Funded by EU's Horizon 2020

D 6.2

PREDICTIVE MODELS FOR COGNITIVE DECLINE

DISCLAIMER

AKNOWLEDGEMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.769661.

Supporting Active Ageing through Multimodal coaching

DELIVERABLE DOCUMENTATION SHEET

Deliverable:	D 6.2 Predictive Models for Cognitive Decline
WP №	6
Title:	Specialised Well-Being Monitoring and Assessment
Editor(s):	Fasih Haider
Contributor(s):	Saturnino Luz
Type:	Report
Version:	1.0
Submission Due Date:	31.05.2019
Dissemination level:	Public
Copyright:	©SAAM Project - All rights reserved

☒	Approved by the WP Leader
☒	Approved by the Technical/Exploitation Manager[1]
☒	Approved by the Coordinator
☒	Approved by the PSC

[1] Choose Technical Manager for Deliverables in WP1-7,10 and Exploitation Manager in WP 8-10

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

PUBLISHABLE SUMMARY

In the framework of SAAM project, we are developing models to predict the cognitive decline in elderly people. Cognitive decline could lead to the severe mental conditions such as dementia and speech analysis could provide an indicator of Alzheimer's disease and help develop clinical tools for automatically detecting and monitoring disease progression. While previous studies have employed speech (acoustic) features for characterisation of Alzheimer's dementia, these studies focused on a few common prosodic features, often in combination with lexical and syntactic features which require transcription. This Deliverable presents a detailed study of the predictive value of purely acoustic and dialogical features automatically extracted from spontaneous speech for Alzheimer's dementia detection, from a computational paralinguistics perspective. We assess several state-of-the-art acoustic feature sets, including the extended Geneva minimalistic acoustic parameter set (eGeMAPs), the emobase feature set, the ComParE feature set, a multi-resolution cochleagram derived feature set (MRCG) and multiple dialogical features, in conjunction with different dimensionality reduction and machine learning approaches, on the Dementia Bank Pitt spontaneous speech dataset and Carolina Conversation dataset. Results show that classification models based solely on acoustic and dialogical speech features can achieve accuracy levels comparable to those achieved by models that employ higher-level language features.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

QUALITY CONTROL ASSESSMENT SHEET

Version	Date	Comment	Name of author/reviewer/contributor
V0.1	3.05.2019	First Draft	Fasih Haider
	14.05.2019	Contributions	Saturnino Luz
V0.2	29.05.2019	Second Draft	Fasih Haider
	31.05.2019	1st Peer review	Andrej Hrovat
	31.05.2019	2nd Peer review	Bernard Ženko
V1.0	01.06.2019	Final Draft	Fasih Haider
	01.06.2019	WP Leader approval	
	03.06.2019	Coordinator approval	
		EAB Review	n.a.
	05.06.2019	PSC approval	
V1.0	06.06.2019	Submission to EC	

HISTORY OF CHANGES

For updating the Deliverable after submission to the EC if applicable

Version	Date	Change
V2.0		

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

Supporting Active Ageing through Multimodal coaching

PROJECT DOCUMENTATION SHEET

Project Acronym:	SAAM
Project Full Title:	Supporting Active Ageing through Multimodal coaching
Grant Agreement:	GA № 769661
Call identifier:	H2020-SC1-2017-CNECT-1
Topic:	Personalised coaching for well-being and care of people as they age
Action:	Research and Innovation Action
Project Duration:	36 months (1 October 2017 – 30 September 2020)
Project Officer:	Jose Albacete VALVERDE
Coordinator:	Balkan Institute for Labour and Social Policy (BILSP)
Consortium partners:	Jožef Stefan Institute (JSI) University of Edinburgh (UEDIN) Paris-Lodron Universität Salzburg (PLUS) Scale Focus AD (SCALE) Interactive Wear AG (IAW) Univerzitetni rehabilitacijski inštitut Republike Slovenije (SOČA) Nacionalna Katolicheska Federacia CARITAS Bulgaria (CARITAS) Bulgarian Red Cross (BRC) Eurag Osterreich (EURAG)
website:	saam2020.eu
social media:	#saam2020, #saamproject

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

ABBREVIATIONS

AD	Alzheimer's Disease
ADR	Active Data Representation
ASR	Automatic Speech Recognition
DT	Decision Tree
CCC	Carolina Conversations Collection
LDA	Linear Discrimination Analysis
LOSO	Leave-One-Subject-Out
MCI	Mild Cognitive Impairment
MMSE	Mini-Mental State Examination
MRCG	Multi-Resolution Cochleagram
MV	Majority Vote
NLP	Natural Language Processing
SL	Segment-Level
SML	Subjective Memory Loss
SVF	Semantic Verbal Fluency
SVM	Support Vector Machine
WPM	Words per Minute

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

CONTENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

TABLE OF FIGURES

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

1. INTRODUCTION

Dementia is a category of neurodegenerative diseases characterized by long-term and usually gradual decrease of cognitive functioning. It is characterised by a set of symptoms that include memory loss, thought difficulties, defective executive functions (e.g. problem-solving, decision-making, planning), language impairment, motor problems, lack of motivation and emotional distress. Throughout the disease, the severity of these symptoms increases at the expense of the patient's autonomy, as well as their well-being and their caregivers' [1]. Those cognitive symptoms may be a consequence of the neuropathology of different diseases, such as Alzheimer's disease (AD; 50 % of dementia cases), cerebrovascular disease (25 % cases, including those mixed with AD), Lewy body disease (15 % cases), and other brain diseases (5 %), including Parkinson's disease, frontotemporal dementia and stroke [2].

The main and most obvious risk factor for dementia is age, and therefore its greatest incidence is amongst the elderly. Since the population over 65 years old is predicted to triple between years 2000 and 2050 [3], dementia care is projected to have an immense societal impact. In 2015, the WHO [4] estimated approximately 47.5 million cases of dementia worldwide, with longitudinal cohort studies finding an annual incidence between 10 and 15 cases every one thousand people, from which between 5 and 8 would be caused by Alzheimer's disease. The prognosis is difficult, with around 7 years of average life expectancy and less than 3 % patients living longer than 14 years after diagnosis [4].

Due to the severity of the situation worldwide, institutions and researchers are investing countless resources in dementia prevention and early detection, by studying the different stages of the disease. There is a need for cost-effective and scalable methods for detection of dementia from its most subtle forms, such as the preclinical stage of Subjective Memory Loss (SML), to more severe conditions like Mild Cognitive Impairment (MCI) and Alzheimer's Dementia (AD) itself.

The neuropathology of AD consists of several phenomena, including intracellular accumulation of tau-protein fibres [5] and extracellular accumulation of beta-amyloid plaques [6]. Both are responsible for brain damage and neural functional disruption [7] and there is no satisfactory treatment for dementia in general, nor for AD in particular. Furthermore, neuropathology is known to start silently up to 20 years before an individual shows obviously observable cognitive symptoms. Therefore, it is paramount to find strategies to detect the problem as early as possible, in order to enhance therapy effectiveness and quality of life [8].

This deliverable focuses on AD recognition using acoustic and dialogical information extracted from spontaneous speech. Whilst memory loss is frequently considered the most prominent symptom of AD [9], speech and language alterations are also common [10, 11]. Patients with AD usually display early naming and word-finding difficulties (anomia) leading to circumlocution [12], while showing relatively intact speech fluency, auditory comprehension, articulation, prosody and repetition [9].

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

Literature also suggests that patients with AD have difficulty accessing semantic information intentionally, reflecting a general semantic deterioration [13].

The heterogeneity of the symptomatic expression of AD requires diagnosis support methods that are able to capture more subtle aspects than conventional screening tools, which often fail to discriminate these symptoms in pre-clinical AD. Social signal processing technologies are creating opportunities for personal health monitoring and development of tools to predict AD based on processing of behavioural signals [14]. Such tools might aid clinicians in early and accurate differential diagnosis of dementia [15]. Speech and language are ubiquitous sources of cognitive behavioural data, and computational analysis of these signals could form the basis for such tools.

There are several commonly used cognitive assessments for dementia diagnosis that involve linguistic tests - such as the Mini-Mental State Examination (MMSE) [16], the five-word test [17], the frontal assessment battery [18], and the instrumental activities of daily living scale [19]. Speech continuity, for instance, may be assessed through picture description tasks [20] or through countdown tasks [21] and Semantic Verbal Fluency (SVF) usually involves naming tasks [22]. However, whilst valuable for aiding diagnosis, most of these neuropsychological tests offer little insight into early stages of neurodegeneration and hence there is an increasing interest developing alternative methods for early detection. For instance, the study by König et al. [21], recorded a sentence repeating task and evaluated the waveforms with a standardized signal processing technique (dynamic time warping). They looked at the alignment curve between pairs of corresponding waveforms to see whether there is a significant difference between the sentences produced by the clinician (presumably healthy) and the sentences produced by the AD patients [21].

However, one disadvantage of these tests is that they employ speech and language generated under controlled, laboratory conditions, while the use of spontaneous speech, which would be the ideal for monitoring tools, is uncommon. One of the few spontaneous speech resources linked to neuropsychological and clinical assessment for dementia is the Pitt Cookie Theft dataset [23], distributed through DementiaBank[2]. This dataset consists of speech from participants who were recorded while performing the Boston Cookie Theft picture description task, from the Boston diagnostic aphasia examination [24, 25, 26]. Machine learning has been employed on this corpus for detection of cognitive impairment through the analysis of linguistic and para-linguistic features [27, 28, 29]. However, these works focused on linguistic features, taking advantage of the manual transcriptions made available along with the speech data, or used limited, ad hoc acoustic feature sets. Furthermore, they did not adjust the dataset for potential confounders in age and gender imbalances, or the effects of variable audio quality. The work presented in this report addresses these issues by assessing a comprehensive set of acoustic features, exclusively, on a gender- and age-balanced subset of the Pitt and Carolina conversions corpus which has been pre-processed to ensure fairly uniform audio quality across the dataset.

[2] http://dementia.talkbank.org/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

The machine learning models proposed in this report will facilitate SAAM in identifying the speech impairment in elderly people which could be associated with AD and helps the SAAM in suggesting an intervention. This research contributes to research in AD detection by:

1. creating a balanced sub-corpus of the Pitt Cookie Theft Dataset from DementiaBank and making it available to the research community,
2. carrying out a factor analysis to identify (and group) those acoustic features that may be discriminative of manifestations of AD pathology in speech behaviour,
3. assessing acoustic information extracted automatically from spontaneous speech as potential 'digital biomarkers',
4. demonstrating the discriminative power of different feature sets (such as eGeMaps, emobase, ComParE and MRCG) and their fusion for automatic recognition of Alzheimer's disease,
5. demonstrating the discriminative power of dialogical feature sets, and
6. testing these features on different machine learning models to implement automatic classification of patients with respect to a probable AD diagnosis.

2. BACKGROUND

The complex multimodal nature of AD and its behavioural manifestations calls for increasingly interdisciplinary research, combining aspects from social signal processing, artificial intelligence, cognitive psychology, computational linguistics, medicine, neuropsychology, computer science and other disciplines. Although research on language and AD has focused on conventional aspects of language (i.e. lexicon, syntax, semantics), the analysis of continuous speech is progressively attention as a source of information to support diagnosis of MCI, AD and related conditions [30, 27, 28, 29, 31].

Findings from language research propose features like information content, comprehension of complexity, picture naming and word-list generation, as promising predictors of disease progression [32]. A study by Roark et al. [33] used Natural Language Processing (NLP) and Automatic Speech Recognition (ASR) to automatically-annotate and time-align certain spoken language features (pause frequency and duration). Their results show that ASR and NLP were at least as accurate as manual methods to annotate the transcripts and evaluate speech parameters. Jarrold et al. [34] introduced a multi-layered perceptron in their analysis of language samples and achieved an accuracy of 88 % in binary classifications of AD vs. healthy subjects based on lexical and acoustic features. A more recent study by Luz et al. [35] extracted simple dialogical features (turn-taking patterns and speech rate) from the Carolina Conversations Collection [36] and used these features to create an additive logistic regression model [37] which obtained an accuracy of 85 % when distinguishing dialogues involving an AD speaker from non-AD dialogues.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

Acoustic feature analysis in AD research generally entails the use of signal processing methods to extract such features, before their predictive potential is evaluated for prediction tasks. Subtle acoustic signs of neurodegeneration may be imperceptible to human diagnosticians. Toth et al. [38], for instance, found that filled pauses (i.e. sounds such as 'hmmm', etc) could not be reliably detected by human annotators, whereas detection improved by using an ASR system. Furthermore, the study found that ASR-extracted features performed best in combination with machine learning methods including Naive Bayes, Random Forest and Support Vector Machine (SVM) techniques, achieving up to 78.8 % accuracy when distinguishing MCI patients from the healthy control group. These results outperformed manually calculated features (74.0 %) even though only a small, non-standardised set of acoustic features was used. This suggests that speech processing could produce more suitable features for early detection of dementia than high level human-annotated speech features [39, 40]. Similar machine learning methods were used by König et al. [21], who reported an accuracy of 79 % when distinguishing MCI participants from their healthy counterparts; 94 % for AD vs. healthy; and 80 % for MCI versus AD, even though their tests were performed on non-spontaneous speech data gathered under controlled conditions, as part of a neuropsychological test, and included manually transcribed text. In another study, spontaneous speech was elicited through a picture description task. Speech parameters extracted from the descriptions were introduced in SVM and Random Forest Classifiers, distinguishing healthy participants from AD participants with an accuracy of 81 %, which increased to 85 % after incorporating MCI participants to the experiment. However, they used a class-imbalanced data set which may be one of the reasons why the method performed worse on the MCI group of participants (yielding a high rate of false negatives) [41].

Studies in this field continually evidence the heterogeneity with which language and speech impairments are displayed in AD and related diseases. Duong et al. [42] ran a cluster analysis with data from picture narratives and concluded that, rather than a common profile, there were several discourse patterns that could be indicative of differences between healthy aging and AD. This heterogeneity seems to be more evident in AD than in specific language diseases such as primary progressive aphasia [43], especially in early stages of AD [44]. In line with this, we hypothesise that acoustic analysis might help identify such discourse patterns and enhance accuracy of early dementia predictive models.

A current research trend in this area is the collection of spontaneous speech data at scale in order to achieve generalisable results. At present, the Pitt data set is one of the few spontaneous speech resources coupled with clinical data which are available and relatively easy to access. Therefore, it has been used in several studies. Among these, the study by Fraser et al. [27] carried out analysis of a large number of parameters. They computed a number of linguistic and acoustic features from the Pitt Corpus Cookie Theft audio files, as well as introducing variables of transcribed language. A factor analysis revealed the latent linguistic factors using these variables and four clear factors emerged: semantic impairment, acoustic abnormality, syntactic impairment and information impairment.

Despite increasing interest, comparisons across these speech studies are difficult, since they use different datasets (which vary in acoustic quality, content, length experimenter and participant expectations, etc.) and a non-standardised variety of processing methods and feature sets. Another

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

difficulty, from a speech processing perspective, is the fact that most studies combine acoustic features with high-level language features which can only be computed once reliable transcription has been done, making it difficult to assess the extent to which classification performance can be obtained by fully automated means. These issues are among the motivations for the present study, where we propose the use of large and standardised sets of acoustic features on a balanced, acoustically pre-processed version of the Pitt Cookie Theft corpus.

3. METHOD AND ANALYSIS

This section describes the dataset, data pre-processing, statistical analysis and machine learning methods which are used to find the underlying acoustic factors that might be relevant to detect participants' cognitive impairment through speech.

3.1 Pitt Corpus

This study specifically uses the Pitt Corpus, gathered longitudinally between 1983 and 1988 on a yearly basis as part of the Alzheimer Research Program at the University of Pittsburgh [23].

Participants are categorised into three groups such as dementia, control (i.e. healthy), and unknown participant. All participants were required to be above 44 years of age, have at least seven years of education, have no history of nervous system disorders or be taking neuroleptic medication, have an initial MMSE score of 10 or more and be able to provide informed consents.

Extensive neuropsychological and physical assessments conducted on the participants are also included; more detailed information of this cohort can be found in [23].

This study selected only the dementia and control groups for a binary diagnosis of AD and non-AD. The Pitt Corpus contains participants' responses to the following:

- Cookie theft stimulus picture for the Control and Dementia groups.
- Word fluency task for the dementia group only.
- Story recall task for the dementia group only.
- Sentence construction task for the dementia group only.

The study specifically chose the cookie theft description task subset. Table 1 lists the data available in this set. Participants were shown the Cookie Theft picture and were asked to describe the picture in their own words.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

Table 1 Statistics of the DementiaBank Pitt corpora

	Control	AD*
Number of patients	99	194
Number of visits (recordings)	242	307
with 1 visit	26	117
with 2 visits	28	53
with 3 visits	28	12
with 4 visits	9	9
with 5 visits	8	3

3.2 Experiment on the Full Pitt Corpus

We stipulate as a basic requirement that the input data for the proposed method must consist of features that can be either readily entered by the user or easily acquired in natural interactive settings with acceptable accuracy. Therefore we assume a scenario where the user enters personal information (e.g. age, gender) and the subsequent monitoring consists of tracking of low-level paralinguistic features and functionals of the user's spontaneous speech, such as the timing and duration of vocalisations and pauses [57, 58] speaking rate, and voice quality measures. In this experiment we focus on vocalisation events and speech rate. These features are easily extracted through basic signal processing and are reasonably robust to environmental noise and diarisation issues [58].

In order to test the potential usefulness of these features in predicting ATD from speech data, we define a categorisation task of differentiating between speech from participants with ATD (represented as *C=a*, or *a*) and speech from control participants (ā). The feature set $F = \{F_1, \dots\dots, F_n\}$ consists of numerical features corresponding to summary statistics (mean, variance, minimum and maximum, entropy) for vocalisation events, speech rate and number of utterances over a discourse event (described below), and the abovementioned nominal features. As the purpose of this work was to assess baseline performance, we chose to use a simple probabilistic model rather than carry out extensive classifier comparisons. We have also not performed any feature selection procedures or parameter tuning. Formally, the probability of a patient being diagnosed with ATD is represented as:

$$P(a|F) \quad \propto \quad P(F_1 = v_1, \dots, F_n = v_n|\alpha) \qquad (1)$$

$$= \quad \prod_{i=1}^{n} P(F_i = v_i|\alpha) \qquad (2)$$

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

Where (2) represents the conditional independence assumption, and $P(F_i=v_i|b)$ are modelled through Gaussian kernels, if F_i is a continuous variable, or estimated by maximum likelihood and incorporated to a multinomial model, if F_i is a nominal variable.

This model was tested on the Pitt data set, available through the DementiaBank[3]. Although the Pitt data set consists of speech recordings from different neuropsychological tests (see [23] for details), we use only the spontaneous speech data gathered for the Boston "cookie theft" picture description task. In this task, the subject is shown a picture and asked to describe the scene depicted. The 551 audio recordings of 247 different participants performing this task included recordings from elderly controls (n=242), people with probable Alzheimer's disease (n=235), and others (n=74). Data were gathered longitudinally, on a yearly basis. The data set also included timed, manually produced transcripts, which have been used in previous work on ATD prediction [27]. We ignored these transcripts for the purposes of this experiment. After excluding recordings with missing data, and other conditions (MCI, vascular dementia etc), we used n=214 ATD and n=184 control recordings.

Vocalisations produced by the instructor were removed. The participant's vocalisations were segmented using a simple silence filter based on the amplitude ratio $\left(\left(\frac{R_s}{R_n}\right)^2\right)$ where R_s and R_n stand for the root mean square amplitude of signal and noise respectively) of the speech signal, setting the silence amplitude rate threshold empirically to -25dB and a duration threshold of 1 second, according to standard practice in vocalisation analysis [6]. Speaking rate was estimated based on a syllable nuclei detection algorithm [59].

Ten-fold cross validation was performed, and accuracy scores were computed averaging over the different folds.

3.2.1. Results

Figure 1 shows the receiver operating characteristic (ROC) curve for the relation between sensitivity (the proportion of ATD patients correctly identified as such) and the probability of a false alarm (the complement of specificity) in the above described model. This model achieved an overall accuracy of 68%, with F_1 scores of 70% for the control class and 64% for the ATD class, as measured against a gold standard diagnostic established through a combination of neuropsychological and neurologic tests [23].

Although inadequate for diagnosis, these results are 36% higher than the baseline, suggesting to us that simple paralinguistic features extracted from noisy audio files have predictive value for ATD

[3] http://www.talkbank.org/DementiaBank/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

detection, and encouraging further exploration of such features, as described in the remaining sections of this deliverable.

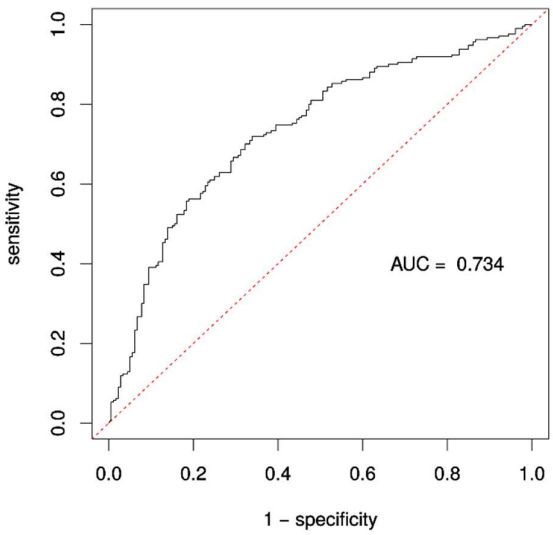


Figure 1 ROC curve for ATD classifier

3.3 Subset Selection from Pitt Corpus

The Pitt Corpus includes both the manual transcripts of the clinical sessions and the corresponding audio recordings for both participant (i.e. AD and non-AD) groups. The transcripts comprise both the speech of the Investigator *INV*) and the Participant (*PAR*). Based on the information provided by DementiaBank, the AD and non-AD groups were not matched with age, gender or education. Therefore, our next step was to create a sub-dataset matched for age and gender to eliminate bias.

This section describes the steps (as shown in Figure 2) taken to create a sub-corpora for statistical analysis and machine learning methods.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

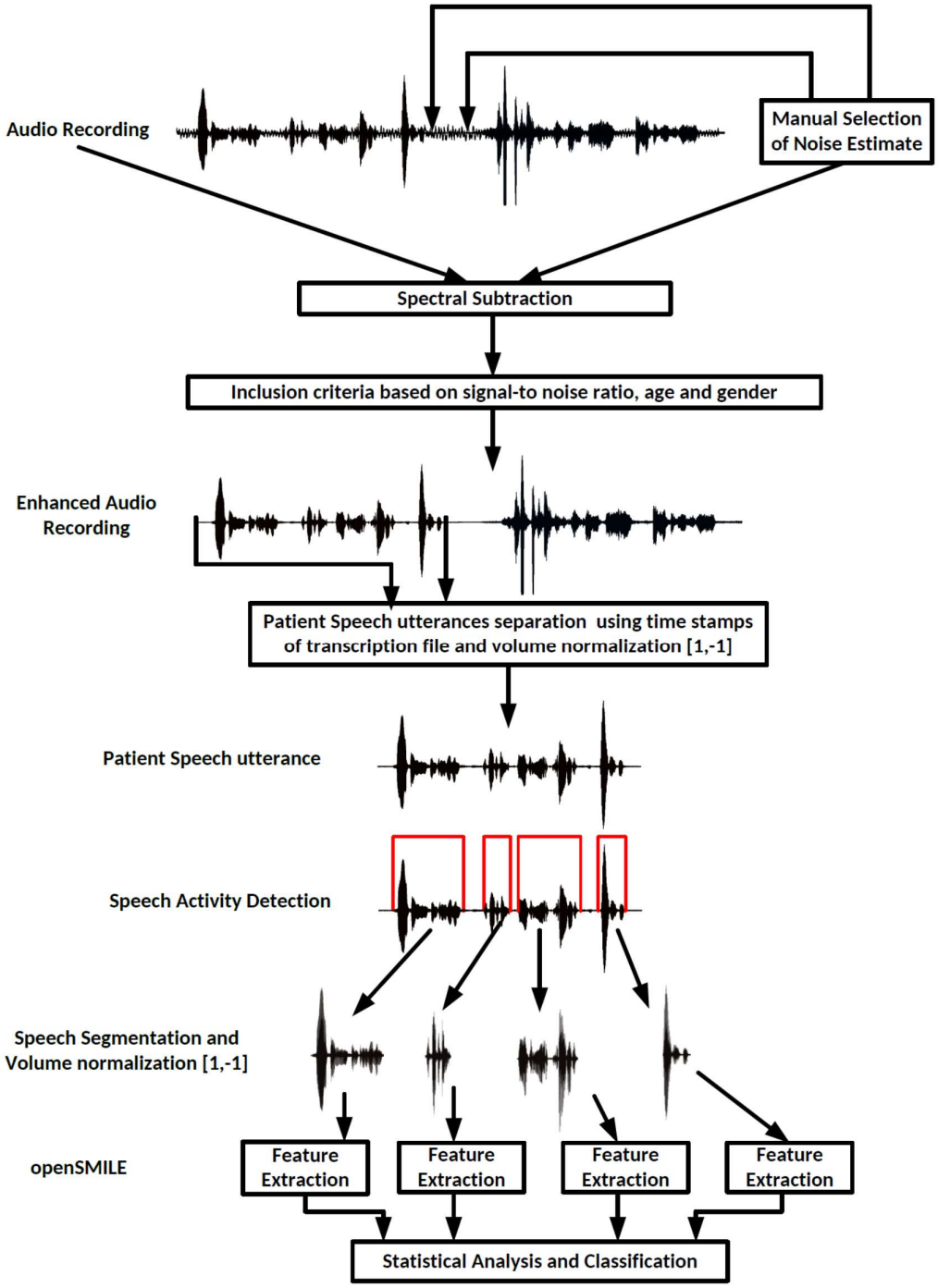


Figure 2 data pre-processing steps

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 769661

3.3.1 Audio Enhancement and Inclusion Criteria

We have manually selected a short interval from each audio recording which contains only the noise and then deploy spectral subtraction method to eliminate that noise. After removing the noise, some audio files still contained some noises such as people talking in the background, ambulance sirens door slamming, etc. That is why we selected the audio files which have higher SNR (i.e. 0 to -17 dB) than the remaining files (***SNR<-17dB***). In case of multiple audio files per subject (i.e. multiple visits) then we selected the most recent audio file of the subject.

3.3.2 Matching the Data for Gender and Age

To eliminate the confounder from the analysis, the dataset should match the AD and non-AD groups for age and gender. Age and gender are considered major risk factors for dementia diagnosis [71]. Therefore, there is a substantial difference in the occurrence of a possible confounder between AD and Control (non-AD) groups. Along with the inclusion criteria in Section 3.3.1, matching gender and age for both AD and non-AD datasets, the homogeneity of the sample population is assured, reducing confounding and increasing the likelihood of finding a true association between exposure and outcomes. The age ranges were chosen empirically to optimise the number of recordings included in the final dataset. As a result, 164 participants matched the selection criteria to be included in the study. Of these, 82 were healthy and 82 were diagnosed with probable AD. After testing the different ranges of the age intervals, the dataset was balanced and could produce the optimal number of recordings by using the age range from 45 to 80 years with the interval of 5 years. Table 2 presents the demographic data. Participants' age in each group ranged from 50-80 years old.

Table 2 Basic characteristics of the patients in each group (AD/non-AD)

	AD		non-AD	
Age Interval	Male	Female	Male	Female
(50, 55)	2	1	2	1
(55, 60)	7	8	7	8
(60, 65)	4	9	4	9
(65, 70)	10	14	10	14
(70, 75)	9	11	9	11
(75, 80)	4	3	4	3
Total	36	46	36	46

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

3.3.3 Speech Segmentation

Speech segmentation was performed on the audio files based on the selection criteria from the previous section. The study only focuses on the participants' speech; therefore, the investigators' speeches are unused and excluded from the entire process.

First, we extracted the participants' speech utterances using the timestamps (start time and end time) transcribed by DementiaBank Pitt Corpus.

However, it is noticed that the participants' speech utterances have long pauses and low volume. That is why we normalised the volume to the range [-1:+1] dBFS and then used speech activity detection (with an energy threshold of 50 dB) method (i.e. auditok[4]) for speech segmentation i.e. to separate the speech from pauses. Later all the speech segments volume were normalised to the range [-1:+1] dBFS. This volume normalization will help in tackling different recording conditions particularly the distance of microphone from the subject.

3.4 Experiment on a Subset of the Pitt Corpus

3.4.1 Acoustic Feature Extraction and Selection from sub-set of Pitt Corpora

Acoustic feature extraction was performed on the speech segments using the openSMILE v2.1 toolkit which is an open-source software suite for automatic extraction of features from speech, widely used for emotion/affect recognition in speech [50]. The following is a brief description of each of the feature sets constructed in this way:

- **emobase:** This acoustic feature set contains the MFCC, voice quality, fundamental frequency (F0), F0 envelope, LSP and intensity features along with their first and second order derivatives. In addition, many statistical functions are applied to these features, resulting in a total of 988 features for every speech utterance.
- **ComParE:** The *ComParE* [51] feature set includes energy, spectral, Mel-Frequency Cepstral Coefficients (MFCCs), and voicing related Low-Level Descriptors (LLDs). LLDs include logarithmic harmonic-to-noise ratio, voice quality features, Viterbi smoothing for F0, spectral harmonicity and psychoacoustic spectral sharpness. This feature set contains 6373 acoustic features for every speech utterance.
- **eGeMAPS:** The *eGeMAPS* [52] feature set contains the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, Hammarberg index and slope V0 features including many statistical ``functionals'' applied on these features, which resulted in a total of 88 features for every speech utterance.

[4] https://pypi.org/project/auditok/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

- **Multi-Resolution Cochleagram (MRCG):** MRCG features have been proposed by Chen et al. [46] and have since been used in speech related applications such as voice activity detection [47] speech separation [46], and more recently for attitude recognition [48]. MRCG features are based on cochleagrams [49]. A cochleagram is generated by applying the gammatone filter to the audio signal, decomposing it in the frequency domain so as to mimic the human auditory filters. MRCG uses the time-frequency representation to encode the multi-resolution power distribution of an audio signal. Four cocheleagram features are generated at different levels of resolution. The high-resolution level encodes local information while the remaining three lower resolution levels capture spectrotemporal information. A total of 768 features are extracted from each frame: 256 MRCG features (frame length of 20ms and frame shift of 10 ms), along with 256 Δ MRCG and 256 ΔΔ MRCG. These features are meant to capture temporal dynamics of the signal [46]. The statistical functionals (mean, standard deviation, minimum, maximum, range, mode, median, skewness and kurtosis) are applied on the 768 MRCG features which resulted in total of 6912 features for every speech utterance.

In sum, we have extracted 88 eGeMAPS, 988 emobase, 6373 ComParE and 6912 MRCG features from 4,077 speech chunks (i.e speech segments). Pearson's correlation test was performed to remove the acoustic features that were significantly correlated with duration (when ***R > 0.2***). Hence, 75 eGeMAPs, 741 emobase, 4510 ComParE and 4691 MRCG features were not correlated with the duration of the speech chunks, and thus, these those features were selected for the machine learning task.

3.4.2 Factor Analysis using eGeMAPs Feature Set

To explore the heterogeneity of speech impairment among participants, we performed an exploratory factor analysis. The 'MVN' package for assessing multivariate normality was used as the dataset does not meet the assumption of normality; PAF analysis was conducted. 75 features were included in PAF with the oblique factor rotation method 'promax' as discussed previously. By conducting analysis for both groups will allow us to investigate the significant differences in acoustic features between AD and non-AD groups.

The parallel analysis (Figure 3) suggests seven factors which can explain the majority of the variance. The exploratory analyses were conducted by rotating seven or six factors. The seven- and six-factor rotations did not, however, produce any theoretically relevant factors. It was decided to retain the five-factor rotation as described below.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

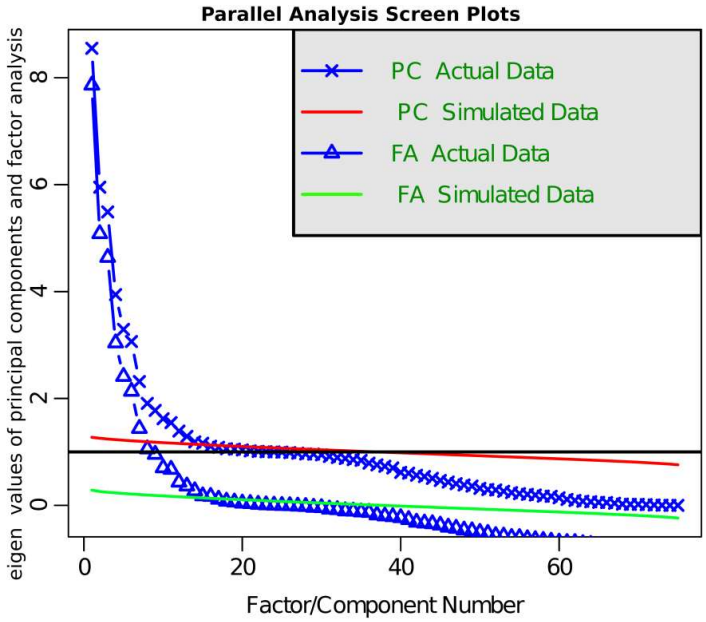


Figure 3 Screen plot for both AD and non-AD groups

- **Factor 1:** This factor reflects the **energy intensity**. It chiefly includes the energy parameters and spectral parameters and has high loadings on loudness variables and spectral flux measurements.
- **Factor 2:** This factor represents the **voice quality** of speech among the patient groups. It equally reflects the frequency and the spectral parameters. The alpha ratio and spectral slope from 0-500HZ are the high loading variables in this factor. Thus, it can reflect the spectral balance based on the energy of speech and further assess the quality of speech.
- **Factor 3:** This factor depicts the dimension of **signal vibration.** It primarily includes the formant amplitude and frequency parameters and has high loadings specifically on the formant amplitude measurements.
- **Factor 4:** This factor represents the characteristic of the **spectral characteristics** which is reflected by the MFCCs.
- **Factor 5:** This factor reflects the characteristics of **frequency intensity**. The high loadings include the formant frequency parameters.

Table 3 presents the results of the exploratory factor analyses for both patient groups. Principal component analysis, using the criterion of eigenvalues greater than 1, reduced the 75 variables to five factors. These five factors explained 35 % of the variance in the data. Some factors emerged from this analysis were identical to that of from the first analysis; however, the main focus of this analysis was on the comparison of these factors between two groups. Out of the five factors emerged from the analysis due to their higher loadings in variance, three show statistically significant differences between AD and non-AD groups as shown in Table 4. This suggests that eGeMAPS features are likely to be the relevant features for predicting the differences between AD and non-AD groups.

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 769661

Table 3 Five factors from both AD and non-AD groups

Factor	Variables	Loadings
1	Loudness percentile20.0	0.94
	Loudness percentile50.0	0.89
	Spectral Flux	0.88
	Loudness	0.87
	Spectral FluxV	0.7
	Loudness Peaks Per Sec	0.61
	Loudness percentile80.0	0.59
	Equivalent Sound Level dBp	0.51
	Spectral Flux UV	0.35
2	Alpha Ratio V	0.9
	SlopeV500.1500	0.59
	F0 semitone From 27.5Hz pctlrange0.2	0.53
	Voiced Segments Per Sec	0.51
	F0 semitone From 27.5Hz (stddevNorm)	0.46
	F1 bandwidth (stddevNorm)	0.42
	Spectral Flux UV	0.37
3	F1 bandwidth (stddevNorm)	0.36
	F3 amplitude LogRel F0	0.78
	F2 amplitude LogRel F0	0.74
	F2 bandwidth (stddevNorm)	0.47
	Alpha Ratio UV	0.44
	F3frequency (stddevNorm)	0.43
	F1frequency (stddevNorm)	0.43
	F3bandwidth (stddevNorm)	0.36
	Loudnesss Rising Slope	0.35
4	mfcc3V	0.73
	mfcc3	0.71
	mfcc4V	0.67
	mfcc4	0.64
	mfcc2V	0.63
	mfcc2	0.59
5	F0 semitone From 27.5Hz percentile20.0	0.54
	F2 frequency	0.86
	F3 frequency	0.83
	F1 frequency	0.8
	F0 semitone From 27.5Hz	0.68
	F0 semitone From 27.5Hz percentile 50.0	0.67
	F0 semitone From 27.5Hz percentile 80.0	0.65

Loadings shown in each factor are specific to that factor

Loadings smaller than 0.35 are excluded

The variables are measured as their arithmetic means if not specifically indicated as other measures

stddev: the standard deviation

stddevNorm: the standard deviation normalised by the arithmetic mean

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 769661

Table 4 results from student's t-test

Factor	t	df	p-value	95% CI	AD mean*	non-AD mean*
1	-2.64	36680	< 0.01	(-0.27, -0.04)	-0.35	-0.19
3	3.20	36444	< 0.01	(0.35, 1.45)	-9.67	-10.57
4	-9.49	24401	< 0.001	(-1.72, -1.13)	4.09	5.52

3.4.3 Predictive Model for Cognitive Decline Using Pitt Corpora

3.4.3.1 Active Data Representation

We have devised a novel method called 'active data representation', which we employed to represent the acoustic features used in this study. Briefly, generating active data representation encompasses the following steps:

- Segmentation and feature extraction: each audio recording Ai (i=1:N, where N represents the total number of audio segments) is divided into n segments $S_{k,Ai}$ using voice activity detection, where k varies from 1 to n. Hence $S_{k,Ai}$ is the k^{th} segment of the i^{th} audio recording, and acoustic features are extracted over such speech segments, rather than over the full audio recording. The system architecture is depicted in Figure 4.
- Clustering of segments: we employed self-organising maps (SOM) [53] for clustering segments $S_{k,Ai}$ into n clusters (C_1 ,C_2,....,C_n) using audio features. Here n represents the number of cluster for SOM.
- Generation of the Active Data Representation (ADR_{Ai}) vector is done by first calculating the number of segments in each cluster for each audio recording (Ai) i.e a histogram representation of the number of segments ($nADR_{Ai}$). Then, to model temporal dynamics we calculate mean and standard deviation values of the rate of change of cluster numbers for each audio recording ($vADR_{Ai}$). Finally, we calculated the duration of segments in each cluster for each audio recording (Ai) i.e histogram representation of duration of segments ($dADR_{Ai}$).

Normalisation: as the number/duration of segments is different for each audio recording (i.e. the duration of all audio recordings is not constant), we normalise the feature vector by dividing it by the total number/duration of subject's segments present in each audio recording (i.e. the L1 norm of $nADR_{Ai}$ and $dADR_{Ai}$), as shown below:

$$nADR_{Ai_{norm}} = \frac{nADR_{Ai}}{\left|\left|nADR_{Ai}\right|\right|_1}, \qquad dADR_{Ai_{norm}} = \frac{dADR_{Ai}}{\left|\left|dADR_{Ai}\right|\right|_1}$$

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

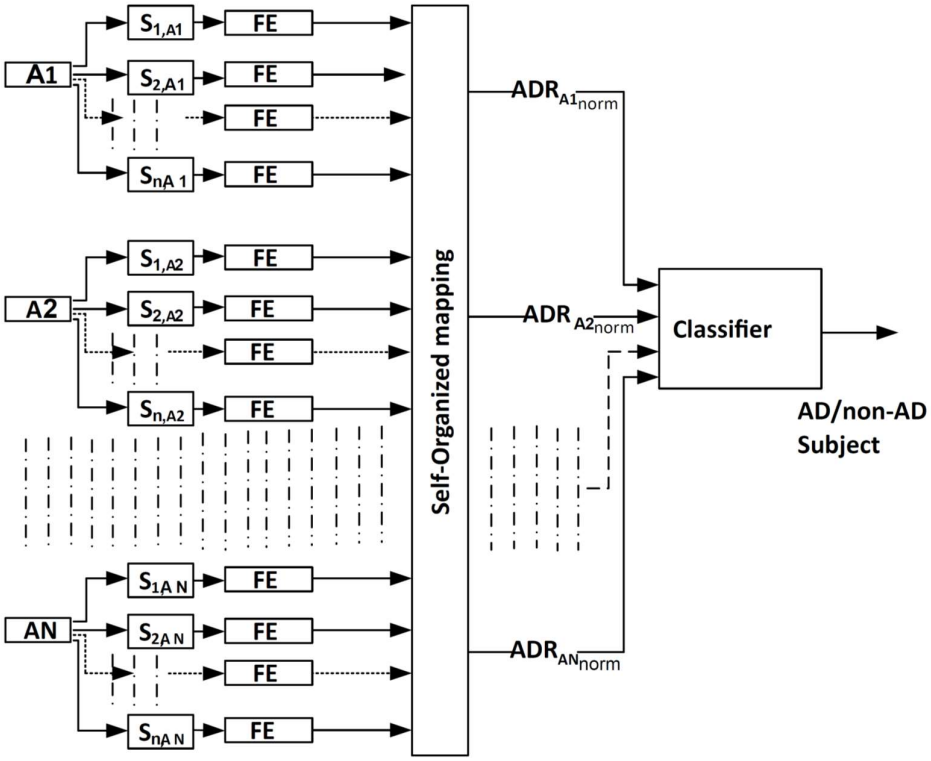


Figure 4 Automatic detection of AD and non-AD subject using the Active Data Representation (ADRAinorm) method where FE represents the extraction of low level features (such as eGeMaps) from speech segments.

3.4.3.2 Classification Methods

The classification is performed using five different methods namely Decision Tree (DT, with leaf size of 20), Nearest Neighbour (KNN with K=1), Linear Discrimination Analysis (LDA), Random Forest (RF with 50 trees and a leaf size of 20) and Support Vector Machines (SVM, with a linear kernel with box constraint of 0.1 and SMO solver). The classification methods are employed in MATLAB[5] using the statistics and machine learning toolbox in the Leave-One-Subject-Out (LOSO) cross-validation setting, where the training data do not contain any information of validation subjects.

3.4.3.3 AD Detection

We conducted three classification experiments to detect cognitive impairment due to AD, namely:

[5] http://uk.mathworks.com/products/matlab/ (December 2018)

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 769661

- Segment-Level (SL) Classification: in this experiment we trained and tested our classifiers in a LOSO setting, with acoustic features, age and gender to predict whether the speech segments were uttered by a non-AD or AD patient.
- Majority Vote (MV): after segment-level classification, we calculated the number of segments detected as AD and non-AD for each subject and then took a majority vote to assign an overall label to the subject.
- Active Data Representation: We generate the ADR using acoustic features as described in section 3.4.3.1, and then used ADR_{Ainorm} for classification as before.

3.4.3.4 Results and Discussion

The results of segment level, majority vote classification and ADR are shown in Table 5, Table 6 and Table 7, respectively. These results show that the ADR (77.44 %) provides better results than majority vote (61.59 %) for all classifiers, with LDA being the best classifier for AD detection. For further insight, the confusion matrices of the best results of each experiment (i.e. segment level, majority vote and ADR) are also shown in Figure 5, Figure 6 and Figure 7.

Table 5 Segment Level Classification

	LDA	DT	1NN	SVM	RF
eGeMaps	49.98	50.64	48.90	49.78	55.03
emobase	52.53	55.10	50.07	55.05	**56.55**
compare	54.88	48.06	51.64	52.63	53.51
MRCG	50.22	52.01	51.57	52.67	54.17

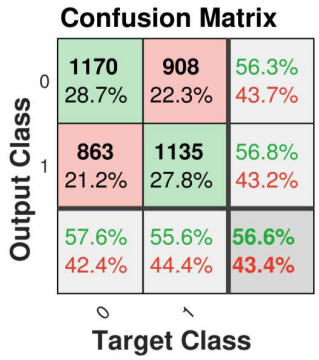


Figure 5 confusion matrices of best results of emobase features

From the results, we note that even though LDA provides the best result (77.44 %) DT also exhibits promising performance, being in fact more stable across all feature sets than the other classifiers (the best average accuracy of 69.05 %). To better understand the relationship between the feature sets for the DT classifier, we drew the Venn diagram depicted in Figure 8. In that Figure, the blue area (labelled 'Target') represents the annotated labels, the yellow area represents the predicted

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

labels when the *ComParE* feature set was used, the green ellipse represents the predicted labels when *eGeMAPs* was used, the red ellipse represents the prediction obtained with the *emobase* feature set, and finally the brown ellipse represents labels predicted with the MRCG features.

Table 6 Majority Vote Classification

	LDA	DT	1NN	SVM	RF
eGeMAPs	50.61	51.83	48.17	50.61	60.98
emobase	53.66	56.10	48.17	56.71	57.93
compare	**61.59**	46.95	53.05	54.88	58.54
MRCG	50.61	54.88	54.27	56.10	56.10

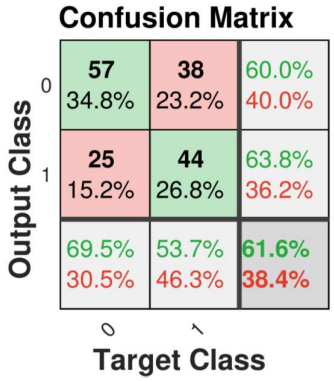


Figure 6 confusion matrices of best results of compare features

From the overlaps in this Venn diagram, it is observed that there are 6 instances (1 of non-AD, and 5 of AD) which have not been recognised by any of the feature sets. However there are 46 instances (32 of non-AD and 14 of AD) which have been detected by all four feature sets. The Venn diagram suggest that although the accuracy results for all feature sets do not vary by a large margin, the information captured by them is not similar, as only 46 out of 164 instances are detected by all the feature sets. This suggests that the fusion of the results could improve overall accuracy. We implemented a simple 'hard fusion' strategy by taking a vote among all four feature sets, breaking ties by considering them as implying an AD label. As hypothesised, fusion provides the best results, with an accuracy of 78.7 % as shown in Figure 8.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

Table 7 Active Data representation Classification

	LDA	DT	1NN	SVM	RF
eGeMaps	77.44, 85	71.34, 30	54.27, 65	52.44, 20	71.34, 30
emobase	56.10, 30	66.46, 20	54.88, 80	45.12, 15	60.98, 25
compare	57.93, 35	68.90, 95	55.49, 100	59.76, 35	60.37, 95
MRCG	59.76, 5	69.51, 15	52.44, 95	59.76, 15	63.41, 15
mean	62.81	69.05	54.27	54.27	64.03

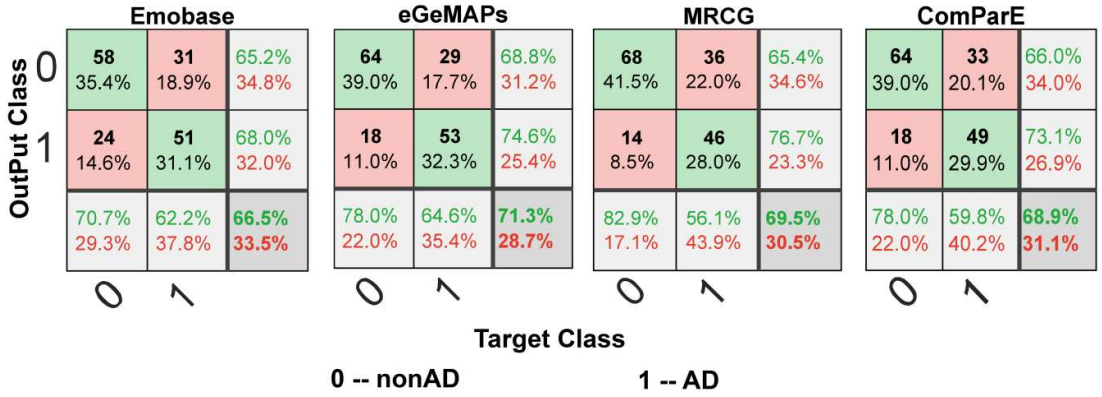


Figure 7 confusion matrices of decision tree obtained using active data representation

In this study, the machine learning models are trained using the acoustic features of speech segments. The reported results are quite promising, indicating that the paralinguistic feature sets may be detecting the presence of abnormality in the voices of AD patients.

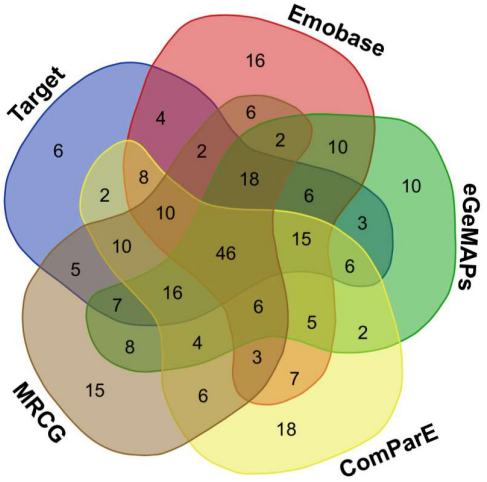
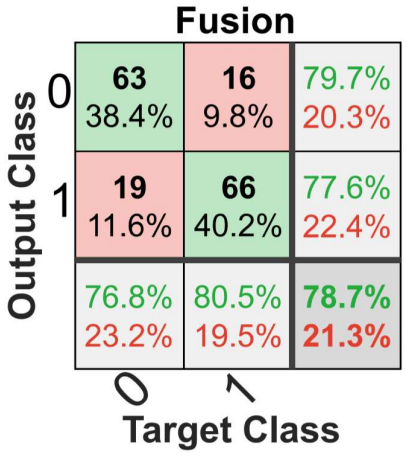


Figure 8 Venn diagram of the best results of four feature sets

Figure 9 Hard Decision fusion of decision tree results



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

3.5 Modelling Cognitive Decline Using Dialogue Data.

In line with the SAAM project settings and overall design philosophy, we aim to provide cognitive monitoring based on spontaneous speech, gathered in natural activities of daily living. The above described studies employed narrative monologues. While these data represent a form of spontaneous speech, monologue is not a frequent occurrence in daily life. We therefore sought to explore the use of dialogues, which are by far the most natural and frequently occurring form of speech. This section reports the results of our analysis of a dataset of dialogues featuring people with cognitive impairment, namely the Carolina Conversations Collection [36]. The dataset is a digital collection of recordings of conversations about health, including both audio and video data, with corresponding transcriptions. The corpus consists of natural conversations involving an older person (over the age of 65) with a medical condition. Several demographic and clinical variables are also available, including: age range, gender, occupation prior to retirement, disease diagnosed, and level of education (in years).

The interviewers were gerontology and linguistic students or researchers to whom the patients spoke at least twice a year. A unique alias was assigned to each patient to protect their identity. Access to the data was provided after complying with the ethical requirements of the University of Edinburgh and the Medical University of South Carolina. In order to ensure that the results described here are reproducible we will provide, on request, the identifiers for the dialogues used in our experiments so that interested researchers can recreate our dataset upon being granted access to the CCC. The source code used for processing the data is available at a University of Edinburgh gitlab server[6].

For the research described here we selected a total of 38 patient dialogues: 21 patients had a diagnosis of Alzheimer's disease (15 females, 6 males), and 17 patients (12 females, 5 males) had other diseases (diabetes, cardiac issues, etc., excluding neuropsychological conditions), but not AD. These groups were selected for matching age ranges and gender frequencies so as to avoid statistical bias. The dataset also included time-aligned transcripts, which we did not use except for the computation of an alternative speech rate feature as described below.

3.5.1 Data Preparation

The speech data selected as previously described were pre-processed in order to generate vocalisation graphs — that is, Markov diagrams encoding the first-order conditional transition probabilities between vocalisation events and steady-state probabilities [58].Vocalisation events are classified as speech by either the patient or the interviewer/ others, joint talk (overlapping speech), or silence events (also known as 'floor' events, which are further in the diagrams as pauses and

[6]https://cybermat.tardis.ed.ac.uk/pial/CCCdataset

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

switching pauses, according to whether the floor is taken by the same speaker or another speaker, respectively). An example of vocalisation graph is shown in Figure 10.

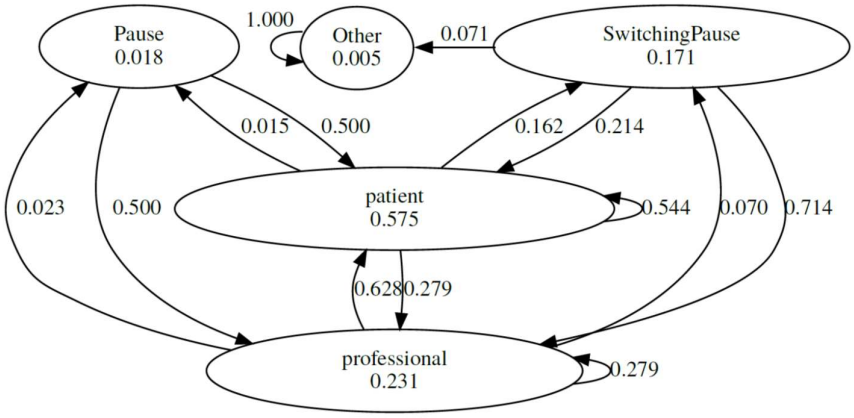


Figure 10 Vocalisation diagramme for a patient dialogue.

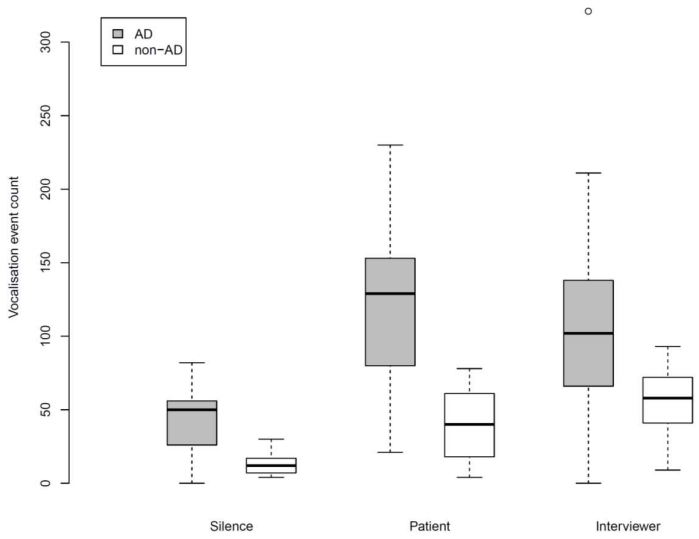


Figure 11 Distribution of vocalisation event counts for patients with and without AD in CCC dialogues.

Vocalisation and pause patterns have been successfully employed in the analysis of dialogues in a mental-health context [60], segmentation [61] and classification of dialogues, and more recently on characterisation of participant role and performance in collaborative tasks [58]. Models that employ basic turn-taking statistics have also been proposed for dementia diagnosis [62], though not in a systematic content-fee framework as in our proposed method. The distribution of vocalisation event counts is shown in Figure 11. It can be observed that patients with AD tend to produce more vocalisation events than their interviewers (and, consequently, produce more silence events). This is consistent with findings in the literature on language changes in AD [1].

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 769661

Table 8 Descriptive statistics on dialogue turn-taking (duration given in seconds)

Feature	non-AD	AD
Dialogue duration	4107.3	7628.4
Dialogue duration TTS	7618.8	7618.8
Avg turn duration	97.3	255.8
Total turn duration	1654.3	4348.3
Norm. total turn duration	3.0	4.1
Avg turn duration TTS	107.6	238.0
Total turn duration TTS	1829.7	4046.1
Norm. total turn duration TTS	3.0	4.2
Avg number of words	314.6	742.5
Total number of words	5348.0	12622.0
Avg words per minute	155.9	166.5

Speech rate was estimated using De Jong's syllable nuclei detection algorithm [59], which is an unsupervised method – that is, it can be applied directly to the acoustic signal, with no need of human annotation. However, as the audio quality of the CCC recordings is uneven, and as the dataset provides no gold standard against which one could assess syllable count, we decided to validate the use of De Jong's method against the time-stamped transcripts provided. Using these transcripts one could, in principle, estimate average words per minute (WPM) for individual utterances, as is sometimes done [63]. However, this method of measuring WPM based on transcription has a number of limitations. Words have variable length, and their articulation can vary greatly due to a number of speech-related phenomena, such as phonological stress, frequency, contextual predictability, and repetition [64]. In or- der to mitigate these problems, we instead produced speech rate ratio estimates normalised through a speech synthesizer, employing the methods proposed by Hayakawa et al. [63]. These estimates represent deviations from a "normalised" pace of 160 words per minute (WPM) synthesised using the MaryTTS system [65]. We therefore computed the ratio of the synthesised speech to the actual duration of the patient's speech. The speech rate ratio correlated well with the syllable per minute rate extracted using only the recorded audio ($\rho = 0.502$, $t(30) = 3.19$, $p = 0.003$) indicating that speech rate can be estimated with an acceptable level of reliability through the unsupervised method, even in fairly noisy settings.

$$\hat{F}(x) = sign\left[\sum_{m=1}^{M} \hat{f}_m(x)\right]$$

A Python script was employed to extract basic speaker turn time stamps, speaker role information, and transcriptions from the original XML-encoded CCC data. The resulting data were then processed using the R language in order to detect silence intervals, and categories turn transitions

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

and pause events.

Some descriptive statistics on the dialogues can be seen in Table 1. These statistics include: average turn duration (how many seconds a participant speaks on average), total turn duration (how many seconds did the participant's turns lasted in total), normalised turn duration (the ratio of a participant's turn duration to the total duration of AD or non-AD dialogues, according the participant's class), number of words generated (total per class and on average per class' participant), and number of words per minute (aver- age per class participant).

Contrary to our expectations, we did not observe a statistically significant difference between the speech rate in syllables per minute between patients with and without AD (Welch two sample t-test t(30.5) = 1.15, p = 0.28), even though the mean for non-AD (M = 180.8 syllables/min, sd = 28.4) was higher than that for patients with AD (M = 168 syllables/min, sd = 35.6).

Two alternative data representations were generated. The first (henceforth referred to as VGO) was based on the vocalisation graphs only. That is, VGO encodes the probabilities of each possible pair of transitions, including self- transitions, which tend to dominate Markov chains sampled, and the steady-state probabilities for each vocalisation event. The second form of representation (VGS) sim- ply consists of the VGO with information about the participant's speech rate (mean and variance) added to the vocalisation statistics. With the exception of speech rate ratio, which necessitates transcription, all the information needed to build VGO and VGS instances can be extracted through straightforward signal processing methods.

3.5.2 Predictive Modelling

The data instances in the two alternative representation schemes were annotated for presence or absence of Alzheimer's disease (AD). A supervised learning procedure was employed in order to train classifiers to predict such annotations on unseen data.

We trained a boosting model [66] using decision stumps (i.e. decision trees with a single split node) as weak learners. The training process consisted of 10 iterations whereby, for each training instance (xi), a weak classifier f_m was fitted using weights on the data which were iteratively computed so that the instances misclassified in the preceding step had their weights increased by a factor proportional to the weighted training error. In this case class probability estimates P (ad = 1jdata) were used to compute these weights and to weigh the final classification decision (additive logistic regression) following the Real Adaboost algorithm [37]: Classification performance was assessed through a 10-fold cross validation procedure. As the dataset is reasonably balanced, results were assessed in terms of accuracy, precision (the ratio of the number of true positives to the number of instances classified as AD), recall (or sensitivity, the ratio of true positives to the number of AD cases) and F1 score (the harmonic mean of precision and recall). Micro (μ) and macro (M) averages for these scores are given by taking means over the entire set of classification decisions and over individual classifiers respectively, across the 10 folds. As the data set is fairly small, we also ran a leave-one- out cross validation (LOOCV) procedure to obtain better estimates of generalisation accuracy. This consisted of selecting one instance for testing, and building a classification model on the remaining instances, and iterating this procedure until all instances have been selected as

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

testing instances. Macro averages are uninformative in LOOCV, so we only report overall accuracy figures for this procedure. ROC curves showing the relationship between true positive and false positive rates as the classification threshold is varied were also plotted. Simulation was employed in order to smooth these ROC curves by running 10 rounds of 10- fold cross validation tests with a randomised selection of instances making up the hold-out sets.

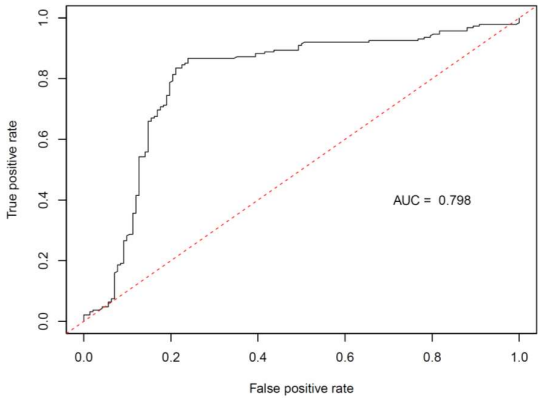
3.5.3 Results

Our first approach, based on the VGO data representation scheme, produced promising results. Accuracy levels were well above the baseline, with overall accuracy reaching 81.1%, showing that turn taking patterns can provide useful cues to the detection of AD in dialogues. The results for the VGO-based classification are shown in Table 9. The corresponding ROC curve is shown in Figure 12.

Table 9 AD detection results for the VGO data representation scheme

AD		non-AD	
Accuracy_{μ}	0.812	Accuracy_{μ}	0.714
Precision_{μ}	0.765	Precision_{μ}	0.769
Recall_{μ}	0.812	Recall_{μ}	0.714
$F_{1,\mu}$	0.788	$F_{1,\mu}$	0.741
Precision_{M}	0.667	Precision_{M}	0.792
Recall_{M}	0.722	Recall_{M}	0.729
$F_{1,M}$	0.685	$F_{1,M}$	0.721
Overall accuracy (LOOCV): 0.811			

Figure 12 ROC curve for VGO-based classifiers.

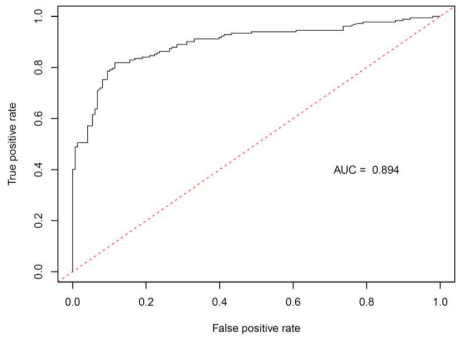


Adding speech rate information (VGS representation) contributed to further enhancing AD detection, bringing the overall accuracy score to about 86.5%. Detailed evaluation metrics are shown in Table 3. The ROC curve for the VGS-based classification approach is shown in Figure 13.

Table 10 results for the VGS data representation scheme

AD		non-AD	
Accuracy_{μ}	0.882	Accuracy_{μ}	0.769
Precision_{μ}	0.833	Precision_{μ}	0.833
Recall_{μ}	0.882	Recall_{μ}	0.769
$F_{1,\mu}$	0.857	$F_{1,\mu}$	0.800
Precision_{M}	0.796	Precision_{M}	0.708
Recall_{M}	0.833	Recall_{M}	0.708
$F_{1,M}$	0.811	$F_{1,M}$	0.700
Overall accuracy (LOOCV): 0.865			

Figure 13 ROC curve for VGS-based classifiers.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

The addition of features for mean and variance of speech rate ratio over dialogues had the effect of improving classification trade-offs, particularly reducing the false positives while increasing the true positives at low threshold cut-offs. For comparison we ran the same testing procedure using some of the other classifiers employed in the literature, namely, logistic regression, naïve Bayes (Gaussian kernel), decision trees (C4.5 algorithm), SVM trained using sequential minimal optimisation, with a polynomial kernel [67], and random forests [68], Weka implementation [69]. The overall (LOOCV) accuracy figures are shown in Table 11. There is little difference in performance between our chosen method (Real Adaboost) and other methods used in the literature, except for logistic regression, which underperforms the machine learning methods. Real Adaboost slightly outperforms SVM and random forests classifiers, and matches C4.5 decision trees, with a slight advantage over the latter on the target AD class (Fm = 0:878 vs. Fm = 0:872). Although there is considerable room for improvement upon this level of classification performance, the levels obtained with these simple models are comparable to the accuracy of approaches that employ more detailed linguistic information, which are presumably harder to acquire in everyday conversational situations, as they would involve a level of speech recognition accuracy which is beyond the capabilities of current systems for spontaneous speech in noisy environments.

Table 11 Compared accuracy results obtained with different classification algorithms, on VGS-based datasets

Classification method	Accuracy (LOOCV)
Logistic regression	75.7%
Real Adaboost	86.5%
Decision trees	86.5%
SVM	83.7%
Random forests	81.1%

4. CONCLUSION

Dementia prevention and life quality in elderly care are important societal challenges. Automatic detection of signs of AD in speech can provide useful tools for the design of technologies for care-giving and cognitive health monitoring to help address these challenges. In the framework of SAAM projects the proposed models will help in predicting the cognitive decline through speech information and helps the SAAM system in suggesting an intervention.

This report demonstrates the relevance of acoustic features for cognitive impairment detection in the context of Alzheimer's diagnosis. Machine learning methods operating on automatically extracted voice features provide accuracy of up to 78.7 %, well above baseline and comparable to other results reported in the literature obtained with manually designed feature sets. Although

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

memory impairment might be the definitive symptom for AD diagnosis, studies have shown that language becomes abnormal relatively early in AD [54] and can serve as a sensitive index of disease severity over time. The outcomes of our study could be used to develop an intervention for detecting early dementia symptoms and mitigating disease progression. This could be achieved through further establishing a sensitive index of speech function and response to cognitive intervention. In future, we aim to extend the research presented here by incorporating different acoustic features for prediction of Mini-Mental State Examination scores. We are also in the process of collecting an extended dataset of spontaneous dialogue data from healthy people in mid-life [55] who are at risk of developing AD due to genetic, clinical and family history factors, and intend to employ the methods introduced in this section to these data to investigate possible 'voice biomarkers' of risk.

This report also presented initial results of a new method to automatically recognise the first signs of disrupted communication using dialogue features. This method obtained an overall accuracy of 0.83, with a micro F-measure of 0.83 and a macro F-measure of 0.76 on the classification of patients as 'AD' and 'non-AD'. Although it is difficult to compare these results directly to related works [27, 70], our accuracy figures are situated within a similar range, 0.70-0.80, with a smaller discrepancy between the classification of the two groups, while relying on features that can be more robustly extracted from spontaneous speech. Thanks to the increasingly important role of social technology, longitudinal studies may become richer in terms of the amount of variables measured, frequency of measurements and places where measures are taken (living settings), allowing for larger datasets. As more data are gathered in natural settings, we expect to obtain more reliable and generalisable results. There are several linguistic parameters that are promising for the assessment of cognitive functioning. In current approaches, these features have been typically extracted from data collected through structured interviews, storytelling or picture descriptions. The work presented here contributes a new perspective to feature extraction by focusing on spontaneous dialogues. Dialogue processing provides a convenient framework for the analysis of natural conversations, in which readily available predictors, such as turn taking behaviour, have already yielded satisfactory results. We plan to further analyse verbal and non-verbal parameters to obtain a better characterisations of AD in order to infer neurosychological assessment results through speech and language processing, and subsequently to combine such features with actual neuropsychological evaluations and other relevant variables, building accurate models to achieve detection of AD at the time of onset. The data set used in the present study has some limitations. Due to its constraints, the study was performed on a restricted subset of 21+17 sessions. In addition, the interview setting includes a degree of bias, as the interviewer's objective is to get the patient to perform a certain task (e.g. description of a picture, driving the discussion) therefore influencing the patient's speech. To mitigate these limitations, we plan to collect further data in more spontaneous dialogue in the near future.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661

5. REFERENCES

[1] American Psychiatric Association, “Delirium, dementia, and amnestic and other cognitive disorders,” in Diagnostic and Statistical Manual of Mental Disorders, Text Revision (DSM-IV-TR), fourth edi ed., American Psychiatric Association, Ed., Arlington, VA, 2000, ch. 2.

[2] A. Burns and S. Iliffe, “Dementia,” BMJ, vol. 338, 2009.

[3] World Health Organization, “Mental Health Action Plan 2013-2020,” WHO Library Cataloguing-in-Publication Data Library Cataloguing-in-Publication Data, pp. 1–44, 2013.

[4] W. H. Organization et al., “First who ministerial conference on global action against dementia: meeting report, who headquarters, Geneva Switzerland, 16-17 march 2015,” 2015.

[5] R. B. Maccioni, G. Farıas, I. Morales, and L. Navarrete, “The revitalized tau hypothesis on Alzheimer’s disease,” Archives of medical research, vol. 41, no. 3, pp. 226–231, 2010.

[6] J. Hardy and D. J. Selkoe, “The amyloid hypothesis of Alzheimer’s disease: progress and problems on the road to therapeutics,” science, vol. 297, no. 5580, pp. 353–356, 2002.

[7] H. Braak and E. Braak, “Evolution of the neuropathology of Alzheimer’s disease,” Acta Neurologica Scandinavica, vol. 94, no. S165, pp. 3–12, 1996.

[8] S. Norton, F. E. Matthews, D. E. Barnes, K. Yaffe, and C. Brayne, “Potential for primary prevention of Alzheimer’s disease: an analysis of population-based data,” The Lancet Neurology, vol. 13, no. 8, pp. 788–794, 2014.

[9] J. L. Cummings and D. F. Benson, “Dementia: a clinical approach (2nd edition),” International Journal of Geriatric Psychiatry, vol. 7, no. 12, 1992.

[10] G. W. Ross, J. L. Cummings, and D. F. Benson, “Speech and language alterations in dementia syndromes: Characteristics and treatment,” Aphasiology, vol. 4, no. 4, pp. 339–352, 1990.

[11] H. S. Kirshner, “Primary Progressive Aphasia and Alzheimer’s disease: Brief History, Recent Evidence,” Current Neurology and Neuroscience Reports, vol. 12, no. 6, pp. 709–714, 2012.

[12] M. F. Mendez, J. L. Cummings, and J. L. Cummings, Dementia: a clinical approach (3rd edition). Butterworth-Heinemann, 2003.

[13] M. W. Bondi, D. P. Salmon, and A. W. Kaszniak, “The neuropsychology of dementia.” in Neuropsychological assessment of neuropsychiatric disorders, 2nd ed. New York, NY, US: Oxford University Press, 1996,
pp. 164–199.

[14] P. N. Dawadi, D. J. Cook, and M. Schmitter-Edgecombe, “Automated cognitive health assessment using smart home monitoring of complex tasks,” IEEE transactions on systems, man, and cybernetics: systems,

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 769661

vol. 43, no. 6, pp. 1302–1313, 2013.

[15] A. J. Braaten, T. D. Parsons, R. Mccue, A. Sellers, and W. J. Burns, “Neurocognitive Differential Diagnosis Of Dementing Diseases: Alzheimer’s Dementia, Vascular Dementia, Frontotemporal Dementia, And Major Depressive Disorder,” International Journal of Neuroscience, vol. 116, no. 11, pp. 1271–1293, 2006.

[16] 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. 189–98, 1975.

[17] P. Robert, S. Schuck, B. Dubois, J. Lepine, T. Gallarda, J. Oli´e, S. Goni, and S. Troy, “Validation of the Short Cognitive Battery (B2C). Value in screening for Alzheimer’s disease and depressive disorders in psychiatric practice,” Encephale, vol. 29, no. 3 Pt 1, pp. 266–72.

[18] B. Dubois, A. Slachevsky, I. Litvan, and B. Pillon, “The FAB: a Frontal Assessment Battery at bedside.” Neurology, vol. 55, no. 11, pp. 1621–6, 2000.

[19] P. S. Mathuranath, A. George, P. J. Cherian, R. Mathew, and P. S. Sarma, “Instrumental activities of daily living scale for dementia screening in elderly people.” International psychogeriatrics, vol. 17, no. 3, pp. 461–74, 2005.

[20] 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.

[21] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert et al., “Automatic speech analysis for the assessment of patients with predementia and alzheimer’s disease,” Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring, vol. 1, no. 1, pp. 112–124, 2015.

[22] A. König, N. Linz, J. Tröger, M. Wolters, J. Alexandersson, and P. Robert, “Fully automatic speech-based analysis of the semantic verbal fluency task,” Dementia and Geriatric Cognitive Disorders, vol. 45, no. 3-4, pp. 198–209, 2018.

[23] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, “The Natural History of Alzheimer’s Disease,” Archives of Neurology, vol. 51, no. 6, p. 585, 1994.

[24] H. Goodglass and E. Kaplan, “The assessment of aphasia and related disorders,” Philadelphia, 1983.

[25] H. Goodglass, E. Kaplan, and B. Barresi, The assessment of aphasia and related disorders. Lippincott Williams & Wilkins, 2001.

[26] H. Goodglass, Boston diagnostic aphasia examination: Short form record booklet. Lippincott Williams & Wilkins, 2000.

[27] 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, 2016.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 769661

[28] S. Luz, “Longitudinal monitoring and detection of alzheimer’s type dementia from spontaneous speech data,” in Computer-Based Medical Systems (CBMS), 2017 IEEE 30th International Symposium on. IEEE, 2017, pp. 45–46.

[29] S. Luz and S. D. la Fuente, “A method for analysis of patient speech in dialogue for dementia detection,” in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), D. Kokkinakis, Ed. Paris, France: European Language Resources Association (ELRA), may 2018.

[30] L. Toth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatloczki, Z. Banreti, M. P´ak´aski, and J. Kalman, “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.

[31] K. Lopez-de Ipina, M. Faundez-Zanuy, J. Sole-Casals, F. Zelarin, and P. Calvo, “Multi-class Versus One-Class Classifier in Spontaneous Speech Analysis Oriented to Alzheimer Disease Diagnosis,” in Recent Advances in Nonlinear Speech Processing, E. et Al., Ed. Springer International Publishing, 2016, vol. 48, pp. 63–72.

[32] J. Reilly, A. D. Rodriguez, M. Lamy, and J. Neils-Strunjas, “Cognition, language, and clinical pathological features of non Alzheimer’s dementias: an overview,” Journal of communication disorders, vol. 43, no. 5, pp. 438–452, 2010.

[33] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, “Spoken language derived measures for detecting mild cognitive impairment,” IEEE transactions on audio, speech, and language processing, vol. 19, no. 7, pp. 2081–2090, 2011.

[34] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, “Aided diagnosis of dementia type through computer-based analysis of spontaneous speech,” in Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2014, pp. 27–37.

[35] S. Luz, S. Fuente, and P. Albert, “A method for analysis of patient speech in dialogue for dementia detection,” in Proceedings of the LREC 2018 Workshop, pp. 35–42.

[36] C. Pope and B. H. Davis, “Finding a balance: The carolinas conversation collection,” Corpus Linguistics and Linguistic Theory, vol. 7, no. 1, pp. 143–161, 2011.

[37] J. Friedman, T. Hastie, R. Tibshirani, and Others, “Additive logistic regression: a statistical view of boosting,” The annals of statistics, vol. 28, no. 2, pp. 337–407, 2000.

[38] L. Toth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatloczki, Z. Banreti, M. Pakaski, and J. Kalman, “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.

[39] A. Satt, R. Hoory, A. König, P. Aalten, P. H. Robert, N. Sophia, C. Memoire, D. Ressources, and C. H. U. D. Nice, “Speech- Based Automatic and Robust Detection of Very Early Dementia,” in Interspeech, Singapore, 2014, pp. 2538–2542.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 769661

[40] A. Satt, A. Sorin, O. Toledo-Ronen, O. Barkan, I. Kompatsiaris, A. Kokonozi, and M. Tsolaki, “Evaluation of speech-based protocol for detection of early-stage dementia.” in INTERSPEECH, 2013, pp. 1692–1696.

[41] L. Hernandez-Domınguez, S. Ratte, G. Sierra-Martınez, and A. Roche-Bergua, “Computer-based evaluation of Alzheimer’s disease and mild cognitive impairment patients during a picture description task,” Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring,
vol. 10, pp. 260–268, 2018.

[42] A. Duong, F. Giroux, A. Tardif, and B. Ska, “The heterogeneity of picture-supported narratives in Alzheimer’s disease,” Brain and language, vol. 93, no. 2, pp. 173–184, 2005.

[43] 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.

[44] J. R. Hodges and K. Patterson, “Is semantic memory consistently impaired early in the course of Alzheimer’s disease? Neuroanatomical and diagnostic implications,” Neuropsychologia, vol. 33, no. 4, pp. 441–459, 1995.

[45] J. Corey Bloom and A. Fleisher, “The natural history of Alzheimer’s disease,” Dementia, vol. 34, pp. 405–15, 2000.

[46] J. Chen, Y. Wang, and D. Wang, “A feature study for classification-based speech separation at low signal-to-noise ratios,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1993–2002, 2014.

[47] J. Kim and M. Hahn, “Voice activity detection using an adaptive context attention model,” IEEE Signal Processing Letters, 2018.

[48] F. Haider and S. Luz, “Attitude recognition using multi-resolution cochleagram features,” in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2019, pp. 3737–3741.

[49] D. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” in Speech separation by humans and machines. Springer, 2005, pp. 181–197.

[50] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the Munich versatile and fast open-source audio feature extractor,” in Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010, pp. 1459–1462.

[51] 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. ACM, 2013, pp. 835–838.

[52] F. Eyben, K. R. Scherer, B.W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan 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.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 769661

[53] T. Kohonen, “The self-organizing map,” Neurocomputing, vol. 21, no. 1-3, pp. 1–6, 1998.

[54] 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. 356–61, 2002.

[55] S. de la Fuente, C. 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, 2019.

[56] M. Mortamais, J. A. Ash et al., “Detecting cognitive changes in preclinical Alzheimer’s disease: A review of its feasibility,” Alzheimer’s & Dementia, in press.

[57] S. Luz, “Automatic identification of experts and performance prediction in the multimodal math data corpus through analysis of speech interaction,” in Procs. of the International Conference on Multimodal Interaction. 2013 Dec 9 (pp. 575-582). ACM.

[58] S. Luz, “The non-verbal structure of patient case discussions in multidisciplinary medical team meetings,” ACM Transactions on Information Systems, vol. 30, no. 3, pp. 17:1–17:24, 2012.

[59] N. H. d. 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, May 2009.

[60] Jaffe, J. and Feldstein, S. (1970). Rhythms of dialogue. Personality and Psychopathology. Academic Press, New York.

[61] Luz, S. and Su, J. (2010). The relevance of timing, pauses and overlaps in dialogues: Detecting topic changes in scenario based meetings. In Proceedings of INTER- SPEECH 2010, pages 1369–1372, Chiba, Japan. ISCA

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

[63] Hayakawa, A., Vogel, C., Luz, S., and Campbell, N. (2017). Speech rate comparison when talking to a sys- tem and talking to a human: A study from a speech-to- speech, machine translation mediated map task. In Proc. Interspeech 2017, pages 3286–3290.

[64] Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. 60(1):92–111.

[65] Schröder, M. and Trouvain, J. (2003). The German text-to- speech synthesis system MARY: A tool for research, development and teaching. International Journal of Speech Technology, 6(4):365–377.

[66] Schapire, R. E. and Freund, Y. (2014). Boosting: Foundations and Algorithms. The MIT Press,

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 769661

January.

[67] Platt, J. (1998). Fast training of support vector ma- chines using sequential minimal optimization. In B. Schoelkopf, et al., editors, Advances in Kernel Meth- ods - Support Vector Learning. MIT Press.

[68] Breiman, L. (2001). Random forests. Machine Learning, 45(1):5–32.

[69] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reute- mann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1):10–18.

[70] Guinn, C. I. and Habash, A. (2012). Language analysis of speakers with dementia of the Alzheimer's type. In AAAI Fall Symposium: Artificial Intelligence for Gerontechnology, pages 8–13.

[71] J. Dukart, M. L. Schroeter, K. Mueller, A. D. N. Initiative, et al., Age correction in dementia matching to a healthy brain, PloS one 6 (7) (2011) e22193.

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 769661