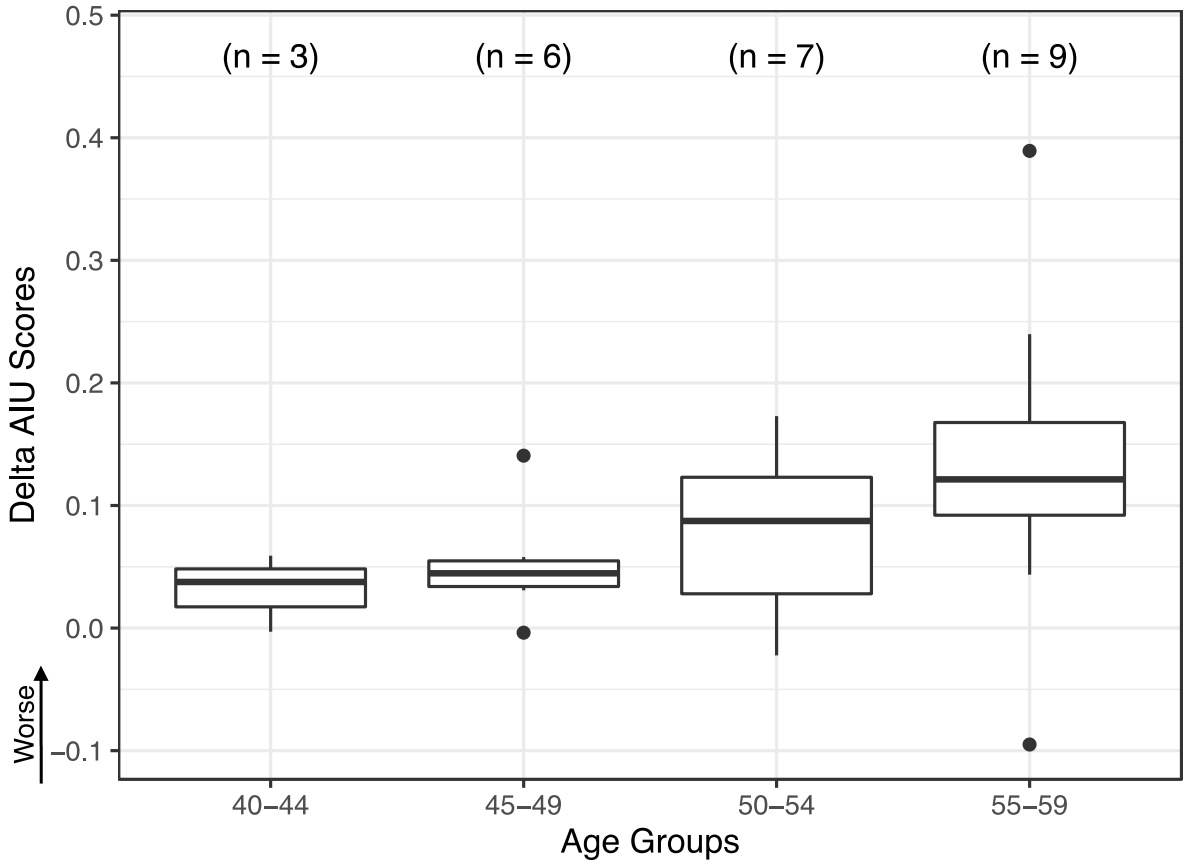


Figure A.5 Distribution of average AIU recall scores in each age group.



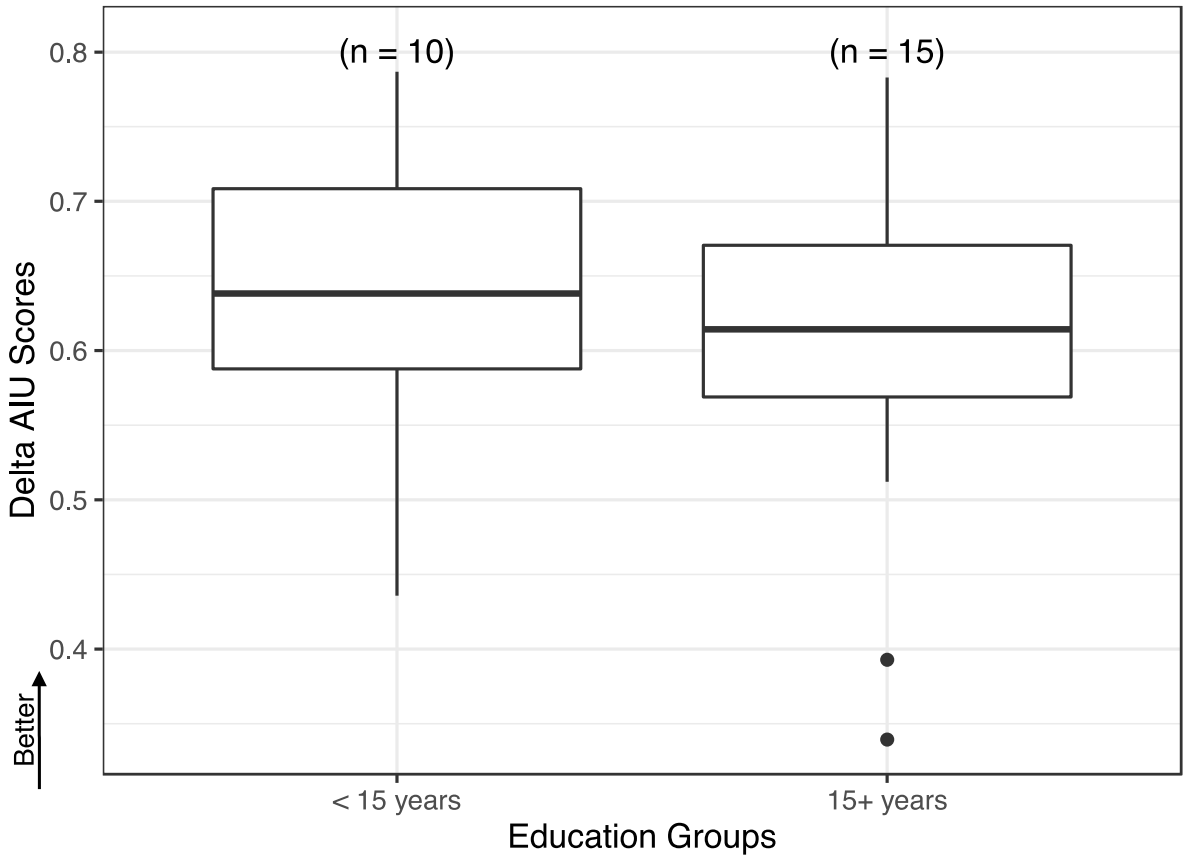
The edge of the box represents the first quartile (Q1) and third quartile (Q3), the box represents the interquartile range (IQR), the horizontal line inside the box represents the median value, whiskers represent minimum (Q1 - 1.5 x IQR), and maximum (Q3 + 1.5 x IQR) and dots represent the values outside the range of minimum and maximum. AIU - Automated Information Units

Figure A.6 Distribution of average delta AIU scores in each age group.



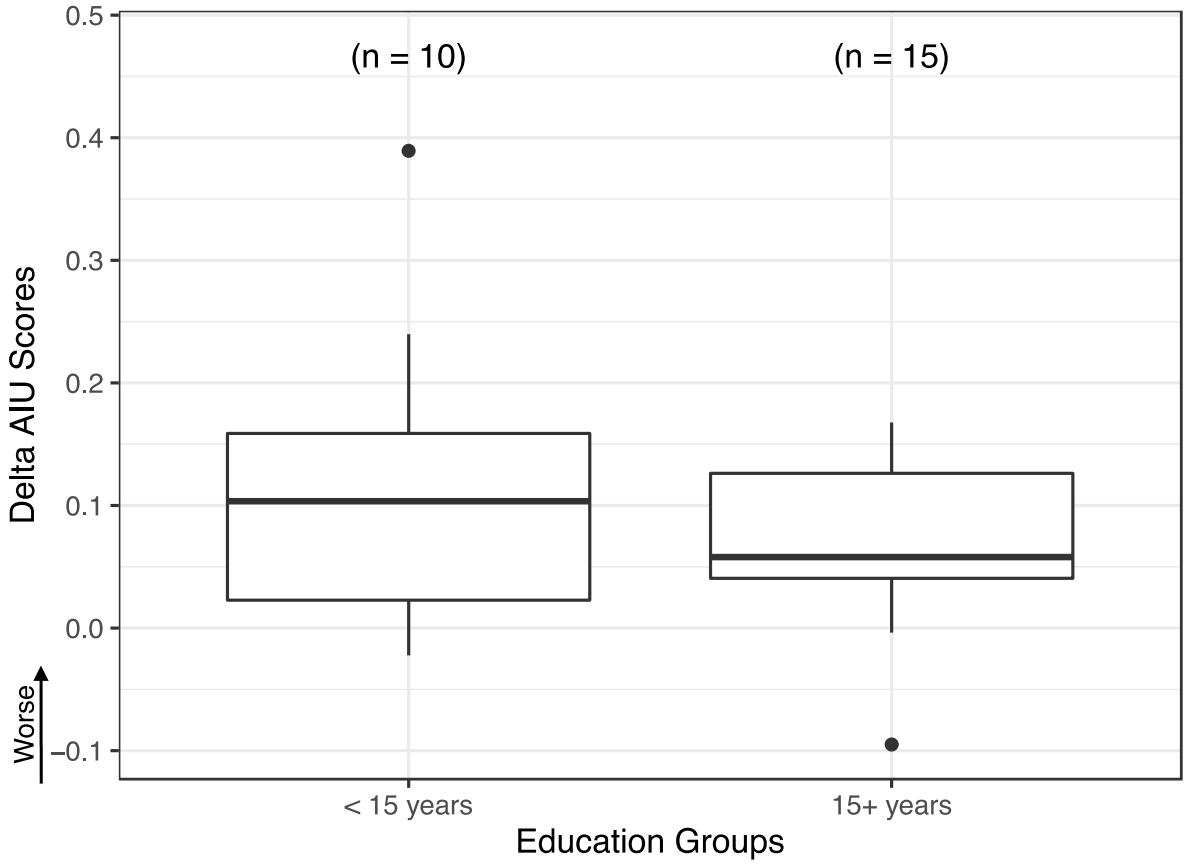
The edge of the box represents the first quartile (Q1) and third quartile (Q3), the box represents the interquartile range (IQR), the horizontal line inside the box represents the median value, whiskers represent minimum (Q1 - 1.5 x IQR), and maximum (Q3 + 1.5 x IQR) and dots represent the values outside the range of minimum and maximum. AIU - Automated Information Units

Figure A.7 Distribution of average AIU recall scores in each education group.



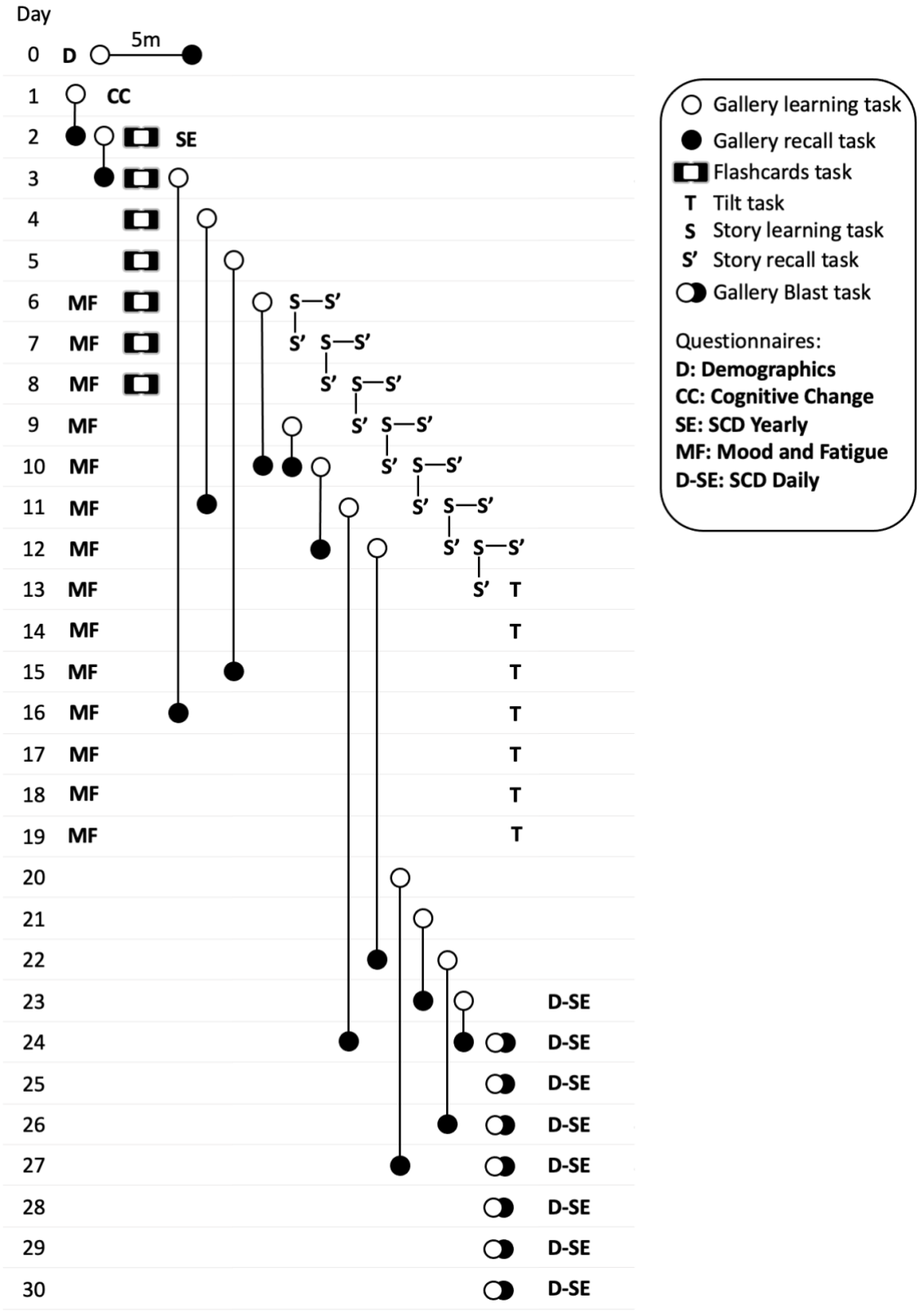
The edge of the box represents the first quartile (Q1) and third quartile (Q3), the box represents the interquartile range (IQR), the horizontal line inside the box represents the median value, whiskers represent minimum (Q1 - 1.5 x IQR), and maximum (Q3 + 1.5 x IQR) and dots represent the values outside the range of minimum and maximum. AIU - Automated Information Units

Figure A.8 Distribution of average delta AIU scores in each education group.



The edge of the box represents the first quartile (Q1) and third quartile (Q3), the box represents the interquartile range (IQR), the horizontal line inside the box represents the median value, whiskers represent minimum (Q1 - 1.5 x IQR), and maximum (Q3 + 1.5 x IQR) and dots represent the values outside the range of minimum and maximum. AIU - Automated Information Units

Figure A.9 The task schedule of the GameChanger study showing various cognitive tasks and self-report questionnaires as they appear in the 30 days schedule.



SCD - Subjective Cognitive Decline

Figure A.10 Distribution of the number of participants fully completing (learning, immediate recall, and delayed recall) different number of stories in GameChanger.

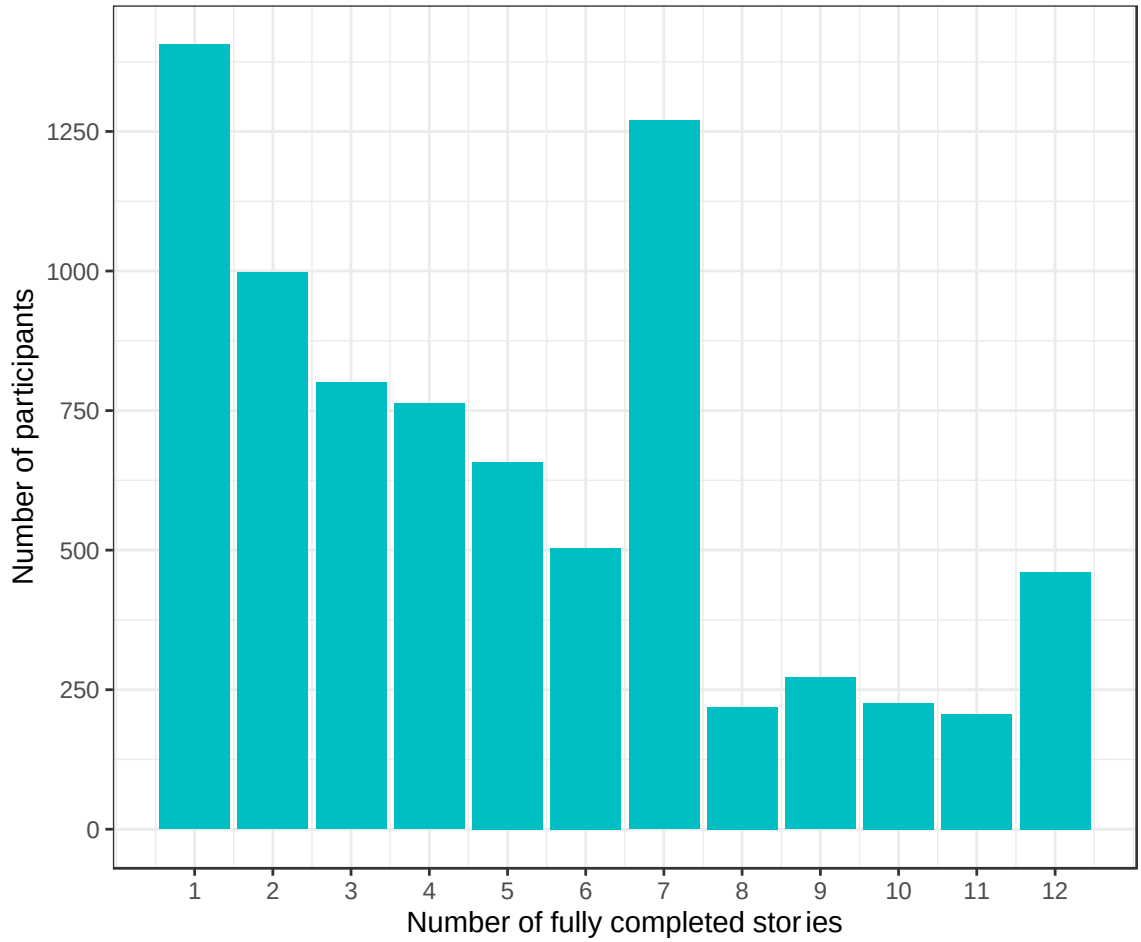


Figure A.11 Correlation of the pipeline features extracted from the Story Time recall task.

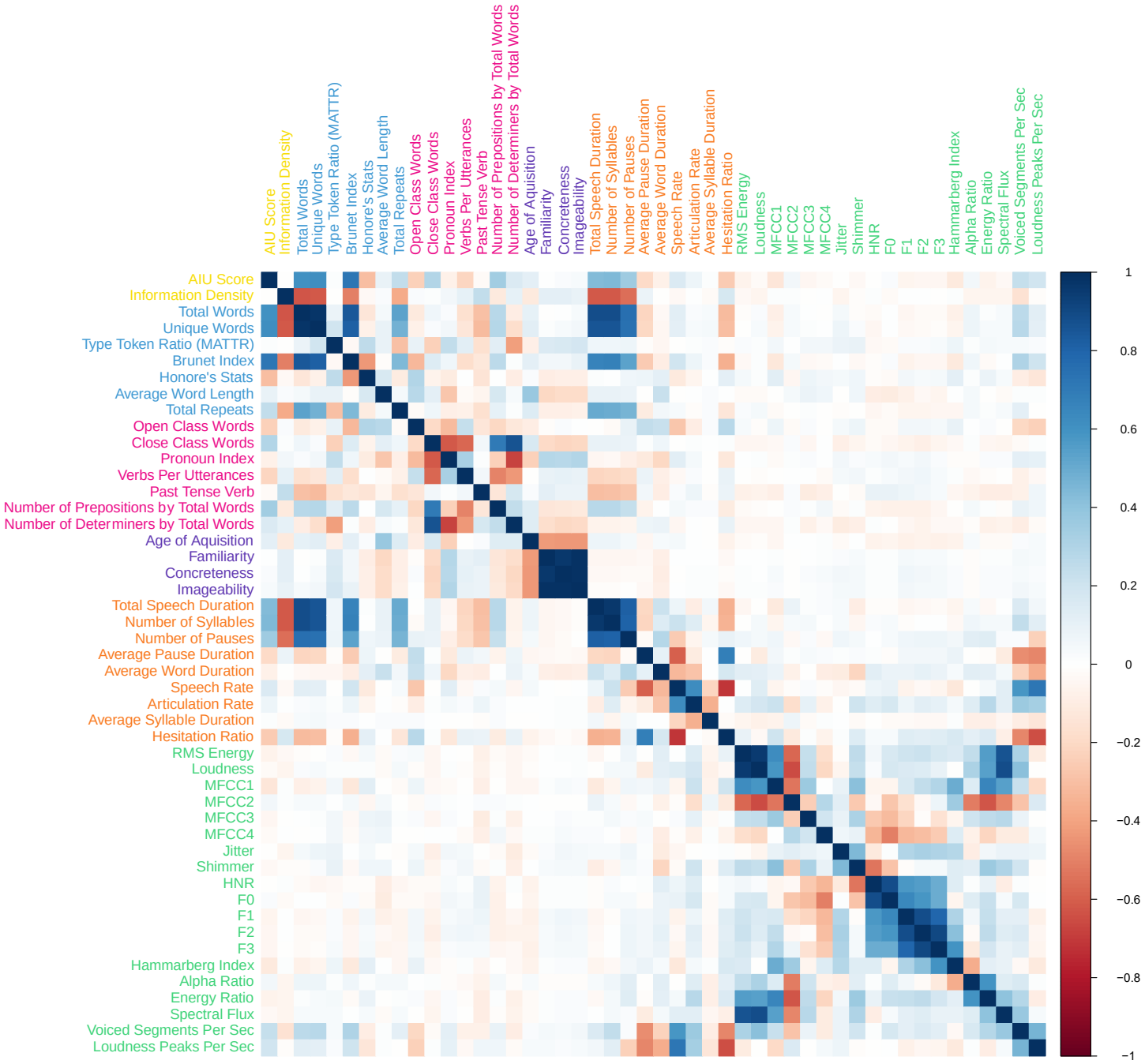


Figure A.12 Density plot of various psycholinguistic features in the GameChanger story responses at recalls.

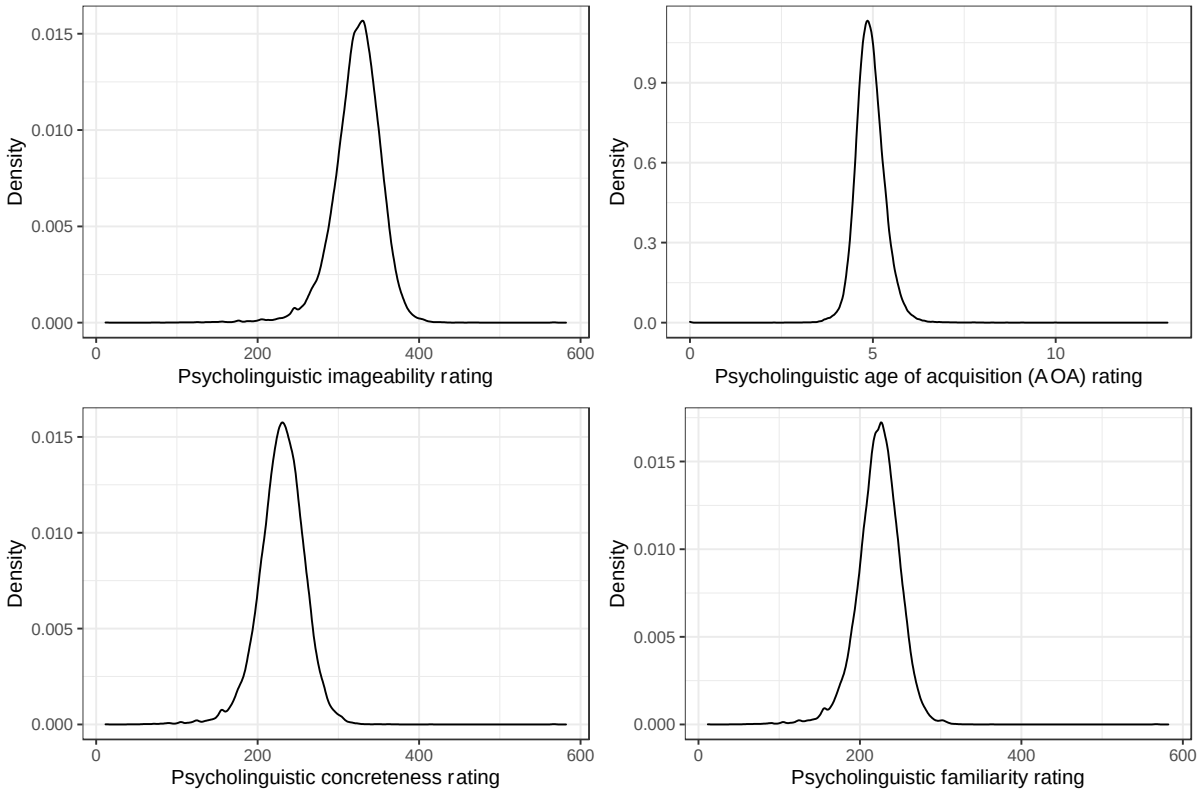
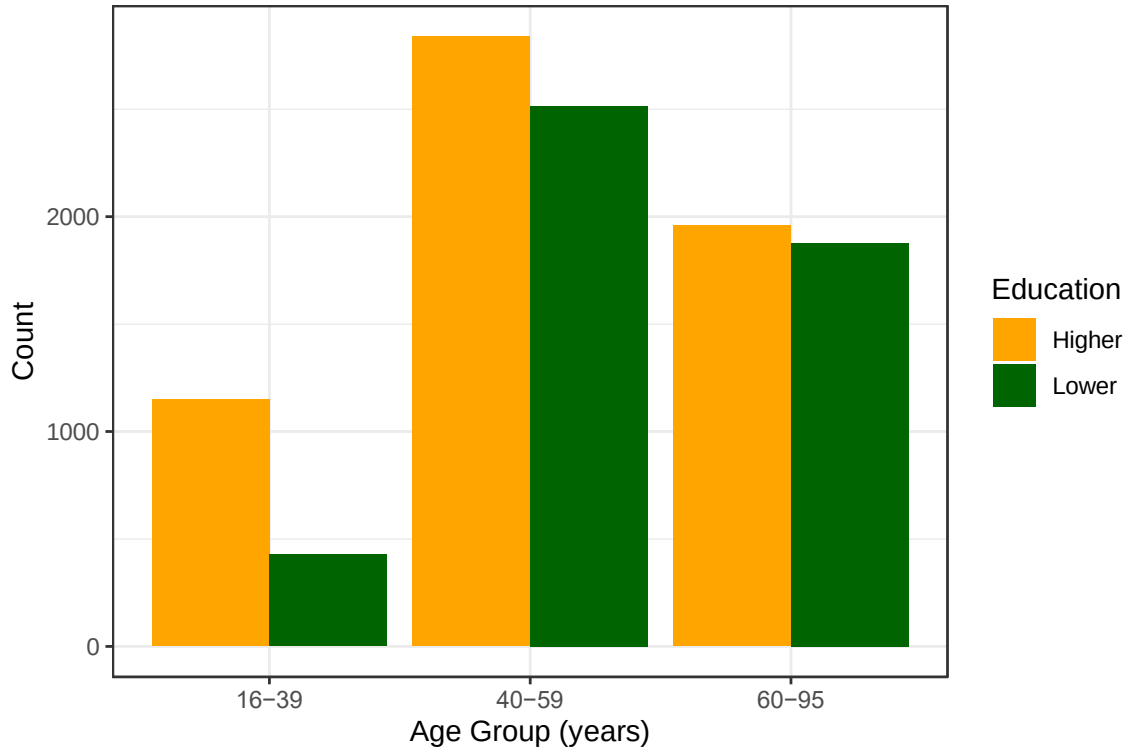


Figure A.13 Bar plot showing the number of participants in various age groups stratified by education in the GameChanger study.



Participants with at least an undergraduate degree were coded higher and remaining as lower.

Appendix B. Supplementary Data File

File B.1

Filename: List of expected phrases and information units in Story Time narration.xlsx

Description: The accompanying Excel spreadsheets provide the text data that was prepared for ASR model optimisation. These text data were used as input to optimise ASR.

File B.2

Filename: Text for building custom language model.xlsx

Description: The accompanying Excel spreadsheets provide the text data that was prepared for ASR model optimisation. These text data were used as input to optimise ASR.

File B.3

Filename: Counterfitted_Story_Time_synonym_pairs.txt

Description: The accompanying text file provides lexical variants that were used to retrofit the generic British National Corpus word2vec model.

Appendix C. RADAR-AD cohort

RADAR-AD is an ongoing cross-sectional observational study with participants including healthy control, pre-clinical AD, prodromal AD, and mild-to-moderate AD, aimed to access the remote monitoring technologies that can provide accurate functional measurement across pre-clinical to MCI stages [86]. This study provides an opportunity to evaluate speech-based features on the Story Time task in the earlier disease stages.

There were fourteen clinical sites across thirteen European countries where the participants were recruited. During enrolment, participants underwent AD biomarker evidence confirmation of Aβ burden with either PET or CSF examination. RADAR-AD participants used remote monitoring technologies, including wrist-worn wearables (Fitbit and Axivity AX3) and smartphone applications (Altoida, Mezurio, and pRMT) and provided the data during the 8-week study period. Detail on the participant's inclusion and exclusion criteria, devices used in the study, and baseline assessments, are available in Muurling et al., 2021 [86].

As outlined above, RADAR-AD contains participants from various European countries who complete Story Time tasks in their native language. Multilingual responses obtained in the Story Time task limited the extraction of language-based features, including the AIU scoring (see Chapter 5). This is because speech-to-text accuracy in languages other than English has yet to reach the level obtained in English. Current NLP research focuses on developing multilingual ASR with improved accuracy by utilising deep learning models [357], [358]. Also,

word embedding models are publicly available only for a few European languages, and to train the Word2Vec model from scratch, a large corpus of language text is required. Substantially more work, including domain adaptation and validation, is necessary to improve the accuracy of automated transcription outputs and Word2Vec models, as discussed in the development of the feature extraction pipeline chapter (Chapter 5).

Baseline characteristics of RADAR-AD participants

In this subset of the RADAR-AD cohort, there were 175 participants. Of these, 121 participants provided speech recordings for at least one recall sub-tasks (immediate and delayed) of six stories within the Story Time task. The overall mean (SD) age of the participants was 67.9 (7.2) years, and the mean (SD) education was 14.8 (3.7) years (**Table C.1**). The clinical groups did not differ in age, sex or years of education.

Table C.1 Demographic characteristic of RADAR-AD participants per clinical group providing Story Time data.

	Healthy Controls	**Pre-clinical AD**	**Prodromal AD**	**Mild-to-moderate AD**
	(n=53)	(n=20)	(n=33)	(n=15)
Age	67.17 (7.84)	68.95 (4.67)	68.88 (7.20)	65.47 (7.74)
Female, n (%)	28 (52.8)	13 (65.0)	14 (42.4)	7 (46.7)
Years of Education	14.66 (3.42)	15.25 (3.12)	14.88 (4.66)	14.07 (3.24)
MMSE Score	29.28 (0.86)	29.20 (1.15)	26.76 (2.32)	21.47 (3.58)
Animal Fluency	21.56 (6.26)	23.75 (6.02)	16.94 (4.71)	11.33 (4.56)
Clinical sites, n (%)				
Amsterdam	14 (26.4)	11 (55.0)	8 (24.2)	4 (26.7)
London	5 (9.4)	0 (0.0)	1 (3.0)	0 (0.0)
Oxford	6 (11.3)	1 (5.0)	0 (0.0)	0 (0.0)
Stockholm	0 (0.0)	3 (15.0)	6 (18.2)	1 (6.7)
Thessaloniki	4 (7.5)	0 (0.0)	1 (3.0)	0 (0.0)
Ljubljana	2 (3.8)	0 (0.0)	0 (0.0)	0 (0.0)
Lisbon	6 (11.3)	0 (0.0)	4 (12.1)	3 (20.0)
Brescia	6 (11.3)	0 (0.0)	2 (6.1)	0 (0.0)
Mannheim	0 (0.0)	0 (0.0)	4 (12.1)	2 (13.3)
Barcelona	5 (9.4)	1 (5.0)	5 (15.2)	2 (13.3)
Stavanger	5 (9.4)	4 (20.0)	2 (6.1)	3 (20.0)

Numbers show mean (SD) unless specified otherwise; MMSE - Mini-Mental State Examination.

References

[1] "Dementia," *World Health Organisation*, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia. [Accessed: 15-Apr-2022].

[2] R. A. Sperling *et al.*, "Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's and Dementia*, vol. 7, no. 3, pp. 280–292, 2011.

[3] C. Fennema-Notestine, L. K. McEvoy, D. J. Hagler Jr, M. W. Jacobson, A. M. Dale, and A. D. N. Initiative, "Structural neuroimaging in the detection and prognosis of pre-clinical and early AD," *Behavioural neurology*, vol. 21, no. 1–2, pp. 3–12, 2009.

[4] B. Dubois *et al.*, "Revising the definition of Alzheimer's disease: a new lexicon," *The Lancet Neurology*, vol. 9, no. 11, pp. 1118–1127, 2010.

[5] D. Maquet *et al.*, "Gait analysis in elderly adult patients with mild cognitive impairment and patients with mild Alzheimer's disease: simple versus dual task: a preliminary report," *Clinical physiology and functional imaging*, vol. 30, no. 1, pp. 51–56, 2010.

[6] G. Livingston *et al.*, "Dementia prevention, intervention, and care," *The Lancet*, vol. 390, no. 10113, pp. 2673–2734, 2017.

[7] D.-D. Li, Y.-H. Zhang, W. Zhang, and P. Zhao, "Meta-analysis of randomized

controlled trials on the efficacy and safety of donepezil, galantamine, rivastigmine, and memantine for the treatment of Alzheimer's disease," *Frontiers in neuroscience*, vol. 13, p. 472, 2019.

[8] A. Abbott, "Treating Alzheimer's Before It Takes Hold," *Nature*, vol. 603, pp. 216–219, 2022.

[9] C. H. Van Dyck *et al.*, "Lecanemab in early Alzheimer's disease," *New England Journal of Medicine*, vol. 388, no. 1, pp. 9–21, 2023.

[10] M. A. Mintun *et al.*, "Donanemab in early Alzheimer's disease," *New England Journal of Medicine*, vol. 384, no. 18, pp. 1691–1704, 2021.

[11] W. H. Organization, *First WHO ministerial conference on global action against dementia: Meeting report*. Geneva, Switzerland: WHO Library Cataloguing-in-Publication Data, 2015.

[12] P. Cuijpers, "Depressive disorders in caregivers of dementia patients: A systematic review," *Aging and Mental Health*, vol. 9, no. 4, pp. 325–330, 2005.

[13] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "'Mini-mental state': a practical method for grading the cognitive state of patients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[14] M. Lozano-Gallego, J. Vilalta-Franch, J. Llinàs-Reglà, and S. López-Pousa, "The Cambridge Cognitive Examination as a tool for detection of dementia," *Revista de*

neurologia, vol. 28, no. 4, pp. 348–352, 1999.

[15] E. Mioshi, K. Dawson, J. Mitchell, R. Arnold, and J. R. Hodges, "The Addenbrooke's Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening," *International Journal of Geriatric Psychiatry*, vol. 21, pp. 1078–1085, 2007.

[16] Z. S. Nasreddine *et al.*, "The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[17] B. Reisberg, S. H. Ferris, M. J. de Leon, and T. Crook, "The Global Deterioration Scale for assessment of primary degenerative dementia.," *The American Journal of Psychiatry*, vol. 139, no. 9, pp. 1136–1139, 1982.

[18] C. P. Hughes, L. Berg, W. L. Danziger, L. A. Coben, and R. L. Martin, "A new clinical scale for the staging of dementia," *British Journal of Psychiatry*, vol. 140, no. 6, pp. 566–572, 1982.

[19] W. G. Rosen, R. C. Mohs, and K. L. Davis, "A new rating scale for Alzheimer's disease.," *The American Journal of Psychiatry*, vol. 141, no. 11, pp. 1356–1364, 1984.

[20] J. Mcconathy and Y. I. Sheline, "Imaging Biomarkers Associated with Cognitive Decline: A Review," *Biological Psychiatry*, vol. 77, no. 8, pp. 685–692, 2015.

[21] F. Cuetos, J. C. Arango-Lasprilla, C. Uribe, C. Valencia, and F. Lopera, "Linguistic changes in verbal expression: A preclinical marker of Alzheimer's disease," *Journal of*

the International Neuropsychological Society, vol. 13, no. 3, pp. 433–439, 2007.

[22] M. Crous-Bou, C. Minguillón, N. Gramunt, and J. L. Molinuevo, "Alzheimer's disease prevention: From risk factors to early intervention," *Alzheimer's Research and Therapy*, vol. 9, no. 1, pp. 1–9, 2017.

[23] P. S. Aisen, G. A. Jimenez-Maggiora, M. S. Rafii, S. Walter, and R. Raman, "Early-stage Alzheimer disease: getting trial-ready," *Nature Reviews Neurology*, vol. 18, no. 7, pp. 389–399, 2022.

[24] B. Dubois, A. Padovani, P. Scheltens, A. Rossi, and G. Dell'Agnello, "Timely diagnosis for Alzheimer's disease: a literature review on benefits and challenges," *Journal of Alzheimer's disease*, vol. 49, no. 3, pp. 617–631, 2016.

[25] F. Öhman, J. Hassenstab, D. Berron, M. Schöll, and K. V. Papp, "Current advances in digital cognitive assessment for preclinical Alzheimer's disease," *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, vol. 13, no. 1, pp. 1–19, 2021.

[26] D. M. Rentz, M. A. Parra Rodriguez, R. Amariglio, Y. Stern, R. Sperling, and S. Ferris, "Promising developments in neuropsychological approaches for the detection of preclinical Alzheimer's disease: A selective review," *Alzheimer's Research and Therapy*, vol. 5, no. 6, 2013.

[27] R. Hill, L. R. Betts, and S. E. Gardner, "Older adults experiences and perceptions of

digital technology: (Dis)empowerment, wellbeing, and inclusion," *Computers in Human Behavior*, vol. 48, pp. 415–423, 2015.

[28] R. Milne and C. Brayne, "We need to think about data governance for dementia research in a digital era," *Alzheimer's Research and Therapy*, vol. 12, no. 1, pp. 10–12, 2020.

[29] R. A. Armstrong, "Risk factors for alzheimer disease," *Brain and Nerve*, vol. 57, no. 2, pp. 87–105, 2019.

[30] O. M. Bubu *et al.*, "Sleep, cognitive impairment, and Alzheimer's disease: a systematic review and meta-analysis," *Sleep*, vol. 40, no. 1, p. zsw032, 2017.

[31] K. D. Mueller, B. Hermann, J. Mecollari, and L. S. Turkstra, "Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks," *Journal of Clinical and Experimental Neuropsychology*, vol. 40, no. 9, pp. 917–939, 2018.

[32] V. B. Fleming, "Early detection of cognitive-linguistic change associated with mild cognitive impairment," *Communication Disorders Quarterly*, vol. 35, no. 3, pp. 146–157, 2014.

[33] H. Schriefers and G. Vigliocco, *Psychology of Speech Production*, 2nd ed., vol. 23. Elsevier, 2015.

[34] S. H. Vos *et al.*, "Syntactic parsing and working memory : The effects of syntactic complexity, reading span, and concurrent load," *Language and Cognitive Processes*,

vol. 16, no. 1, pp. 65–103, 2001.

[35] R. J. Hartsuiker and P. N. Barkhuysen, “Language production and working memory: The case of subject-verb agreement,” *Language and Cognitive Processes*, vol. 21, no. 1–3, pp. 181–204, 2006.

[36] M. F. Elias, A. Beiser, P. A. Wolf, R. Au, R. F. White, and R. B. D’Agostino, “The preclinical phase of Alzheimer disease: A 22-year prospective study of the Framingham cohort,” *Archives of Neurology*, vol. 57, no. 6, pp. 808–813, 2000.

[37] V. Berisha, S. Wang, A. LaCross, and J. Liss, “Tracking Discourse Complexity Preceding Alzheimer’s Disease Diagnosis: A Case Study Comparing the Press Conferences of Presidents Ronald Reagan and George Herbert Walker Bush,” *Journal of Alzheimer’s Disease*, vol. 45, no. 3, pp. 959–963, 2015.

[38] M. Bilgel *et al.*, “Trajectories of Alzheimer disease-related cognitive measures in a longitudinal sample,” *Alzheimer’s and Dementia*, vol. 10, no. 6, pp. 735–742, 2014.

[39] V. Taler and N. A. Phillips, “Language performance in Alzheimer’s disease and mild cognitive impairment: A comparative review,” *Journal of Clinical and Experimental Neuropsychology*, vol. 30, no. 5, pp. 501–556, 2008.

[40] D. A. Snowdon, L. H. Greiner, and W. R. Markesbery, “Linguistic ability in early life and the neuropathology of Alzheimer’s disease and cerebrovascular disease: Findings from the Nun Study,” *Annals of the New York Academy of Sciences*, vol. 903, no. 1, pp.

34–38, 2000.

[41] C. Laske *et al.*, “Innovative diagnostic tools for early detection of Alzheimer’s disease,” *Alzheimer’s and Dementia*, vol. 11, no. 5, pp. 561–578, 2015.

[42] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, “Linguistic features identify Alzheimer’s disease in narrative speech,” *Journal of Alzheimer’s Disease*, vol. 49, no. 2, pp. 407–422, 2015.

[43] M. Yancheva and F. Rudzicz, “Vector-space topic models for detecting Alzheimer’s disease,” in *Proceedings of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 2337–2346.

[44] S. Ahmed, A. M. F. Haigh, C. A. de Jager, and P. Garrard, “Connected speech as a marker of disease progression in autopsy-proven Alzheimer’s disease.,” *Brain*, vol. 136, no. 12, pp. 3727–3737, 2013.

[45] R. Lunsford and P. A. Heeman, “Using linguistic indicators of difficulty to identify mild cognitive impairment,” in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2015, pp. 658–662.

[46] M. Nicholas, L. K. Obler, M. L. Albert, and N. Helm-Estabrooks, “Empty speech in Alzheimer’s disease and fluent aphasia,” *Journal of Speech and Hearing Research*, vol. 28, no. 3, pp. 405–410, 1985.

[47] B. Croisile *et al.*, “Comparative study of oral and written picture description in patients

with Alzheimer's disease," *Brain and Language*, vol. 53, no. 1, pp. 1–19, 1996.

[48] E. Aramaki, S. Shikata, M. Miyabe, and A. Kinoshita, "Vocabulary size in speech May Be an early indicator of cognitive impairment," *PLoS ONE*, vol. 11, no. 5, pp. 1–13, 2016.

[49] K. D. Mueller, R. L. Koscik, B. P. Hermann, S. C. Johnson, and L. S. Turkstra, "Declines in connected language are associated with very early mild cognitive impairment: Results from the Wisconsin Registry for Alzheimer's Prevention," *Frontiers in Aging Neuroscience*, vol. 9, p. 437, 2018.

[50] A. Almor, D. Kempler, M. C. MacDonald, E. S. Andersen, and L. K. Tyler, "Why do Alzheimer patients have difficulty with pronouns? Working memory, semantics, and reference in comprehension and production in Alzheimer's disease," *Brain and Language*, vol. 67, no. 3, pp. 202–227, 1999.

[51] K. E. Forbes, A. Venneri, and M. F. Shanks, "Distinct patterns of spontaneous speech deterioration: An early predictor of Alzheimer's Disease," *Brain and Cognition*, vol. 48, no. 2–3, pp. 356–361, 2002.

[52] S. Ahmed, C. A. de Jager, A. M. Haigh, and P. Garrard, "Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed alzheimer's disease," *Neuropsychology*, vol. 27, no. 1, pp. 79–85, 2013.

[53] X. Le, I. Lancashire, G. Hirst, and R. Jokel, "Longitudinal detection of dementia through

lexical and syntactic changes in writing: A case study of three British novelists," *Literary and Linguistic Computing*, vol. 26, no. 4, pp. 435–461, 2011.

[54] F. Gayraud, H. R. Lee, and M. Barkat-Defradas, "Syntactic and lexical context of pauses and hesitations in the discourse of Alzheimer patients and healthy elderly subjects," *Clinical Linguistics and Phonetics*, vol. 25, no. 3, pp. 198–209, 2011.

[55] J. A. Thomas *et al.*, "Assessing the Utility of Language and Voice Biomarkers to Predict Cognitive Impairment in the Framingham Heart Study Cognitive Aging Cohort Data," *Journal of Alzheimer's Disease*, vol. 76, no. 3, pp. 905–922, 2020.

[56] S. Kemper, C. Anagnopoulos, K. Lyons, and W. Heberlein, "Speech accommodations to dementia," *Journals of Gerontology*, vol. 49, no. 5, pp. 223–229, 1994.

[57] G. Blanken, J. Dittmann, J.-C. Haas, and C.-W. Wallesch, "Spontaneous speech in senile dementia and aphasia : Implications for a neurolinguistic model of language production," *Cognition*, vol. 27, pp. 247–274, 1987.

[58] J. Appell, A. Kertesz, and M. Fisman, "A study of language functioning in Alzheimer patients," *Brain and Language*, vol. 17, no. 1, pp. 73–91, 1982.

[59] K. Lyons, S. Kemper, E. LaBarge, F. R. Ferraro, D. Balota, and M. Storandt, "Oral Language and Alzheimer's Disease: A Reduction in Syntactic Complexity," *Aging and Cognition*, vol. 1, no. 4, pp. 271–281, 1994.

[60] K. E. Forbes-McKay and A. Venneri, "Detecting subtle spontaneous language decline

in early Alzheimer's disease with a picture description task," *Neurological Sciences*, vol. 26, no. 4, pp. 243–254, 2005.

[61] S. Ash, P. Moore, L. Vesely, and M. Grossman, "The decline of narrative discourse in Alzheimer's disease," *Brain and Language*, vol. 103, no. 1–2, pp. 181–182, 2007.

[62] E. Can and G. Kuruoğlu, "Language Changes in Late-Onset Alzheimer's Disease," *Psycholinguistics*, vol. 25, no. 2, pp. 50–68, 2019.

[63] T. Bschor, K. P. Kühl, and F. M. Reischies, "Spontaneous speech of patients with dementia of the Alzheimer type and mild cognitive impairment," *International Psychogeriatrics*, vol. 13, no. 3, pp. 289–298, 2001.

[64] H. Goodglass and E. Kaplan, *Boston diagnostic aphasia examination booklet*. Lea & Febiger, 1983.

[65] C. K. Tomoeda, K. A. Bayles, M. W. Trosset, T. Azuma, and A. McGeagh, "Cross-sectional analysis of Alzheimer disease effects on oral discourse in a picture description task," *Alzheimer Disease and Associated Disorders*, vol. 10, no. 4. pp. 204–215, 1996.

[66] G. Kavé and M. Goral, "Word retrieval in picture descriptions produced by individuals with Alzheimer's disease," *Journal of Clinical and Experimental Neuropsychology*, vol. 38, no. 9, pp. 958–966, 2016.

[67] A. Pistono, J. Pariente, C. Bézy, B. Lemesle, J. Le Men, and M. Jucla, "What happens when nothing happens? An investigation of pauses as a compensatory mechanism in

early Alzheimer's disease," *Neuropsychologia*, vol. 124, pp. 133–143, 2019.

[68] D. Wechsler, *WAIS-III: Administration and scoring manual: Wechsler adult intelligence scale*. Psychological Corporation, 1997.

[69] D. Wechsler, "Wechsler memory scale," *Psychological Corporation*, 1945.

[70] K. A. Bayles, C. K. Tomoeda, and R. I. Dharmaperwira-Prins, "ABCD: The Arizona Battery for Communication Disorders of Dementia," *Canyonlands Publishing*, 1993.

[71] M. Mayer, *Frog, where are you?* New York: NY: Dial Press, 1969.

[72] V. Fleming and J. Harris, "Complex discourse production in mild cognitive impairment: Detecting subtle changes," *Aphasiology*, vol. 22, no. 7–8, pp. 729–740, 2008.

[73] S. De La Fuente Garcia, C. W. Ritchie, and S. Luz, "Protocol for a conversation-based analysis study: PREVENT-ED investigates dialogue features that may help predict dementia onset in later life," *BMJ Open*, vol. 9, no. 3, pp. 1–10, 2019.

[74] M. S. Cannizzaro, N. Reilly, J. C. Mundt, and P. J. Snyder, "Remote capture of human voice acoustical data by telephone: A methods study," *Clinical Linguist Phonology*, vol. 19, no. 8, pp. 649–658, 2005.

[75] G. Fairbanks, "Pitch," in *Voice and articulation drillbook*, 2nd ed., New York: Harper & Row, 1960, pp. 124–134.

[76] C. De Looze, F. Kelly, L. Crosby, A. Vourdanou, B. Lawlor, and R. Reilly, "Prosodic

Chunking in Out-Loud Reading: a Window Into Cognitive Function in Mild-To-Moderate Alzheimer'S Disease," *Alzheimer's & Dementia*, vol. 13, no. 7, pp. P699–P700, 2017.

[77] L. E. Nicholas and R. H. Brookshire, "A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia," *Journal of Speech and Hearing Research*, vol. 36, no. 2, pp. 338–350, 1993.

[78] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, 2000.

[79] V. Boschi, E. Catricalà, M. Consonni, C. Chesi, A. Moro, and S. F. Cappa, "Connected speech in neurodegenerative language disorders: A review," *Frontiers in Psychology*, vol. 8, p. 269, 2017.

[80] E. Vuorinen, M. Laine, and J. Rinne, "Common pattern of language impairment in vascular dementia and in Alzheimer disease," *Alzheimer Disease and Associated Disorders*, vol. 14, no. 2, pp. 81–86, 2000.

[81] S. Ash, P. Moore, S. Antani, G. McCawley, M. Work, and M. Grossman, "Trying to tell a tale: Discourse impairments in progressive aphasia and frontotemporal dementia," *Neurology*, vol. 66, no. 9, pp. 1405–1413, 2006.

[82] S. Ash *et al.*, “Impairments of speech fluency in Lewy body spectrum disorder,” *Brain and Language*, vol. 120, no. 3, pp. 290–302, 2012.

[83] C. W. Ritchie and K. Ritchie, “The PREVENT study: A prospective cohort study to identify mid-life biomarkers of late-onset Alzheimer’s disease,” *BMJ Open*, vol. 2, no. 6, pp. 1–6, 2012.

[84] “GameChanger,” *GameChanger*. [Online]. Available: https://joingamechanger.org/. [Accessed: 15-Nov-2022].

[85] I. Koychev *et al.*, “Deep and Frequent Phenotyping study protocol: An observational study in prodromal Alzheimer’s disease,” *BMJ Open*, vol. 9, no. 3, pp. 1–16, 2019.

[86] M. Muurling *et al.*, “Remote monitoring technologies in Alzheimer ’ s disease : design of the RADAR- AD study,” *Alzheimer’s Research and Therapy*, vol. 13, no. 1, pp. 1–13, 2021.

[87] A.-L. Frey *et al.*, “Harnessing the potential of digital technologies for the early detection of neurodegenerative diseases,” 2019.

[88] C. Lancaster, I. Koychev, J. Blane, A. Chinner, L. Wolters, and C. Hinds, “Evaluating the feasibility of frequent cognitive assessment using the Mezurio smartphone app: Observational and interview study in adults with elevated dementia risk,” *JMIR mHealth and uHealth*, vol. 8, no. 4, 2020.

[89] P. Machado, J. Romero, M. Nadal, A. Santos, J. Correia, and A. Carballal,

"Computerized measures of visual complexity," *Acta Psychologica*, vol. 160, pp. 43–57, 2015.

[90] J. Simola, J. Torniainen, M. Moisala, M. Kivikangas, and C. M. Krause, "Eye movement related brain responses to emotional scenes during free viewing," *Frontiers in Systems Neuroscience*, vol. 7, p. 41, 2013.

[91] S. C. J. Verfaillie *et al.*, "High amyloid burden is associated with fewer specific words during spontaneous speech in individuals with subjective cognitive decline," *Neuropsychologia*, vol. 131, pp. 184–192, 2019.

[92] W. Simpson, L. Kaufman, M. Detke, C. Lynch, A. Butler, and S. Dominy, "Utility of Speech-Based Digital Biomarkers for Evaluating Disease Progression in Clinical Trials of Alzheimer'S Disease," *Alzheimer's & Dementia*, vol. 15, no. 7, p. P1524, 2019.

[93] C. Themistocleous, M. Eckerström, and D. Kokkinakis, "Voice quality and speech fluency distinguish individuals with Mild Cognitive Impairment from Healthy Controls," *PLoS ONE*, vol. 15, pp. 1–13, 2020.

[94] B. Zellner, "Pauses and the temporal structure of speech," in *Fundamentals of speech synthesis and speech recognition*, E. Keller, Ed. John Willey and Sons Ltd., 1994, pp. 41–62.

[95] O. Ivanova, J. J. G. Meilán, F. Martínez-Sánchez, I. Martínez-Nicolás, T. E. Llorente, and N. C. González, "Discriminating speech traits of Alzheimer's disease assessed

through a corpus of reading task for Spanish language," *Computer Speech and Language*, vol. 73, p. 101341, 2022.

[96] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, and J. A. S. Martin, "Acoustic Markers Associated with Impairment in Language Processing in Alzheimer's Disease," *The Spanish Journal of Psychology*, vol. 15, no. 2, pp. 487–494, 2012.

[97] S. Nasreen, M. Rohanian, J. Hough, and M. Purver, "Alzheimer's Dementia Recognition From Spontaneous Speech Using Disfluency and Interactional Features," *Frontiers in Computer Science*, vol. 3, no. June, pp. 1–15, 2021.

[98] S. Al-Hameed, M. Benaissa, H. Christensen, B. Mirheidari, D. Blackburn, and M. Reuber, "A new diagnostic approach for the identification of patients with neurodegenerative cognitive complaints," *PLoS ONE*, vol. 14, no. 5, pp. 1–18, 2019.

[99] A. König *et al.*, "Detecting apathy in older adults with cognitive disorders using automatic speech analysis," *Journal of Alzheimer's Disease*, vol. 69, no. 4, pp. 1183–1193, 2019.

[100] A. König *et al.*, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, vol. 1, no. 1, pp. 112–124, 2015.

[101] J. P. Teixeira and A. Gonçalves, "Algorithm for Jitter and Shimmer Measurement in Pathologic Voices," *Procedia Computer Science*, vol. 100, pp. 271–279, 2016.

[102] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana, “Speech in alzheimer’s disease: Can temporal and acoustic parameters discriminate dementia?,” *Dementia and Geriatric Cognitive Disorders*, vol. 37, no. 5–6, pp. 327–334, 2014.

[103] F. Eyben *et al.*, “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing,” *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[104] B. Hammarberg, B. Fritzell, J. Gaufin, J. Sundberg, and L. Wedin, “Perceptual and acoustic correlates of abnormal voice qualities,” *Acta Oto-Laryngologica*, vol. 90, no. 1–6, pp. 441–451, 1980.

[105] A. A. Wisler, A. R. Fletcher, and M. J. McAuliffe, “Predicting montreal cognitive assessment scores from measures of speech and language,” *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 6, pp. 1752–1761, 2020.

[106] T. Alhanai, R. Au, and J. Glass, “Spoken language biomarkers for detecting cognitive impairment,” in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2018, no. 2, pp. 409–416.

[107] B. Sumali *et al.*, “Speech quality feature analysis for classification of depression and dementia patients,” *Sensors (Switzerland)*, vol. 20, no. 12, pp. 1–17, 2020.

[108] K. C. Fraser, F. Rudzicz, and G. Hirst, “Detecting late-life depression in Alzheimer’s

disease through analysis of speech and language," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, 2016, pp. 1–11.

[109] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, 2013, pp. 835–838.

[110] M. Nasrolahzadeh, J. Haddadnia, and S. Rahnamayan, "Multi-Objective Optimization of Wavelet-Packet-Based Features in Pathological Diagnosis of Alzheimer Using Spontaneous Speech Signals," *IEEE Access*, vol. 8, pp. 112393–112406, 2020.

[111] L. J. P. Altmann, D. Kempler, and E. S. Andersen, "Speech errors in Alzheimer's disease: Reevaluating morphosyntactic preservation," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 5, pp. 1069–1082, 2001.

[112] J. Kornfilt, "Parts of speech, lexical categories, and word classes in morphology," in *Oxford Research Encyclopedia of Linguistics*, 2020.

[113] L. L. Murray, "Distinguishing clinical depression from early Alzheimer's disease in elderly people: Can narrative analysis help?," *Aphasiology*, vol. 24, no. 6–8, pp. 928–939, 2010.

[114] D. N. Ripich, T. Fritsch, E. Ziol, and E. Durand, "Compensatory strategies in picture description across severity levels in Alzheimer's disease: A longitudinal study," *American Journal of Alzheimer's Disease and Other Dementias*, vol. 15, no. 4, pp. 217–

228, 2000.

[115] C. Guinn and A. Habash, “Language analysis of speakers with dementia of the Alzheimer’s type,” *AAAI Fall Symposium Series*, pp. 8–13, 2012.

[116] K. Croot, J. R. Hodges, J. Xuereb, and K. Patterson, “Phonological and articulatory impairment in Alzheimer’s disease: A case series,” *Brain and Language*, vol. 75, no. 2, pp. 277–309, 2000.

[117] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, “Detecting signs of dementia using word vector representations,” in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2018, pp. 1893–1897.

[118] J. C. Dunn, O. P. Almeida, L. Barclay, A. Waterreus, and L. Flicker, “Latent Semantic Analysis: A new method to measure prose recall,” *Journal of Clinical and Experimental Neuropsychology*, vol. 24, no. 1, pp. 26–35, 2002.

[119] A. Rozeva and S. Zerkova, “Assessing semantic similarity of texts - Methods and algorithms,” *AIP Conference Proceedings*, vol. 1910, 2017.

[120] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” pp. 1–12, 2013.

[121] J. Pennington, R. Socher, and C. D. Manning, “GloVe : Global Vectors for Word Representation,” in *Proceedings of the 2014 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[122] E. Altszyler, M. Sigman, S. Ribeiro, and D. F. Slezak, "Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database," pp. 1–14, 2016.

[123] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[124] A. Roshanzamir, H. Aghajan, and M. Soleymani Baghshah, "Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, pp. 1–14, 2021.

[125] J. O. de Lira, T. S. C. Minett, P. H. F. Bertolucci, and K. Z. Ortiz, "Analysis of word number and content in discourse of patients with mild to moderate Alzheimer's disease," *Dementia & Neuropsychologia*, vol. 8, no. 3, pp. 260–265, 2014.

[126] K. Sirts, O. Piguet, and M. Johnson, "Idea density for predicting Alzheimer's disease from transcribed speech," no. ID, pp. 322–332, 2017.

[127] J. M. J. Vonk *et al.*, "Semantic network function captured by word frequency in nondemented APOE ε4 carriers," *Neuropsychology*, vol. 33, no. 2, pp. 256–262, 2019.

[128] K. E. Forbes-Mckay, A. W. Ellis, M. F. Shanks, and A. Venneri, "The age of acquisition of words produced in a semantic fluency task can reliably differentiate normal from

pathological age related cognitive decline," *Neuropsychologia*, vol. 43, no. 11, pp. 1625–1632, 2005.

[129] H. Stadthagen-gonzalez and C. J. Davis, "The Bristol norms for age of acquisition, imageability, and familiarity," *Behavior Research Methods*, vol. 38, no. 4, pp. 598–605, 2006.

[130] S. Sherratt, "Multi-level discourse analysis: A feasible approach," *Aphasiology*, vol. 21, no. 3–4, pp. 375–393, 2007.

[131] T. L. Mitzner and S. Kemper, "Oral and written language in late adulthood: Findings from the Nun Study," *Experimental Aging Research*, vol. 29, no. 4, pp. 457–474, 2003.

[132] B. H. Davis and M. MacLagan, "Examining pauses in alzheimer's discourse," *American Journal of Alzheimer's Disease and other Dementias*, vol. 24, no. 2, pp. 141–154, 2009.

[133] A. Kertesz, *Western aphasia battery test manual*. Psychological Corporation, 1982.

[134] M. Bros, *Juvenile Collection.. The Circus procession*. Mcloughlin Bros, 1888.

[135] J.-L. Nespoulous *et al.*, "Protocole Montréal-Toulouse d'examen linguistique de l'aphasie (MT86)," *Isbergues, France: L'Ortho-Edition*, 1992.

[136] M. Mayer, *Frog, where are you?* Penguin, 2003.

[137] J. S. Ehrlich, L. K. Obler, and L. Clark, "Ideational and semantic contributions to narrative production in adults with dementia of the Alzheimer's type," *Journal of*

Communication Disorders, vol. 30, no. 2, pp. 79–99, 1997.

[138] R. C. Marshall and H. H. Wright, "Developing a clinician-friendly aphasia test," 2007.

[139] M. Shimada *et al.*, "Impaired verbal description ability assessed by the Picture Description Task in Alzheimer's disease," *Archives of Gerontology and Geriatrics*, vol. 27, no. 1, pp. 57–65, 1998.

[140] J. O. De Lira, T. S. C. Minett, P. H. F. Bertolucci, and K. Z. Ortiz, "Evaluation of macrolinguistic aspects of the oral discourse in patients with Alzheimer's disease," *International Psychogeriatrics*, vol. 31, no. 9, pp. 1343–1353, 2019.

[141] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "Identifying Mild Cognitive Impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features," *Computer Speech and Language*, vol. 53, pp. 181–197, 2019.

[142] G. Walker, L. A. Morris, H. Christensen, B. Mirheidari, M. Reuber, and D. J. Blackburn, "Characterising spoken responses to an intelligent virtual agent by persons with mild cognitive impairment," *Clinical Linguistics and Phonetics*, vol. 00, no. 00, pp. 1–16, 2020.

[143] S. Kiran, J. L. Harris, T. P. Marquardt, and M. Palmieri, "Communication Wellness Check-ups: Age-related changes in communication," in *San Diego, CA: ASHA National Convention*, 2005.

[144] S. J. Hays, B. E. Niven, H. P. D. Godfrey, and R. J. Linscott, “Clinical assessment of pragmatic language impairment: A generalisability study of older people with Alzheimer’s disease,” *Aphasiology*, vol. 18, no. 8, pp. 693–714, 2004.

[145] E. Giles, K. Patterson, and J. R. Hodges, “Performance on the Boston Cookie Theft picture description task in patients with early dementia of the Alzheimer’s type: Missing information,” *Aphasiology*, vol. 10, no. 4, pp. 395–408, 1996.

[146] E. G. March, P. Pattison, and R. Wales, “The role of cognition in context-dependent language use: Evidence from Alzheimer’s disease,” *Journal of Neurolinguistics*, vol. 22, no. 1, pp. 18–36, 2009.

[147] M. Asgari, J. Kaye, and H. Dodge, “Predicting mild cognitive impairment from spontaneous spoken utterances,” *Alzheimer’s and Dementia: Translational Research and Clinical Interventions*, vol. 3, no. 2, pp. 219–228, 2017.

[148] I. Hoffmann, D. Nemeth, C. D. Dye, M. Pákáski, T. Irinyi, and J. Kálmán, “Temporal parameters of spontaneous speech in Alzheimer’s disease,” *International Journal of Speech-Language Pathology*, vol. 12, no. 1, pp. 29–34, 2010.

[149] D. Van Lancker Sidtis, J. Choi, A. Alken, and J. J. Sidtis, “Formulaic Language in Parkinson’s Disease and Alzheimer’s Disease: Complementary Effects of Subcortical and Cortical Dysfunction,” *Journal of Speech, Language, and Hearing Research*, vol. 58, pp. 1493–1507, 2015.

[150] S. Baillon *et al.*, “Neuropsychological performance in Alzheimer’s disease and vascular dementia: comparisons in a memory clinic population,” *International Journal of Geriatric Psychiatry*, vol. 19, no. 12, pp. 1140–1147, 2004.

[151] A. Beales, A. Whitworth, J. Cartwright, P. K. Panegyres, and R. T. Kane, “Determining stability in connected speech in primary progressive aphasia and Alzheimer’s disease,” *International Journal of Speech-Language Pathology*, vol. 20, no. 3, pp. 361–370, 2018.

[152] V. B. Fleming and J. L. Harris, “Test-retest discourse performance of individuals with mild cognitive impairment,” *Aphasiology*, vol. 23, no. 7–8, pp. 940–950, 2009.

[153] S. Suribhatla *et al.*, “Neuropsychological performance in early and late onset Alzheimer’s disease: Comparisons in a memory clinic population,” *International Journal of Geriatric Psychiatry*, vol. 19, no. 12, pp. 1140–1147, 2004.

[154] S. C. J. Verfaillie *et al.*, “High amyloid burden is associated with fewer specific words during spontaneous speech in individuals with subjective cognitive decline,” *Neuropsychologia*, vol. 131, pp. 184–192, 2019.

[155] F. De Anna *et al.*, “Intrusions in story recall: When over-learned information interferes with episodic memory recall. Evidence from Alzheimer’s disease,” *Cortex*, vol. 44, no. 3, pp. 305–311, 2008.

[156] S. Ash and M. Grossman, “Why study connected speech production?,” *Cognitive Neuroscience of Natural Language Use*, pp. 29–58, 2015.

[157] D. Beltrami *et al.*, “Automatic identification of mild cognitive impairment through the analysis of Italian spontaneous speech productions,” in *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 2016, pp. 2086–2093.

[158] L. Brandão, F. G. Castelló, T. A. van Dijk, M. A. de M. P. Parente, and J. Peña-Casanova, “Cognition and discourse production in Alzheimer’s disease: Using informative prompts.,” *Psychology & Neuroscience*, vol. 2, no. 2, pp. 147–155, 2009.

[159] K. A. Bridges and D. Van Lancker Sidtis, “Formulaic language in Alzheimer’s disease,” *Aphasiology*, vol. 27, no. 7, pp. 799–810, 2013.

[160] B. M. Broderick, S. L. Tou, and E. M. Provost, “Cogid: a Speech Recognition Tool for Early Detection of Alzheimer’S Disease,” *Alzheimer’s & Dementia*, vol. 14, pp. 191–192, 2018.

[161] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, “Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance,” *Aphasiology*, vol. 14, no. 1, pp. 71–91, 2000.

[162] E. Can and G. Kuruoğlu, “Spontaneous speech in dementia,” *Psiholingvistika*, vol. 26, no. 2, pp. 175–189, 2019.

[163] S. B. Chapman, A. P. Highley, and J. L. Thompson, “Discourse in fluent aphasia and

Alzheimer's disease: Linguistic and pragmatic considerations," *Journal of Neurolinguistics*, vol. 11, no. 1–2, pp. 55–78, 1998.

[164] S. B. Chapman, J. Zientz, M. Weiner, R. Rosenberg, W. Frawley, and M. H. Burns, "Discourse changes in early Alzheimer disease, mild cognitive impairment, and normal aging," *Alzheimer Disease and Associated Disorders*, vol. 16, no. 3, pp. 177–186, 2002.

[165] Y. W. Chien, S. Y. Hong, W. T. Cheah, L. H. Yao, Y. L. Chang, and L. C. Fu, "An Automatic Assessment System for Alzheimer's Disease Based on Speech Using Feature Sequence Generator and Recurrent Neural Network," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[166] J. D. Choi, M. Li, F. Goldstein, and I. Hajjar, "Meta-Semantic Representation for Early Detection of Alzheimer's Disease," no. Section 4, pp. 82–91, 2019.

[167] K. Dijkstra, M. S. Bourgeois, R. S. Allen, and L. D. Burgio, "Conversational coherence: Discourse analysis of older adults with and without dementia," *Journal of Neurolinguistics*, vol. 17, no. 4, pp. 263–283, 2004.

[168] A. Duong, A. Tardif, and B. Ska, "Discourse about discourse: What is it and how does it progress in Alzheimer's disease?," *Brain and Cognition*, vol. 53, no. 2, pp. 177–180, 2003.

[169] C. Drummond *et al.*, "Narrative impairment, white matter damage and CSF biomarkers in the Alzheimer's disease spectrum," *Aging*, vol. 11, no. 20, pp. 9188–9208, 2019.

[170] S. T. Farias *et al.*, “Idea density measured in late life predicts subsequent cognitive trajectories: Implications for the measurement of cognitive reserve,” *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, vol. 67 B, no. 6, pp. 677–686, 2012.

[171] M. Farivar, Z. G. Anaraki, F. Derakhshandeh, N. Baharloei, and M. Poorjavad, “Narrative discourse in Persian-Speaking patients with mild Alzheimer’s disease,” *Dementia e Neuropsychologia*, vol. 13, no. 2, pp. 225–231, 2019.

[172] Y. Faroqi-Shah, A. Treanor, N. B. Ratner, B. Ficek, K. Webster, and K. Tsapkini, “Using narratives in differential diagnosis of neurodegenerative syndromes,” *Journal of Communication Disorders*, vol. 85, no. April, p. 105994, 2020.

[173] K. C. Fraser, K. L. Fors, M. Eckerström, F. Öhman, and D. Kokkinakis, “Predicting MCI status from multimodal language data using cascaded classifiers,” *Frontiers in Aging Neuroscience*, vol. 10, no. JUL, pp. 1–18, 2019.

[174] K. C. Fraser, K. Lundholm Fors, and D. Kokkinakis, “Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment,” *Computer Speech and Language*, vol. 53, pp. 121–139, 2019.

[175] K. Forbes-McKay, M. F. Shanks, and A. Venneri, “Profiling spontaneous speech decline in Alzheimer’s disease: A longitudinal study,” *Acta Neuropsychiatrica*, vol. 25, no. 6, pp. 320–327, 2013.

[176] M. C. Gély-Nargeot, B. Ska, and J. Touchon, “Text structure and content modulate the recall of patients with dementia of the Alzheimer’s type,” *Brain and Cognition*, vol. 48, no. 2–3, pp. 371–375, 2002.

[177] G. Glosser, “Patterns of Discourse Production among Neurological Patients with Fluent Language Disorders,” *Brain and Language*, vol. 40, pp. 67–88, 1990.

[178] K. Groves-Wright, J. Neils-Strunjas, R. Burnett, and M. J. O’Neill, “A comparison of verbal and written language in Alzheimer’s disease,” *Journal of Communication Disorders*, vol. 37, no. 2, pp. 109–130, 2004.

[179] C. Guinn, B. Singer, and A. Habash, “A comparison of syntax, semantics, and pragmatics in spoken language among residents with Alzheimer’s disease in managed-care facilities,” *IEEE SSCI 2014 - 2014 IEEE Symposium Series on Computational Intelligence - CICARE 2014: 2014 IEEE Symposium on Computational Intelligence in Healthcare and e-Health, Proceedings*, pp. 98–103, 2015.

[180] F. Haider, S. De La Fuente, and S. Luz, “An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer’s Dementia in Spontaneous Speech,” *IEEE Journal on Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2020.

[181] J. L. Harris, S. Kiran, T. Marquardt, and V. Fleming, “Communication wellness check-up©: Age-related changes in communicative abilities,” *Aphasiology*, vol. 22, no. 7–8, pp. 813–825, 2008.

[182] R. B. Heller, A. R. Dobbs, and B. G. Rule, “Communicative function in patients with questionable Alzheimer’s disease.,” *Psychology and aging*, vol. 7, no. 3. pp. 395–400, 1992.

[183] L. Hernández-Domínguez, S. Ratté, G. Sierra-Martínez, and A. Roche-Bergua, “Computer-based evaluation of Alzheimer’s disease and mild cognitive impairment patients during a picture description task,” *Alzheimer’s and Dementia: Diagnosis, Assessment and Disease Monitoring*, vol. 10, pp. 260–268, 2018.

[184] I. Hoffmann, D. Nemeth, C. D. Dye, M. Pákáski, T. Irinyi, and J. Kálmán, “Temporal parameters of spontaneous speech in Alzheimer’s disease,” *International Journal of Speech-Language Pathology*, vol. 12, no. 1, pp. 29–34, 2010.

[185] J. Illes, “Neurolinguistic features of spontaneous language production dissociate three forms of neurodegenerative disease: Alzheimer’s, Huntington’s, and Parkinson’s,” *Brain and Language*, vol. 37, no. 4, pp. 628–642, 1989.

[186] G. Kavé and A. Dassa, “Severity of Alzheimer’s disease and language features in picture descriptions,” *Aphasiology*, vol. 32, no. 1, pp. 27–40, 2018.

[187] S. Kemper, M. Thompson, and J. Marquis, “Longitudinal Change in Language Production: Effects of Aging and Dementia on Grammatical Complexity and Propositional Content,” *Pychology and Aging*, vol. 16, no. 4, pp. 600–614, 2001.

[188] A. Khodabakhsh, S. Kusxuoglu, and C. Demiroglu, “Natural language features for

detection of Alzheimer's disease in conversational speech," in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2014, pp. 581–584.

[189] A. Khodabakhsh, F. Yesil, E. Guner, and C. Demiroglu, "Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, 2015.

[190] B. S. Kim, Y. B. Kim, and H. H. Kim, "Discourse measures to differentiate between mild cognitive impairment and healthy aging," *Frontiers in Aging Neuroscience*, vol. 11, p. 221, 2019.

[191] Y. H. Lai and Y. Te Lin, "Discourse markers produced by Chinese-speaking seniors with and without Alzheimer's disease," *Journal of Pragmatics*, vol. 44, no. 14, pp. 1982–2003, 2012.

[192] Y. H. Lai, "Discourse features of chinese-speaking seniors with and without alzheimer's disease," *Language and Linguistics*, vol. 15, no. 3, pp. 411–434, 2014.

[193] M. Laine, M. Laakso, E. Vuorinen, and J. Rinne, "Coherence and informativeness of discourse in two dementia types," *Journal of Neurolinguistics*, vol. 11, no. 1–2, pp. 79–87, 1998.

[194] M. Lehr, E. Prud'Hommeaux, I. Shafran, and B. Roark, "Fully automated

neuropsychological assessment for detecting mild cognitive impairment," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[195] J. Liu, J. Zhao, and X. Bai, "Syntactic Impairments of Chinese Alzheimer's Disease Patients from a Language Dependency Network Perspective," *Journal of Quantitative Linguistics*, vol. 28, no. 3, pp. 253–281, 2020.

[196] S. Lock and L. Armstrong, "Cohesion analysis of the expository discourse of normal, fluent aphasic and demented adults: A role in differential diagnosis?," *Clinical Linguistics and Phonetics*, vol. 11, no. 4, pp. 299–317, 1997.

[197] K. López-de-Ipiña *et al.*, "On Automatic Diagnosis of Alzheimer's Disease Based on Spontaneous Speech Analysis and Emotional Temperature," *Cognitive Computation*, vol. 7, no. 1, pp. 44–55, 2013.

[198] K. D. Mueller *et al.*, "Connected Language in Late Middle-Aged Adults at Risk for Alzheimer's Disease," *Journal of Alzheimer's Disease*, vol. 54, no. 4, pp. 1539–1550, 2016.

[199] B. Mirheidari *et al.*, "Toward the Automation of Diagnostic Conversation Analysis in Patients with Memory Complaints," *Journal of Alzheimer's Disease*, vol. 58, no. 2, pp. 373–387, 2017.

[200] M. Nasrolahzadeh, Z. Mohammadpoory, and J. Haddadnia, "A novel method for early diagnosis of Alzheimer's disease based on higher-order spectral estimation of

spontaneous speech signals," *Cognitive Neurodynamics*, vol. 10, no. 6, pp. 495–503, 2016.

[201] N. Nevler, S. Ash, D. Irwin, M. Liberman, and M. Grossman, "Apoe E4 Allele Effects on Speech in Ad and Lvppa," *Alzheimer's & Dementia*, vol. 15, pp. P1020–P1021, 2019.

[202] M. Nicholas, L. K. Obler, M. L. Albert, and N. Helm-Estabrooks, "Empty speech in Alzheimer's disease and fluent aphasia," *Journal of Speech and Hearing Research*, vol. 28, no. 3, pp. 405–410, 1985.

[203] R. Nitrini, "Immediate recall of short stories depends on educational level," *Dementia & Neuropsychologia*, vol. 2, no. 4, pp. 310–314, 2008.

[204] S. O. Orimaye, J. S. M. Wong, K. J. Golden, C. P. Wong, and I. N. Soyiri, "Predicting probable Alzheimer's disease using linguistic deficits and biomarkers," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–13, 2017.

[205] S. O. Orimaye, J. S. M. Wong, and C. P. Wong, "Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia," *PLoS ONE*, vol. 13, no. 11, pp. 1–15, 2018.

[206] R. Ostrand and J. Gunstad, "Using Automatic Assessment of Speech Production to Predict Current and Future Cognitive Function in Older Adults," *Journal of Geriatric Psychiatry and Neurology*, 2020.

[207] R. Perri, L. Fadda, C. Caltagirone, and G. A. Carlesimo, "Word list and story recall elicit

different patterns of memory deficit in patients with Alzheimer's disease, frontotemporal dementia, subcortical ischemic vascular disease, and lewy body dementia," *Journal of Alzheimer's Disease*, vol. 37, no. 1, pp. 99–107, 2013.

[208] A. Pistono, M. Jucla, C. Bézy, B. Lemesle, J. Le Men, and J. Pariente, "Discourse macrolinguistic impairment as a marker of linguistic and extralinguistic functions decline in early Alzheimer's disease," *International Journal of Language and Communication Disorders*, vol. 54, no. 3, pp. 390–400, 2019.

[209] A. Pompili, A. Abad, D. M. De Matos, and I. P. Martins, "Pragmatic Aspects of Discourse Production for the Automatic Identification of Alzheimer's Disease," *IEEE Journal on Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 261–271, 2020.

[210] A. L. Powell, J. L. Cummings, M. A. Hill, and D. F. Benson, "Speech and language alterations in multi-infarct dementia," *Neurology*, vol. 38, no. 5, pp. 717–719, 1988.

[211] E. T. Prud'hommeaux and B. Roark, "Extraction of narrative recall patterns for neuropsychological assessment," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[212] E. Prud'hommeaux and B. Roark, "Graph-Based Word Alignment for Clinical Language Evaluation," *Association for Computational Linguistics*, vol. 41, no. 4, pp. 549–578, 2015.

[213] V. Rentoumi, L. Raoufian, S. Ahmed, C. A. De Jager, and P. Garrard, "Features and

machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology," *Journal of Alzheimer's Disease*, vol. 42, pp. S3–S17, 2014.

[214] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken Language Derived Measures for Detecting Mild Cognitive Impairment," *IEEE Trans Audio Speech Lang Processing*, vol. 19, no. 7, pp. 2081–2090, 2011.

[215] B. Romero and A. Kwz, "Deterioration of spontaneous speech in ad patients during a 1-year follow-up: Homogeneity of profiles and factors associated with progression," *Dementia and Geriatric Cognitive Disorders*, vol. 7, no. 1, pp. 35–40, 1996.

[216] S. A. Sajjadi, K. Patterson, M. Tomek, and P. J. Nestor, "Abnormalities of connected speech in semantic dementia vs Alzheimer's disease," *Aphasiology*, vol. 26, no. 6, pp. 847–866, 2012.

[217] J. Shi *et al.*, "The Chinese version of story recall: A useful screening tool for mild cognitive impairment and Alzheimer's disease in the elderly," *BMC Psychiatry*, vol. 14, no. 1, pp. 1–10, 2014.

[218] W. Simpson, L. Kaufman, A. Balagopalna, and J. Novikova, "Comparison of a Speech-Based Digital Biomarker With the Moca in a Naturalistic Cohort of Seniors," *Alzheimer's & Dementia*, vol. 15, no. 2, pp. P1523–P1523, 2019.

[219] R. A. Sluis *et al.*, "An Automated Approach to Examining Pausing in the Speech of

People With Dementia," *American journal of Alzheimer's disease and other dementias*, vol. 35, pp. 1–8, 2020.

[220] S. R. Smith, H. J. Chenery, and B. E. Murdoch, "Semantic abilities in dementia of the Alzheimer type. II. Grammatical semantics," *Brain and Language*, vol. 36, no. 4, pp. 533–542, 1989.

[221] F. Smolik, H. Stepankova, M. Vyhnalek, T. Nikolai, K. Horakova, and S. Matejka, "Propositional Density in Spoken and Written Language of Czech-Speaking Patients With Mild Cognitive Impairment," *Journal of Speech, Language and Hearing Research and Hearing Research*, vol. 59, no. December, pp. 1461–1470, 2016.

[222] S. J. Stevens, "Differentiating the language disorder in dementia from dysphasia - The potential of a screening test," *European Journal of Disorders of Communication*, vol. 27, pp. 275–288, 1992.

[223] D. Stück *et al.*, "Novel Digital Voice Biomarkers of Dementia From the Framingham Study," *Alzheimer's & Dementia*, vol. 14, no. 7, pp. P778–P779, 2018.

[224] V. Taler, P. S. R. Davidson, C. Sheppard, and J. Gardiner, "A discourse-theoretic approach to story recall in aging and mild cognitive impairment," *Aging, Neuropsychology, and Cognition*, 2020.

[225] C. Thomas, V. Kešelj, N. Cercone, K. Rockwood, and E. Asp, "Automatic Detection and Rating of Dementia of Alzheimer Type through Lexical Analysis of Spontaneous

speech," in *Proceedings of the IEEE International Conference on Mechatronics & Automation*, 2005, vol. 5, pp. 1569–1574.

[226] C. M. Toledo *et al.*, "Analysis of macrolinguistic aspects of narratives from individuals with Alzheimer's disease, mild cognitive impairment, and no cognitive impairment," *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, vol. 10, pp. 31–40, 2018.

[227] L. Toth *et al.*, "A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech," *Current Alzheimer Research*, vol. 15, no. 2, pp. 130–138, 2018.

[228] M. V. Treviso, L. B. dos Santos, C. Shulby, L. C. Hubner, L. L. Mansur, and S. M. Aluiso, "Detecting mild cognitive impairment in narratives in Brazilian Portuguese: first steps towards a fully automated system," *Letras de Hoje*, vol. 53, no. 1, pp. 48–58, 2018.

[229] E. Tsantali, D. Economidis, and M. Tsolaki, "Could language deficits really differentiate Mild Cognitive Impairment (MCI) from mild Alzheimer's disease?," *Archives of Gerontology and Geriatrics*, vol. 57, no. 3, pp. 263–270, 2013.

[230] V. Vincze *et al.*, "Detecting mild cognitive impairment by exploiting linguistic information from transcripts," *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Short Papers*, no. 2004, pp. 181–187, 2016.

[231] A. H. Wicklund, N. Johnson, A. Rademaker, B. B. Weitner, and S. Weintraub, "Word

list versus story memory in Alzheimer disease and frontotemporal dementia," *Alzheimer Disease and Associated Disorders*, vol. 20, no. 2, pp. 86–92, 2006.

[232] B. Yu, J. R. Williamson, J. C. Mundt, and T. F. Quatieri, "Speech-based automated cognitive impairment detection from remotely-collected cognitive test audio," *IEEE Access*, vol. 6, pp. 40494–40505, 2018.

[233] G. Kavé and Y. Levy, "Morphology in picture descriptions provided by persons with Alzheimer's disease," *Journal of Speech, Language, and Hearing Research*, vol. 46, no. 2, pp. 341–352, 2003.

[234] D. Beltrami, G. Gagliardi, R. R. Favretti, E. Ghidoni, F. Tamburini, and L. Calzà, "Speech analysis by natural language processing techniques: A possible tool for very early detection of cognitive decline?," *Frontiers in Aging Neuroscience*, vol. 10, no. November, 2018.

[235] J. Keller and T. Rech, "Towards a modular description of the deficits in spontaneous speech in dementia," *Journal of Pragmatics*, vol. 29, no. 3, pp. 313–332, 1998.

[236] D. Povey *et al.*, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[237] L. Tóth *et al.*, "Automatic detection of Mild cognitive impairment from spontaneous speech using ASR," in *Proceedings of International Speech Communication Association*, 2015.

[238] C. Fellbaum, “WordNet : An Electronic Lexical Database,” *Language*, vol. 76, no. 3, pp. 706–708, 2000.

[239] J. C. Dunn, O. P. Almeida, L. Barclay, A. Waterreus, and L. Flicker, “Latent Semantic Analysis: A new method to measure prose recall,” *Journal of Clinical and Experimental Neuropsychology*, vol. 24, no. 1, pp. 26–35, 2002.

[240] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O’Reilly Media, Inc, 2009.

[241] K. Fraser and G. Hirst, “Detecting semantic changes in Alzheimer’s disease with vector space models,” in *Proceedings of LREC 2016 Workshop: Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016)*, 2016.

[242] P. Boersma and V. van Heuven, “Speak and unSpeak with Praat,” *Glot International*, vol. 5, no. 9–10, pp. 341–347, 2001.

[243] F. Eyben, M. Wöllmer, and B. Schuller, “openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor,” in *18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[244] T. Schrank and B. Schuppler, “Automatic detection of uncertainty in spontaneous German dialogue,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 1972–1976.

[245] P. Juola, “Using the Google N-Gram corpus to measure cultural complexity,” *Literary and linguistic computing*, vol. 28, no. 4, pp. 668–675, 2013.

[246] S. Ahmed and P. Garrard, “Spoken discourse in Alzheimer’ disease: a review,” *Linguistica*, vol. 52, no. 1, pp. 9–25, 2012.

[247] “Prevent Dementia.” [Online]. Available: https://preventdementia.co.uk/. [Accessed: 20-Apr-2020].

[248] C. Lancaster *et al.*, “Gallery Game: Smartphone-based Assessment of Long-Term Memory in Adults at Risk of Alzheimer’s Disease,” *bioRxiv*, vol. 00, no. 00, p. 599175, 2019.

[249] K. Ritchie, R. Guilhem, R. . Craig, B. Alain, and P. Vanessa, “COGNITO: Computerized Assessment of Information Processing,” *Journal of Psychology & Psychotherapy*, vol. 04, no. 02, 2014.

[250] H. Kim, J. E. Sung, and J. H. Jeong, “Non-transcription analysis of connected speech in mild cognitive impairment using an information unit scoring system,” *Journal of Neurolinguistics*, vol. 61, no. October 2021, p. 101035, 2022.

[251] L. Huang, Z. Liu, and Y. Li, “Incompleteness features in the descriptive discourse of Chinese elders with and without Alzheimer’s disease,” *Clinical Linguistics and Phonetics*, vol. 00, no. 00, pp. 1–16, 2022.

[252] M. R. McNeil, P. J. Doyle, T. R. D. Fossett, G. H. Park, and A. J. Goda, “Reliability and

concurrent validity of the information unit scoring metric for the story retelling procedure," *Aphasiology*, vol. 15, no. 10–11, pp. 991–1006, 2001.

[253] T. Smith, M. Noble, S. Noble, G. Wright, D. McLennan, and E. Plunkett, *The English Indices of Deprivation 2015: Research Report*, no. September. 2015.

[254] J. S. Roberts *et al.*, "Who seeks genetic susceptibility testing for Alzheimer's disease? Findings from a multisite, randomized clinical trial," *Genetics in Medicine*, vol. 6, no. 4, pp. 197–203, 2004.

[255] W. Y. Zou, R. Socher, D. Cer, and C. D. Manning, "Bilingual word embeddings for phrase-based machine translation," in *Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1393–1398.

[256] M. Bansal, K. Gimpel, and K. Livescu, "Tailoring continuous word representations for dependency parsing," in *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014*, 2014, vol. 2, pp. 809–815.

[257] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-To-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the AAAI conference on artificial intelligence (Vol. 30, No. 1)*, 2016, pp. 3776–3783.

[258] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020.

[259] B. McFee *et al.*, "librosa: Audio and music signal analysis in python," in *Proceedings*

of the 14th python in science conference, 2015, vol. 8, pp. 18–25.

[260] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, p. e0144610, 2015.

[261] P. Boersma, "Praat: doing phonetics by computer," *http://www. praat. org/*, 2007.

[262] A. Koenecke *et al.*, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, no. 14, pp. 7684–7689, 2020.

[263] M. Johnson *et al.*, "A systematic review of speech recognition technology in health care," *BMC Medical Informatics and Decision Making*, vol. 14, no. 1, 2014.

[264] E. Edwards *et al.*, "Medical speech recognition: Reaching parity with humans," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10458 LNAI, pp. 512–524, 2017.

[265] M. Nakamura, K. Iwano, and S. Furui, "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance," *Computer Speech and Language*, vol. 22, no. 2, pp. 171–184, 2008.

[266] M. A. Ulasik, M. Hürlimann, F. Germann, E. Gedik, F. Benites, and M. Cieliebak, "CEASR: A corpus for evaluating automatic speech recognition," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 6477–6485.

[267] “Amazon Transcribe,” *Amazon*. [Online]. Available: https://aws.amazon.com/transcribe/. [Accessed: 14-Nov-2022].

[268] A. Hannun *et al.*, “Deep Speech: Scaling up end-to-end speech recognition,” pp. 1–12, 2014.

[269] “Google Cloud Speech-To-Text,” *Google*, 2021. [Online]. Available: https://cloud.google.com/speech-to-text. [Accessed: 14-Nov-2022].

[270] “IBM Watson,” *IBM Watson*. [Online]. Available: https://www.ibm.com/cloud/watson-speech-to-text. [Accessed: 14-Nov-2022].

[271] “Kaldi - A speech-to-text system for quick, cost free transcription.” [Online]. Available: https://www.bbc.co.uk/makerbox/tools/kaldi. [Accessed: 20-Nov-2021].

[272] “Azure Speech to Text,” *Microsoft*. [Online]. Available: https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text/. [Accessed: 14-Nov-2022].

[273] T. Audacity, “Audacity,” *The Name Audacity (R) Is a Registered Trademark of Dominic Mazzoni Retrieved from http://audacity. sourceforge. net*, 2017.

[274] B. MacWhinney, *The CHILDES project: The database*, vol. 2. Psychology Press, 2000.

[275] K. Heafield, “KenLM: Faster and Smaller Language Model Queries,” in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.

[276] S. K. Gaikwad, B. W. Gawali, and P. Yannawar, “A Review on Speech Recognition

Technique," *International Journal of Computer Applications*, vol. 10, no. 3, pp. 16–24, 2010.

[277] F. Kelly, O. Forth, A. Atreya, S. Kent, and A. Alexander, "What your voice says about you: Automatic Speaker Profiling using i-vectors," in *International Association of Forensic Phonetics and Acoustics (IAFPA) conference*, 2017.

[278] A. Alexander, O. Forth, A. A. Atreya, and F. Kelly, "VOCALISE: A forensic automatic speaker recognition system supporting spectral, phonetic, and user-provided features," in *Proc. Odyssey*, 2016.

[279] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith, "Retrofitting word vectors to semantic lexicons," in *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, no. i, pp. 1606–1615.

[280] N. Mrkšić *et al.*, "Counter-fitting word vectors to linguistic constraints," *In Proceedings of the NAACL HLT*, pp. 142–148, 2016.

[281] M. Rei and T. Briscoe, "Looking for Hyponyms in Vector Space," pp. 68–77, 2015.

[282] V. Vincze, "Uncertainty detection in Hungarian texts," *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1844–1853, 2014.

[283] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification,"

in *Proceedings of 56th Annual Meeting of the Association for Computational Linguistics*, 2018, vol. 1, pp. 328–339.

[284] M. Richardson, C. J. C. Burges, and E. Renshaw, "MCTest: A challenge dataset for the open-domain machine comprehension of text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 193–203.

[285] E. Casanova, M. V. Treviso, L. C. Hübner, and S. M. Aluísio, "Evaluating sentence segmentation in different datasets of neuropsychological language tests in brazilian Portuguese," *Proceedings of 12th International Conference on Language Resources and Evaluation*, pp. 2605–2614, 2020.

[286] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," *arXiv preprint arXiv:1603.01354*, 2016.

[287] C. J. Van Rijsbergen, "Foundation of evaluation," *Journal of documentation*, vol. 30, no. 4, pp. 365–373, 1974.

[288] A. M. Robertson and P. Willet, "Applications of n-grams in textual information systems," *Journal of Documentation*, vol. 54, no. 1, pp. 48–69, 1998.

[289] C. Silva and B. Ribeiro, "The Importance of Stop Word Removal on Recall Values in Text Categorization," in *Proceedings of the International Joint Conference on Neural Networks*, 2003, vol. 3, pp. 1661–1666.

[290] N. H. de Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech

rate automatically," *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, 2009.

[291] D. Vohra, "Apache parquet," in *Practical Hadoop Ecosystem*, Springer, 2016, pp. 325–335.

[292] S. Hayakawa and F. Itakura, "Text-dependent speaker recognition using the information in the higher frequency band," in *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, 1994, vol. 1, pp. I–137.

[293] J. Daniel, *Data science with Python and Dask*. Simon and Schuster, 2019.

[294] J. M. Bland and D. G. Altman, "Agreement between methods of measurement with multiple observations per individual," *Journal of Biopharmaceutical Statistics*, vol. 17, no. 4, pp. 571–582, 2007.

[295] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[296] F. E. Satterthwaite, "Synthesis of variance," *Psychometrika*, vol. 6, no. 5, pp. 309–316, 1941.

[297] D. Giavarina, "Understanding Bland Altman Analysis," *Biochemia Medica*, vol. 25, no. 2, pp. 141–151, 2015.

[298] D. G. Altman and J. M. Bland, "Measurement in Medicine: the Analysis of Method

Comparison Studies," in *The Statistician*, 1983, vol. 32, pp. 307–317.

[299] J. L. McGaugh, "Memory--a century of consolidation," *Science*, vol. 287, no. 5451, pp. 248–251, 2000.

[300] D. Hakkani-Tür, D. Vergyri, and G. Tur, "Speech-based automated cognitive status assessment," in *Proceedings of the Eleventh Annual Conference of the International Speech Communication Association*, 2010, pp. 258–261.

[301] S. Baudic, G. D. Barba, M. C. Thibaudet, A. Smagghe, P. Remy, and L. Traykov, "Executive function deficits in early Alzheimer's disease and their relations with episodic memory," *Archives of Clinical Neuropsychology*, vol. 21, no. 1, pp. 15–21, 2006.

[302] A. Yeung *et al.*, "Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and Alzheimer's dementia," *Alzheimer's Research and Therapy*, vol. 13, no. 1, pp. 1–10, 2021.

[303] M. Ganguli, B. E. Snitz, C. W. Lee, J. Vanderbilt, J. A. Saxton, and C. C. H. Chang, "Age and education effects and norms on a cognitive test battery from a population-based cohort: The Monongahela-Youghiogheny Healthy Aging Team," *Aging and Mental Health*, vol. 14, no. 1, pp. 100–107, 2010.

[304] A. Ardila, F. Ostrosky-Solis, M. Rosselli, and C. Gómez, "Age-related cognitive decline

during normal aging: The complex effect of education," *Archives of Clinical Neuropsychology*, vol. 15, no. 6, pp. 495–513, 2000.

[305] J. A. Mortimer and A. B. Graves, "Education and other socioeconomic determinants of dementia and alzheimer' disease," *Neurology*, vol. 43, no. 8, pp. S39–S44, 1993.

[306] "Local government geography and history," in *Department for Communities and Local Government*, 2010, pp. 140–141.

[307] M. Tavakol and R. Dennick, "Making sense of Cronbach's alpha," *International journal of medical education*, vol. 2, pp. 53–55, 2011.

[308] J. M. Bland and D. G. Altman, "Statistics notes: Cornbach's alpha," *British Medical Journal*, vol. 314, no. 7080, p. 572, 1997.

[309] C. Schmidt, F. Collette, C. Cajochen, and P. Peigneux, "A time to think: Circadian rhythms in human cognition," *Cognitive Neuropsychology*, vol. 24, no. 7, pp. 755–789, 2007.

[310] H. Wilks *et al.*, "Sharper in the morning: Cognitive time of day effects revealed with high-frequency smartphone testing," *Journal of Clinical and Experimental Neuropsychology*, vol. 43, no. 8, pp. 825–837, 2021.

[311] A. Pratap *et al.*, "Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–10, 2020.

[312] D. Berron *et al.*, “Feasibility of Digital Memory Assessments in an Unsupervised and Remote Study Setting,” *Frontiers in Digital Health*, vol. 4, pp. 1–23, 2022.

[313] D. Kaufman, L. Sauvé, L. Renaud, A. Sixsmith, and B. Mortenson, “Older adults’ digital gameplay: Patterns, benefits, and challenges,” *Simulation & Gaming*, vol. 47, no. 4, pp. 465–489, 2016.

[314] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, “Quantifying Bias in Automatic Speech Recognition,” 2021.

[315] S. P. Woods, D. C. Delis, J. C. Scott, J. H. Kramer, and J. A. Holdnack, “The California Verbal Learning Test - second edition: Test-retest reliability, practice effects, and reliable change indices for the standard and alternate forms,” *Archives of Clinical Neuropsychology*, vol. 21, no. 5, pp. 413–420, 2006.

[316] J. P. Peter, G. A. Churchill, Jr., and T. J. Brown, “Caution in the Use of Difference Scores in Consumer Research,” *Journal of Consumer Research*, vol. 19, no. 4, p. 655, 1993.

[317] D. R. Rogosa and J. B. Willett, “Demonstrating the Reliability of the Difference Score in the Measurement of Change,” *Journal of Educational Measurement*, vol. 20, no. 4, pp. 335–343, 1983.

[318] M. Asperholm, N. Högman, J. Rafi, and A. Herlitz, “What did you do yesterday? A meta-analysis of sex differences in episodic memory.,” *Psychological Bulletin*, vol. 145,

no. 8, p. 785, 2019.

[319] T. M. Hess, L. E. Popham, L. Emery, and T. Elliott, “Mood, motivation, and misinformation: Aging and affective state influences on memory,” *Aging, Neuropsychology, and Cognition*, vol. 19, no. 1–2, pp. 13–34, 2012.

[320] A. Brose, F. Schmiedek, M. Lövdén, and U. Lindenberger, “Daily variability in working memory is coupled with negative affect: The role of attention and motivation,” *Emotion*, vol. 12, no. 3, pp. 605–617, 2012.

[321] D. L. Murman, “The Impact of Age on Cognition,” *Seminars in Hearing*, vol. 36, no. 3, pp. 111–121, 2015.

[322] M. Jitlal *et al.*, “The Influence of Socioeconomic Deprivation on Dementia Mortality, Age at Death, and Quality of Diagnosis: A Nationwide Death Records Study in England and Wales 2001-2017,” *Journal of Alzheimer’s Disease*, vol. 81, no. 1, pp. 321–328, 2021.

[323] P. W. Schofield, K. Marder, G. Dooneief, D. M. Jacobs, M. Sano, and Y. Stern, “Association of subjective memory complaints with subsequent cognitive decline in community-dwelling elderly individuals with baseline cognitive impairment,” *American Journal of Psychiatry*, vol. 154, no. 5, pp. 609–615, 1997.

[324] F. Jessen *et al.*, “Prediction of dementia by subjective memory impairment: effects of severity and temporal association with cognitive impairment,” *Archives of general*

psychiatry, vol. 67, no. 4, pp. 414–422, 2010.

[325] J. L. Molinuevo *et al.*, "Implementation of subjective cognitive decline criteria in research studies," *Alzheimer's & Dementia*, vol. 13, no. 3, pp. 296–311, 2017.

[326] K. D. Mueller *et al.*, "Connected Language in Late Middle-Aged Adults at Risk for Alzheimer's Disease," *Journal of Alzheimers Disease*, vol. 54, no. 4, pp. 1539–1550, 2016.

[327] E. M. Saffran, R. S. Berndt, and M. F. Schwartz, "The quantitative analysis of agrammatic production: Procedure and data," *Brain and Language*, vol. 37, no. 3, pp. 440–479, 1989.

[328] M. Brysbaert, I. Van Wijnendaele, and S. De Deyne, "Age-of-acquisition effects in semantic processing tasks," *Acta Psychologica*, vol. 104, no. 2, pp. 215–226, 2000.

[329] A. Rofes, V. de Aguiar, R. Jonkers, S. J. Oh, G. DeDe, and J. E. Sung, "What Drives Task Performance During Animal Fluency in People With Alzheimer's Disease?," *Frontiers in Psychology*, vol. 11, no. July, pp. 1–11, 2020.

[330] L. L. Saling, N. Laroo, and M. M. Saling, "When more is less: Failure to compress discourse with re-telling in normal ageing," *Acta Psychologica*, vol. 139, no. 1, pp. 220–224, 2012.

[331] O. Juncos-Rabadán, A. X. Pereiro, and M. S. Rodríguez, "Narrative speech in aging: Quantity, information content, and cohesion," *Brain and Language*, vol. 95, no. 3, pp.

423–434, 2005.

[332] J. W. Pennebaker and L. D. Stone, “Words of Wisdom: Language Use Over the Life Span,” *Journal of Personality and Social Psychology*, vol. 85, no. 2, pp. 291–301, 2003.

[333] P. Barrouillet, S. Bernardin, S. Portrat, E. Vergauwe, and V. Camos, “Time and cognitive load in working memory,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 3, pp. 570–585, 2007.

[334] S. V. S. Pakhomov *et al.*, “Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration,” *Cognitive and Behavioral Neurology*, vol. 23, no. 3, pp. 165–177, 2010.

[335] C. T. Ferrand, “Harmonics-to-noise ratio: An index of vocal aging,” *Journal of Voice*, vol. 16, no. 4, pp. 480–487, 2002.

[336] P. Torre III and J. A. Barlow, “Age-related changes in acoustic characteristics of adult speech,” *Journal of communication disorders*, vol. 42, no. 5, pp. 324–333, 2009.

[337] S. Nanga *et al.*, “Review of dimension reduction methods,” *Journal of Data Analysis and Information Processing*, vol. 9, no. 3, pp. 189–231, 2021.

[338] M. Gold *et al.*, “Digital technologies as biomarkers, clinical outcomes assessment, and recruitment tools in Alzheimer’s disease clinical trials,” *Alzheimer’s and Dementia: Translational Research and Clinical Interventions*, vol. 4, pp. 234–242, 2018.

[339] W. N. Whiteley *et al.*, “Are large simple trials for dementia prevention possible ?,” *Age and Ageing*, vol. 49, no. 2, pp. 1–7, 2019.

[340] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, “Alzheimer’s dementia recognition through spontaneous speech: The ADReSS challenge,” in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2020.

[341] P. Mahajan and V. Baths, “Acoustic and Language Based Deep Learning Approaches for Alzheimer’s Dementia Detection From Spontaneous Speech,” *Frontiers in Aging Neuroscience*, vol. 13, p. 623607, 2021.

[342] P. A. Pérez-Toro *et al.*, “Acoustic and linguistic analyses to assess early-onset and genetic Alzheimer’s disease,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8338–8342.

[343] Z. S. Syed, M. S. S. Syed, M. Lech, and E. Pirogova, “Automated Recognition of Alzheimer’s Dementia Using Bag-of-Deep-Features and Model Ensembling,” *IEEE Access*, vol. 9, pp. 88377–88390, 2021.

[344] M. Martinc, F. Haider, S. Pollak, and S. Luz, “Temporal Integration of Text Transcripts and Acoustic Features for Alzheimer’s Diagnosis Based on Spontaneous Speech,” *Frontiers in Aging Neuroscience*, vol. 13, p. 642647, 2021.

[345] A. Valsaraj, I. Madala, N. Garg, and V. Baths, “Alzheimer’s Dementia Detection Using

Acoustic & Linguistic Features and Pre-trained BERT," in *2021 8th International Conference on Soft Computing and Machine Intelligence (ISCMI)*, 2021, pp. 171–175.

[346] M. Rohanian, J. Hough, and M. Purver, "Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs," 2021.

[347] S. K. Pandey, H. S. Shekhawat, S. Bhasin, R. Jasuja, and S. R. M. Prasanna, *Alzheimer's Dementia Recognition Using Multimodal Fusion of Speech and Text Embeddings*. Springer International Publishing, 2022.

[348] Y. Qiao, X. Yin, D. Wiechmann, and E. Kerz, "Alzheimer's disease detection from spontaneous speech through combining linguistic complexity and (Dis)fluency features with pretrained language models," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2021, pp. 2162–2166.

[349] R. Pappagari *et al.*, "Automatic detection and assessment of Alzheimer Disease using speech and language technologies in low-resource scenarios," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2021, pp. 3825–3829.

[350] J. Chen, J. Ye, F. Tang, and J. Zhou, "Automatic detection of Alzheimer's disease using spontaneous speech only," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2021.

[351] R. Haulcy and J. Glass, "Classifying Alzheimer's Disease Using Audio and Text-Based

Representations of Speech," *Frontiers in Psychology*, vol. 11, p. 624137, 2021.

[352] A. Balagopalan, B. Eyre, J. Robin, F. Rudzicz, and J. Novikova, "Comparing Pre-trained and Feature-Based Models for Prediction of Alzheimer's Disease Based on Speech," *Frontiers in Aging Neuroscience*, vol. 13, p. 635945, 2021.

[353] K. D. Mueller *et al.*, "Amyloid beta associations with connected speech in cognitively unimpaired adults," *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, vol. 13, no. 1, pp. 1–10, 2021.

[354] A. König *et al.*, "Screening over Speech in Unselected Populations for Clinical Trials in AD (PROSPECT-AD): Study Design and Protocol," *The Journal of Prevention of Alzheimer's Disease*, pp. 1–8, 2023.

[355] J. Weiner and T. Schultz, "Automatic Screening for Transition into Dementia using Speech," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[356] T. E. Goldberg, P. D. Harvey, K. A. Wesnes, P. J. Snyder, and L. S. Schneider, "Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials," *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, vol. 1, no. 1, pp. 103–111, 2015.

[357] L. Miao, J. Wu, P. Behre, S. Chang, and S. Parthasarathy, "Multilingual Transformer Language Model for Speech Recognition in Low-resource Languages," 2022.

[358] A. Tjandra *et al.*, "Massively Multilingual ASR on 70 Languages: Tokenization,

Architecture, and Generalization Capabilities," 2022.

[359] S. Gregory *et al.*, "Remote data collection speech analysis and prediction of the identification of Alzheimer's disease biomarkers in people at risk for Alzheimer's disease dementia: The Speech on the Phone Assessment (SPeAk) prospective observational study protocol," *BMJ Open*, vol. 12, no. 3, pp. 1–7, 2022.

[360] C. De Looze *et al.*, "Changes in speech chunking in reading aloud is a marker of mild cognitive impairment and mild-to-moderate Alzheimer's disease.," *Current Alzheimer Research*, vol. 15, pp. 828–847, 2018.

[361] J. M. Duchek, D. A. Balota, and V. C. Thessing, "Inhibition of visual and conceptual information during reading in healthy aging and Alzheimer's disease," *Aging, Neuropsychology, and Cognition*, vol. 5, no. 3, pp. 169–181, 1998.