An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech

An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech

Fasih Haider, *Member, IEEE*, Sofia de la Fuente and Saturnino Luz, *Member, IEEE*

***Abstract*—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 acoustic (speech) 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. We present a detailed study of the predictive value of purely acoustic features automatically extracted from spontaneous speech for Alzheimer's dementia detection, from a computational paralinguistics perspective. The effectiveness of several state-of-the-art paralinguistic feature sets for Alzheimer's detection were assessed on a balanced sample of DementiaBank's Pitt spontaneous speech dataset, with patients matched by gender and age. The feature sets assessed were the extended Geneva minimalistic acoustic parameter set (eGeMAPS), the *emobase* feature set, the *ComParE 2013* feature set, and new Multi-Resolution Cochleagram (MRCG) features. Furthermore, we introduce a new active data representation (ADR) method for feature extraction in Alzheimer's dementia recognition. Results show that classification models based solely on acoustic speech features extracted through our ADR method can achieve accuracy levels comparable to those achieved by models that employ higher-level language features. Analysis of the results suggests that all feature sets contribute information not captured by other feature sets. We show that while the *eGeMAPS* feature set provides slightly better accuracy than other feature sets individually (71.34%), "hard fusion" of feature sets improves accuracy to 78.70%.**

***Index Terms*—Affective Computing, Social Signal Processing, Dementia, Alzheimer, Cognitive Decline Detection, Cognitive Impairment Detection**

I. Introduction

DEMENTIA is a category of neurodegenerative diseases that entails a 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, reducing the patient's autonomy and wellbeing, as well as 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% of cases, including those that also manifest AD), Lewy body disease (15% cases), and other brain diseases (5%), including Parkinson's, frontotemporal dementia and stroke [2].

The main risk factor for dementia is age, and therefore its greatest incidence is amongst the elderly. As 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 per one thousand people, where 5 to 8 cases would be caused by Alzheimers 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 considerably on dementia prevention and early detection, focusing on disease progression. 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]. Such neuropathology is known to start silently up to 20 years before an individual shows obvious and observable cognitive symptoms, and there is no satisfactory treatment for them. 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 study focuses on AD recognition using acoustic 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 naming and word-finding difficulties (anomia) leading to circumlocution, as well as difficulty accessing semantic information intentionally, leading to a general semantic deterioration [12]. The heterogeneity of the symptomatic expression of AD requires

F. Haider, S.D.L Fuente and S. Luz are with the Usher Institute of Population Health Sciences and Informatics, Edinburgh Medical School, the University of Edinburgh.

Financial support for this research comes from the European Union's Horizon 2020 research and innovation programme, under the grant agreement No 769661, towards the SAAM project; and from the Medical Research Council (MRC; grant No. MR/N013166/1).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. The material includes a flow diagram of the ADR algorithm. Contact s.luz@ed.ac.uk for further questions about this work.

Manuscript received Aug 20, 2019; revised Oct 20, 2019.

diagnosis support methods that are able to capture more subtler aspects than conventional screening tools, which often fail to discriminate these symptoms in preclinical AD. Social signal processing technologies are creating opportunities for personal health monitoring and development of diagnostic support tools based on automated processing of behavioural signals [13]. Speech and language are rich and ubiquitous sources of cognitive behavioural data, where computational analysis has the potential to aid clinicians in early and accurate diagnosis of dementia [14].

Several commonly used cognitive tests for dementia diagnosis involve linguistic assessment, These include the Mini-Mental State Examination (MMSE) [15], the five-word test [16], the frontal assessment battery [17], and the instrumental activities of daily living scale [18]. Speech continuity, for instance, may be assessed through picture description tasks [19] or through countdown tasks [20], and Semantic Verbal Fluency (SVF) usually involves naming tasks [21]. However, whilst still valuable for diagnosis, most of these neuropsychological tests offer little insight into early stages of neurodegeneration. Hence there is an increasing interest in developing alternative methods for early detection. A recent study [20] on sentence repetition data employed dynamic time warning to evaluate the wave forms, assessing alignment curve between pairs of corresponding wave forms to see whether there is a significant difference between the sentences produced by the clinician and the sentences produced by the AD patients [20].

A disadvantage of these tests is that they employ speech and language generated under controlled laboratory conditions rather than spontaneously, which would be required for practical longitudinal screening and monitoring in daily life. One of the few currently available spontaneous speech datasets linked to clinical neuropsychological assessments for dementia is the picture description task gathered by the Alzheimer and Related Dementias Study at the University of Pittsburgh School of Medicine, often referred to as the "Pitt dataset" [22][1].

The Pitt dataset consists of speech from participants who were recorded while performing the Boston Cookie Theft picture description task, from the Boston diagnostic aphasia examination [23], [24], [25]. A variety of computational methods have been employed on this corpus for detection of Alzheimer's Disease and mild cognitive impairment across different studies (more details in section 2). Most of these works focused on linguistic features [26], [27], [28], [29], taking advantage of the manual transcriptions available with the speech data. While paralinguistic features have so far received less attention, there are good reasons for investigating a paralinguistics approach to AD. Some of these reasons are methodological, such as avoiding the need for transcriptions, and some are related to the nature of the disease, such as the fact that prosodic analysis may lead to detection of motor subtleties in speech production, in addition to subtle linguistic decline. Fraser et al. [29] carried out additional prosodic analysis on the Pitt recordings, extracting 42 mel-frequency cepstral coefficient (MFCC) features [30]. Others have used a similar approach [31], [32]. Another recent study successfully used these recordings to extract low-level speech features (vocalisation events and speech rate) and used them to train a system for AD detection [33]. All these studies use the Pitt corpus, and the majority rely on manual speech transcripts; only [33] relies exclusively on acoustic features. Furthermore, these previous studies did not adjust for age and gender imbalances or the effects of variable audio quality in the data, and employed *ad hoc* paralinguistic feature sets.

[1]Data available at DementiaBank, http://dementia.talkbank.org/

The work presented in this paper addresses these issues by evaluating a comprehensive set of acoustic features which are emerging in the field of computational paralinguistics [34], on a gender- and age-balanced subset of the Pitt corpus, which has been preprocessed to ensure consistent audio quality. It contributes to research in AD detection by:

- evaluating the potential of several feature sets designed for different computational paralinguistics tasks (eGeMAPS [35], emobase [36] and ComParE [37]) along with a recently proposed MRCG derived feature set [38], for AD detection. This is, to the best of our knowledge, the first empirical attempt to use these feature sets as "digital biomarkers" for Alzheimer's disease. We
  1) demonstrate the discriminative power of these feature sets, and their fusion, for automatic recognition of AD
  2) test these features using different machine learning methods to implement automatic classification of patients with and without AD.
- presenting and evaluating a novel method (ADR) of representing these acoustic features, and
- creating an enhanced version of the Pitt corpus which is balanced and acoustically preprocessed.

II. Background

The complex multimodal ways in which AD symptoms may appear calls for increasingly interdisciplinary research. Current work on AD symptomatology combines signal processing, artificial intelligence, cognitive psychology, computational linguistics, medicine, neuropsychology and computer science, among other disciplines. Although linguistics research on AD has focused on formal aspects of language (i.e. lexicon, syntax, semantics), the analysis of continuous speech has been progressively seen by Alzheimer's researchers as a source of information that may support diagnosis of MCI, AD and related conditions [39], [29], [33], [40], [41], [42].

Language research into AD has employed high-level features such as information content, comprehension of complexity, picture naming and word-list generation, as predictors of disease progression [43]. A study by Roark et al. [44] used natural language processing (NLP) and automatic speech recognition (ASR) to automatically annotate and time-align a few spoken language features (pause frequency and duration), also comparing them to manually annotated counterparts. They analysed audio recordings of 74 neuropsychological assessments to classify MCI and healthy elderly participants. Their best SVM classifier obtained an AUC of 0.86 by including a combination of automated speech and language features and cognitive tests scores. Jarrold et al. [45] worked with a dataset consisting of semi-structured interviews from

9 healthy participants, 9 with AD, 9 with frontotemporal dementia, 13 with semantic dementia, and 8 with progressive nonfluent aphasia. With an ASR system, they extracted 41 features, including speech rate, and the mean and standard deviation of the duration of pauses, vowels, and consonants. They used a multilayered perceptron network, achieving a classifier accuracy of 88% for AD vs. healthy subjects based on lexical and acoustic features. A more recent study by Luz et al. [40] extracted graph-based features encoding turn-taking patterns and speech rate [46] from the Carolina Conversations Collection [47] (spontaneous interviews of participants with and without an AD). They used these features to create an additive logistic regression model that obtained 85% accuracy in distinguishing dialogues involving an AD speaker.

These studies combine signal processing and machine learning to detect subtle acoustic signs of neurodegeneration which may be imperceptible to human diagnosticians. Tóth et al. [39], for instance, found that filled pauses (sounds like "hmmm", etc) could not be reliably detected by human annotators, whereas detection improved by using an ASR system. This study analysed the recorded speech of 38 healthy controls and 48 patients with MCI speaking about two short films, extracting several acoustic features (hesitation ratio, speech tempo, length and number of silent and filled pauses, length of utterance). They reported that ASR-extracted features performed best in combination with machine learning methods, particularly with a Random forest classifier (75% accuracy), outperforming manually calculated features (69.1% accuracy). Similar machine learning methods were used by König et al. [20], 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. However, their tests were performed on non-spontaneous speech data gathered under controlled conditions, as part of a neuropsychological assessment that included manually transcribed text.

Satt et al. [48] reported accuracy levels above 80% for different SVM classifiers when distinguishing 89 AD, MCI and control participants who had performed 4 different spoken tasks. Several features were extracted with an ASR system, such as global statistics of the segments, temporal structure of the speech/voice, real cepstrum coefficients, irregularity/errors in pronunciation, response time, speech rate, correctness, and pause patterns. This analysis was carried on a Greek language dataset, and the same research group reported similarly promising results on a French dataset later on [49]. This supports our view that the acoustic-prosodic approach generalises across languages.

Studies in this field continually evidence the heterogeneity with which language and speech impairments are displayed in AD and related diseases. Duong et al. [50] 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 ageing and AD. This heterogeneity seems to be more evident in AD than in specific language diseases such as primary progressive aphasia [51], especially in early stages of AD [52]. Therefore we hypothesise that a comprehensive analysis of state-of-the-art paralinguistic feature sets which have been successfully used in different prediction tasks may help identify such patterns and enhance accuracy of early AD detection.

Although there is a research trend on collection of spontaneous speech data, as opposed to speech elicited through lab-based tasks, the Pitt Corpus remains one of the very few available datasets coupling relatively spontaneous speech (recordings and transcriptions) with clinical information. Hence, this dataset has been used in several studies. One of the best known such studies is [29], which obtained 81.92% accuracy for machine learning classification of individuals with and without AD. A variety of features were employed, identifying four factors: semantic impairment, acoustic abnormality, syntactic impairment and information impairment. In addition to a range of high-level linguistic features, this study employed a basic set of acoustic features, namely, mean, variance, skewness, and kurtosis of the first 42 MFCCs.

Similarly, [31] used a Random Forest classifier to detect AD in the Pitt dataset, achieving 80% accuracy. They developed a vector-space method for automatic topic modelling based on manual transcripts to train this classifier. However, the aforementioned accuracy is only achieved when they add the same lexicosyntactic and acoustic features described by Fraser et al.'s [29] in their topic model.

Along the same lines, the work of Hernandez-Gmez et al. [32] used information coverage measures, linguistic features and acoustic features for automatic classification of dementia. Their best binary model (AD-nonAD) is an SVM classifier which achieved 79% accuracy. Following [29], the acoustic features extracted consisted of the mean, kurtosis, skewness and variance of MFCCs, although they only estimated these for the first 13 MFCC values. After introducing the MCI group, conjoint it with the AD group in the binary classification, the best performance with the same feature set dropped to 77% with a Random Forest classifier [32].

A slightly different approach was adopted by Orimaye et al. [27] who used a deep neural network for the classification task. They advocate for high order n-grams and deeper vocabulary spaces, and report an AUC of 0.79 for 4-grams, and 0.83 for 5-grams, and even 0.94 for 1000-grams features in another study [26]. However, n-grams require expected information units to be operationalised (previously), which is a process influenced by context and subject to variability [53]. The higher the n-gram order, the more these n-grams become tied to the task content. This makes the method hard to generalise as a screening tool, unless speech is collected using this same Cookie-Theft task. To address this problem, a method has been proposed [28] which generates information units across two different languages. SVM models are trained with these features extracted from the Pitt corpus descriptions in English and Swedish to classify MCI and healthy elderly controls. Classification accuracy varied across languages, achieving 72% accuracy in Swedish, and 63% in English. Although these results do not match the performance of previous work [29] in terms of classification accuracy, it is worth mentioning this study because its multilingual approach attempts to palliate an important challenge for clinical language analysis, namely, the predominance of English language data in research and the difficulties to generalise these tools to less frequently

researched languages. In this regard, we consider this to be a priority for global public health, and argue that a model based solely on acoustic analysis, such as the present study, might make multilingual generalisation more straightforward.

Lastly, the study by Luz et al. [33] shows that a simple Bayesian classifier can achieve 68% accuracy classifying Alzheimer's patients and elderly controls without relying on transcribed content. As far as we know, this is the only study on this dataset to exclusively employ speech data for the analysis - without also including the available transcriptions. The Bayesian classifier was trained in low-level acoustic features directly extracted from the recordings, but no extensive use of paralinguistic feature sets was attempted.

The general picture of speech research aiming at dementia detection is heterogeneous and comparisons are difficult to draw. One of the main reasons is not the multiplicity of datasets, some of which have been mentioned, but rather the diversity of collection methods (i.e. semi-structured interviews, recording of neuropsychological testing, picture descriptions, spontaneous conversations), acoustic quality, content, length, experimenter and participant expectations, etc. Even the cited studies that use the same dataset - the Pitt Cookie Theft Corpus - do not use the full dataset, but different samples of it. For instance, [29] used 233 control samples and 240 AD samples, whereas [26] selected a subset consisting of 99 patients with probable AD and 99 healthy controls.

From a speech processing perspective, another difficulty is that most studies combine acoustic features with high-level language features which can only be extracted reliably from transcription, making it difficult to assess the extent to which classification performance can be obtained by fully automated means. Notwithstanding, several studies have successfully carried fully automatic transcriptions to detect dementia from speech. There have been attempts to quantify the potential effects of ASR errors on classification performance. Weiner et al. [54] compared AD detection based on manually and automatically generated transcripts. They obtained nearly identical results with an automatic transcription (unweighted average recall, UAR = 0.623) as with a manual one (UAR = 0.606), using the ISLE dataset, which consists of biographic interview and cognitive diagnoses of 74 German participants. Another study [55] evaluated the effectiveness of ASR transcripts as a source of input features on four different spoken language datasets ([22], [56], [57], [58]), which had available but erroneous transcriptions. This is, to the best of our knowledge, the only study to employ linguistic features extracted from the Pitt dataset through ASR. It reported 58.4% accuracy using ASR-generated transcripts, compared to 69.8% accuracy when using manual transcription. Although these results are not to be overlooked, we propose a model which, not relying on transcription, is free from the constraints inherent to ASR, and might be more easily portable to other languages.

III. Method and Analysis

This section describes the dataset, data preprocessing, acoustic features and machine learning methods employed in this study for detection of AD through spontaneous speech.

A. The Pitt Corpus

The Pitt corpus was gathered longitudinally between 1983 and 1988 on a yearly basis as part of the Alzheimer Research Program at the University of Pittsburgh [59]. Participants are categorised into three groups such as dementia, control (i.e. healthy), and unknown. 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 consent. Extensive neuropsychological and physical assessments conducted on the participants are also included [22]. The study reported in this paper selected only the dementia and control groups for a learning task of distinguishing between AD (including categorised by clinicians as probable AD and possible AD) and non-AD participants.

The Pitt Corpus contains participants' speech data collected by the Alzheimer and Related Dementias Study at the University of Pittsburgh School of Medicine on the following tasks:

1) a description task in which the participant is asked to describe, verbally in their own words, a picture (the "cookie theft" picture from the Boston aphasia examination),
2) word fluency task,
3) a story recall task, and
4) a sentence construction task.

The picture description task has been transcribed for AD and control patients, while the remaining tasks contain AD patient data only. We chose the picture description task sample because it contains spontaneously generated narrative speech. Table I shows the data available in this dataset.

TABLE I
STATISTICS OF THE DEMENTIABANK PITT CORPUS

	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

*One participant (ID:172) has changed the diagnosis from "Control" (in the first visit) to "Dementia" (in the remaining 3 visits).

As the AD and non-AD groups are not matched for age and gender in the original dataset, we created a derived dataset which is matched for age and gender, as shown in Table II, so as to minimise risk of bias in classification. The resulting dataset was segmented for voice activity using using a voice activity detection system based on a signal energy threshold. We set the log energy threshold parameter to 65 with a maximum duration of 10 seconds per speech segment. The segmented dataset contains 2033 speech segments from 82 non-AD subjects and 2043 speech segments from 82 AD subjects. The average number of speech segments produced by each participant in their descriptions was 24.86 (standard deviation $sd = 12.84$). Audio volume was normalised across all speech segments to control for variation caused by recording conditions, such as microphone placement.

TABLE II
BASIC CHARACTERISTICS OF THE PATIENTS IN EACH GROUP.

	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

B. Acoustic Feature Extraction

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 and affect recognition in speech [36]. The following is a brief description of each of the feature sets constructed in this way:

emobase: This acoustic feature set contains the mel-frequency cepstral coefficients (MFCC) voice quality, fundamental frequency (F0), F0 envelope, LSP and intensity features with their first and second order derivatives. Several statistical functions are applied to these features, resulting in a total of 988 features for every speech segment.

ComParE: The *ComParE 2013* [37] feature set includes energy, spectral, MFCC, 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. Statistical functionals are also computed, bringing the total to 6,373 features.

eGeMAPS: The *eGeMAPS* [35] feature set resulted from an attempt to reduce the somewhat unwieldy feature sets above to a basic set of acoustic features based on their potential to detect physiological changes in voice production, as well as theoretical significance and proven usefulness in previous studies [60]. It contains the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, Hammarberg index and slope V0 features, as well as their most common statistical functionals, for a total of 88 features per speech segment.

MRCG functionals: MRCG features were proposed by Chen et al. [30] and have since been used in speech related applications such as voice activity detection [61] speech separation [30], and more recently for attitude recognition [38]. MRCG features are based on cochleagrams [62]. 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 the audio signal. Four cochleagram features were 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 were extracted from each frame: 256 MRCG features (frame length of 20 ms and frame shift of 10 ms), along with 256 Δ MRCG and 256 $\Delta\Delta$ MRCG features. These features are meant to capture temporal dynamics of the signal [30]. The statistical functionals (mean, standard deviation, minimum, maximum, range, mode, median, skewness and kurtosis) were applied on the 768 MRCG features for a total of 6,912 features.

In sum, we extracted 88 eGeMAPS, 988 emobase, 6,373 ComParE and 6,912 MRCG features from 4,077 speech segments. Pearson's correlation test was performed on the whole dataset to remove acoustic features that were significantly correlated with duration (when $R > 0.2$). Hence, 75 eGeMAPS, 711 emobase, 3,899 ComParE and 4,688 MRCG features were not correlated with the duration of the speech chunks, and were therefore selected for the machine learning experiments. Examples of features from the ComParE feature set by the above described procedure include L1-norms of segment length functionals smoothed by a moving average filter (including their means, maxima and standard deviations), and the relative spectral transform applied to auditory spectrum (RASTA) functionals (including the percentage of time the signal is above 25%, 50% and 75% of range plus minimum).

C. Machine Learning Methods

The experiments conducted to test different feature sets and approaches to AD recognition through speech encompassed segment level classification, majority vote classification, and active data representation. experimental settings are described in the following sections, starting with a description of our ADR approach.

D. Active Data Representation

The acoustic information contained in a speech segment, which typically lasts only a few seconds, may not be enough for AD recognition. While segment-level aggregation through voting approaches has been used in computational paralinguistics, this approach does not reflect intersegmental relations other than a basic grouping according to their predicted class. The motivation behind ADR is to model the acoustic information of the full audio recording of a subject using acoustic features of all speech segments, representing an audio recording with a single fixed-dimension feature vector for the classification task. These features are extracted as follows:

1) *Segmentation and feature extraction*: each audio recording Ai ($i = 1 : r$, where r represents the total number of audio recordings or subjects) is divided into n speech segments $S_{k,i}$ as described in Section III-A, where k varies from 1 to n. Hence $S_{k,i}$ 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, at this processing stage. The system architecture depicted in Figure 1 illustrates this point.
2) *Clustering* of segments: self-organising maps (SOM) [63] are employed for clustering segments $S_{k,Ai}$ into m clusters $(C_1, C_2, ..., C_m)$ using audio features. Here m represents the number of SOM clusters. The number of clusters was determined through a grid search cross-validation procedure with a hyperparameter space of $m \in \{5, 10, \ldots, 100\}$.
3) *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), that is, creating a histogram of the number of speech segments ($nADR_{Ai}$) present in each of the m clusters for each audio recording. Then, to model temporal dynamics we calculate the mean and standard deviation of the rate of change with respect to the clusters associated with the speech segments for each audio recording ($cADR_{Ai}$), where the rate of change is given by an approximation of derivative

$$vADR_{Ai} = \frac{\partial cADR_{Ai}}{\partial t},$$

with respect to time (t). Finally, we calculate the duration of segments in each cluster (Ai), building a histogram representation of segment duration ($dADR_{Ai}$).

4) *Normalisation*: as the number and duration of segments is typically different for each audio recording due to inter-subject variability, we normalise the feature vector by dividing it by the total number (duration) of segments present in each audio recording (i.e. the L1 norm of $nADR_{Ai}$ and $dADR_{Ai}$), as shown:

$$nADR_{Ai_{norm}} = \frac{nADR_{Ai}}{\|nADR_{Ai}\|_1} \quad (1)$$

$$dADR_{Ai_{norm}} = \frac{dADR_{Ai}}{\|dADR_{Ai}\|_1} \quad (2)$$

5) *Fusion*: the $ADR_{Ai_{norm}}$ feature set encompasses the features of $nADR_{Ai_{norm}}$, $dADR_{Ai_{norm}}$, $vADR_{Ai}$, age and gender. Therefore a feature vector with dimensionality of $2 \times (m + 2)$ is generated to represent each subject.

A flowchart of the ADR generation procedure is available in supplementary material.

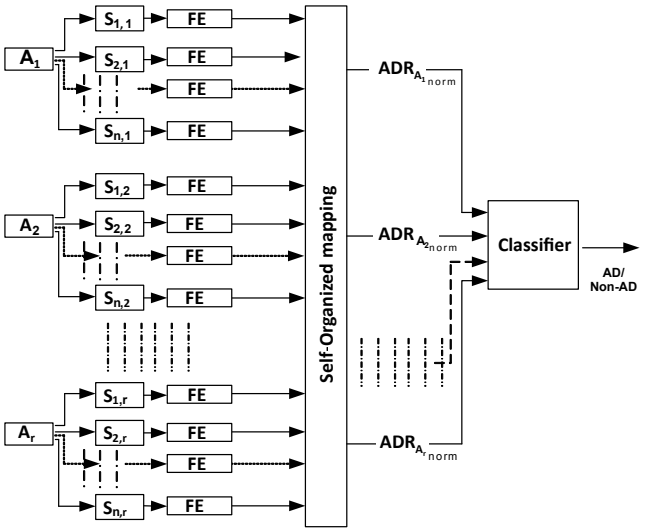


Fig. 1. Automatic detection of AD and non-AD subject using the Active Data Representation ($ADR_{Ai_{norm}}$) method where FE represents the extraction of low level features (such as eGeMAPS) from speech segments.

E. Classification methods

The classification experiments were performed using five different methods, namely decision trees (DT, with leaf size of 20), nearest neighbour (KNN with K=1), linear discriminant analysis (LDA), random forests (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 sequential minimal optimisation solver). The classification methods are implemented in MATLAB [64] using the statistics and machine learning toolbox. A leave-one-subject-out (LOSO) cross-validation setting was adopted, where the training data do not contain any information of validation subjects.

F. AD Detection

As mentioned above, we conducted three classification experiments to detect cognitive impairment due to AD, namely:

1) 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;
2) majority vote (MV) classification: using the results of 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; and
3) active data representation: we generate the ADR using acoustic features as described in section III-D, and then used $ADR_{Ai_{norm}}$ for classification as before.

TABLE III
SEGMENT LEVEL CLASSIFICATION (CHANCE LEVEL 50.12%).

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

TABLE IV
MAJORITY VOTE CLASSIFICATION (CHANCE LEVEL 50.00%).

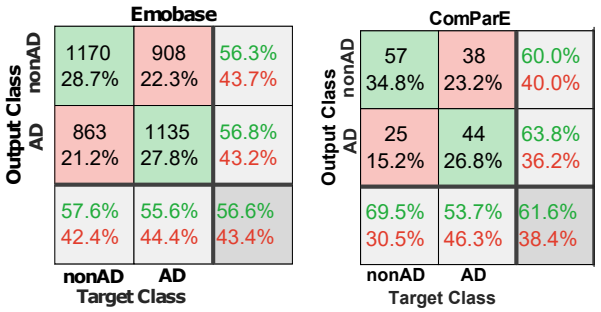
Features	LDA	DT	1NN	SVM	RF
emobase	53.66	56.10	48.17	56.71	57.93
ComParE	**61.59**	46.95	53.05	54.88	58.54
eGeMAPS	50.61	51.83	48.17	50.61	60.98
MRCG	50.61	54.88	54.27	56.10	56.10

IV. Results and Discussion

The classification accuracy of segment level, majority vote and ADR are shown in Table III, IV and V respectively. These results show that the ADR (77.44%) provides better results than majority vote (61.59%) in most of the cases (17 out of 20) , with LDA being the best classifier for AD detection with accuracy well above the chance level of 50%. A possible reason for the better results obtained using ADR of acoustic features in comparison with MV (which relies on segment level classification) is ADR's ability to encode acoustic information of a full audio recording into a single feature vector for model training. The segment level classification accuracy is very low (56.55%, against a random baseline of 50.12%)

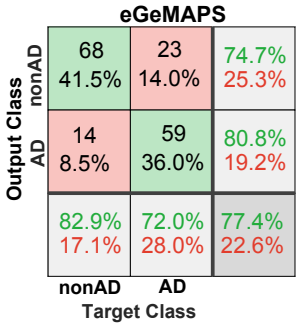
TABLE V
ADR CLASSIFICATION RESULTS WITH NUMBER OF CLUSTERS (m) USED. THE CHANCE LEVEL IS 50.00%

Features	LDA, m	DT, m	1NN, m	SVM, m	RF, m
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
eGeMAPS	**77.44**, 85	71.34, 30	54.27, 65	52.44, 20	71.34, 30
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



(a) Segment level classification (RF)

(b) Majority vote using LDA



(c) ADR of eGeMAPS using LDA

Fig. 2. Confusion matrices of the best results of each experiment along with precision and recall for each class, and overall accuracy.

which suggests that the speech segments carry contradictory or incomplete information at the speaker level, which results in poor machine learning models for classifying AD and non-AD patients. It also suggest that the AD and nonAD subjects have some common speech characteristics at the segment level. 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 2.

From the results shown in Table V and Figure 3, 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%). We have also note that the ADR of eGeMAPS (71.34%) provides the best result for DT, and the ADR of MRCG (69.51 %) yields results close to ADR of eGeMAPS using DT. It is noted that the dimensionality of ADR of eGeMAPS ($m = 30$ will result in 64 features for each audio recording) is higher than dimensionality of ADR of MRCG ($m = 15$ will result in 34 features for each audio recording), even though the former starts from a much smaller feature set than the latter. The ADR of the emobase feature set provides the least accurate results (66.5%), which could be due to the fact that it does not use jitter, shimmer and spectral flux features which indicate speech instability.

A limitation of ADR is that by clustering the original acoustic features and then extracting new features based on these clusters one loses the ability to assess the contributions of the original features to the model's predictions (as one is able to do with factor analysis "loadings", for instance). We intend to tackle this issue in future work. For now, to better understand the relationship between the feature sets as regards the DT classifier, we drew the Venn diagram shown in Figure 4; the blue area (labelled "Target") represents the annotated labels, the yellow area represents the predicted labels when the ComParE feature set was used, the green ellipse represents the predicted labels under eGeMAPS, the red ellipse represents the prediction under the emobase feature set, and finally the brown area represents labels predicted with the MRCG features. From the overlaps in this diagram, we observed that there are 6 instances (one non-AD, and five AD) which have not been recognised by any of the feature sets. Subjective evaluation by one expert confirmed that the AD patients' recordings had good voice quality while the non-AD participant had poor voice quality. This could be due to a natural variability with which patients manifest signs of neurodegeneration [9]. Further clinical information is required to investigate this possibility in greater depth. Unfortunately, this information is not available in this dataset. It is possible that linguistic information might have helped diagnose those subjects. This is a possibility we aim to investigate in future work. The age and gender characteristics of the misclassified patients are: the AD patients were 52, 58, 67, 71, and 74 years old (mean 64.4), two males and three females, and the non-AD patient was a 78 year old male. It is possible the age and gender also acted as confounders for these patients, as AD affects mostly older people, and most often females.

There are 46 instances (32 of non-AD and 14 of AD) which have been detected by all four feature sets. The Venn diagram suggests 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 correctly classified by all the feature sets. For example, the ADR of ComParE does not use features F1, F2, F3 and alpha ratio, which is a possible reason why it is capturing different information than the other ADRs. This suggests that the fusion of the results could improve overall accuracy.

We implemented a simple "hard fusion" procedure by taking a vote among decision tree classifiers for the four feature sets, breaking ties by assuming an AD label. As hypothesised,

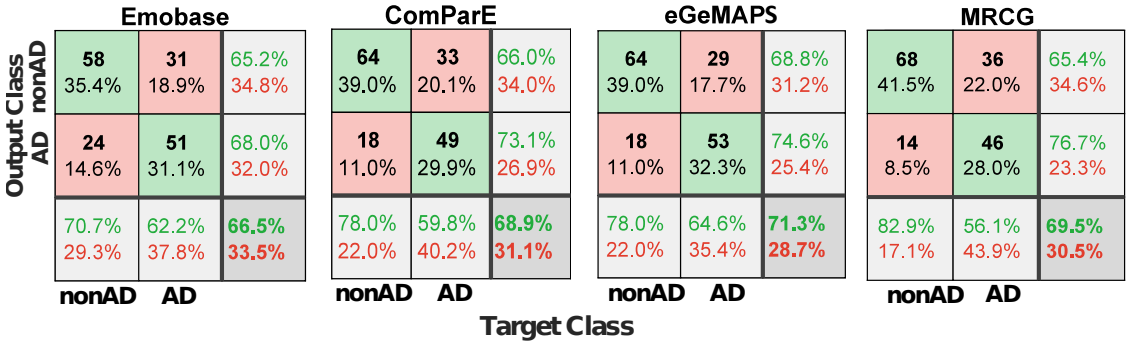


Fig. 3. Confusion matrices of decision trees obtained using active data representation along with precision and recall for each class, and overall accuracy.

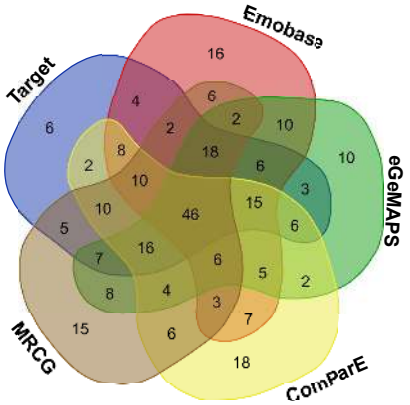


Fig. 4. Venn Diagram of the best results of four feature sets using decision tree classifiers and clinical diagnosis (Target).

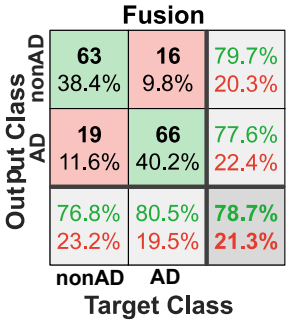


Fig. 5. Hard fusion of decision tree results.

fusion provides the best results, with an accuracy of 78.7% as shown in Figure 5.

These results are comparable to those attained by state-of-the-art models working with the speech recordings available for the Pitt corpus, reviewed in section II. These are presented in [29], [31] and [32], who reported 81.92%, 80% and 79%, respectively, for AD-nonAD classifiers (see Table VI). Although some of these studies report slightly higher accuracy than ours, all of those that do include information from the manual transcripts, and were conducted on an unbalanced data set (in terms of age, gender and number of subjects in the AD and non-AD classes). The performance of a model without the information from transcripts, that is, relying only on acoustic features as we do, is only reported in [32], dropping significantly to an average accuracy of 62% with an SVM. It is also noted that previous studies do not evaluate their methods in a complete subject-independent setting (i.e. they are considering multiple sessions for a subject and classifying a session instead of a subject). This could lead to overfitting, as the model might learn speaker dependent features from a session and then, based on those features, classify the next session of the same speaker. One strength of our method is its speaker independent nature, and as such we evaluated our method in a LOSO cross-validation setting.

TABLE VI
COMPARISON WITH THE STATE OF THE ART

Study	accuracy	modality	fully automatic
This study	78.7%	acoustic	yes
Hernández et AL.[32]	62.0%	acoustic	yes
Luz [33]	68.0%	acoustic	yes
Mirheidari et Al. [55]	62.3%	text	yes (ASR)
Fraser et Al. [29]	81.9%	text/acoustic	no (text)
Yancheva & Rudzicz [31]	80.0%	text/acoustic	no (text)
Hernández et AL.[32]	68.0%	text	no
Mirheidari et Al. [55]	75.6%	text	no

Working on a balanced and standardised subset of the original dataset implies necessary trade-offs. The most significant of these is the information contained in voice loudness. Voice volume is a good indicator of AD, and we are possibly losing that information by normalising the volume of all speech segments. However, a machine learning model trained on normalised volume is robust against variations in distance between microphone and subject (as the volume/energy of speech signal varies if a subject is close to microphone or far from it) which makes the machine learning model more suitable to monitor subjects in far-field settings and under diverse recording conditions.

Furthermore, our fully automated acoustic-prosodic model performs better than the fully automated transcription-based model presented in [55], who reported a maximum accuracy of 62.7% for a combination of convolutional and long short term memory deep neural networks on ASR data, for the DementiaBank dataset. We also achieve higher accuracy than the only available work exclusively reliant speech for automatic AD-nonAD classification of the Pitt dataset (68%) [33].

V. Conclusion

This article demonstrates the relevance of acoustic features of spontaneous speech for cognitive impairment detection in the context of Alzheimer's Disease diagnosis. Machine learning methods operating on automatically extracted voice features provide accuracy of up to 78.7%, well above the chance level of 50%. Our results improve on those from fully automated models on the same dataset such as [55] (transcription based) and [33] (acoustic-prosodic). We argue that there are at least three reasons for these improvements with respect to those previous works. Firstly, that a more comprehensive acoustic feature set is able to capture a wider range of speech subtleties potentially indicative of neurodegeneration or impairment. Secondly, that a larger number of features allows for the use of more sophisticated classifiers, as well as for the implementation of our novel ADR method, which shows an improvement in relation to previous methods. And thirdly, that we used an enhanced version of the dataset.

We expect that these findings will contribute to the development of screening tools for AD that are cost-effective and non-invasive (as compared to imaging and blood biomarkers). Furthermore, since linguistic abilities respond reasonably well to certain treatments for AD [65], our method could also be applied to monitor responsiveness to such interventions.

As the next steps towards these goals, we will extend the research presented here for prediction of MMSE scores [66], available in the corpus. We also intend to apply our method to spontaneous dialogue data, which we are currently collecting following the Prevent-ED protocol [67]. Prevent-ED participants are healthy adults with a comprehensive risk profile (i.e. genetics, cognitive assessments and family history of AD). We hypothesise acoustic-prosodic analysis to be sensitive enough to capture dialogical signs that may be present in preclinical AD. Hence, by applying the method presented in this paper to the Prevent-ED dialogues, we expect to predict risk of AD and offer some insight into the earliest stages of the disease.

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)*, 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 DataLibrary Cataloguing-in-Publication Data*, pp. 1–44, 2013.

[4] ——, "First WHO ministerial conference on global action against dementia: meeting report," *WHO Library Cataloguing-in-Publication DataLibrary Cataloguing-in-Publication Data*, pp. 1–76, 2015.

[5] R. B. Maccioni, G. Farías, 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, pp. 3–12, 1996.

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

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

[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. W. Bondi, D. P. Salmon, and A. W. Kaszniak, "The neuropsychology of dementia," in *Neuropsychological Assessment of Neuropsychiatric Disorders*, 2nd ed., 1996, pp. 164–199.

[13] P. N. Dawadi, D. J. Cook, and M. Schmitter-Edgecombe, "Automated cognitive health assessment using smart home monitoring of complex tasks," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 6, pp. 1302–1313, 2013.

[14] 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.

[15] 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." *J. of Psychiatric Res.*, vol. 12, no. 3, pp. 189–98, 1975.

[16] P. Robert, S. Schuck, B. Dubois, J. Lépine, T. Gallarda, J. Olié, 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, 2003.

[17] B. Dubois, A. Slachevsky, I. Litvan, and B. Pillon, "The FAB: a frontal assessment battery at bedside," *Neurology*, vol. 55, pp. 1621–6, 2000.

[18] 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." *Intl. Psychogeriatrics*, vol. 17, no. 3, pp. 461–74, 2005.

[19] 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.

[20] 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: Diagn., Assess. & Dis. Mon.*, vol. 1, no. 1, pp. 112–124, 2015.

[21] 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.

[22] 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.

[23] H. Goodglass and E. Kaplan, "The assessment of aphasia and related disorders," Philadelphia, 1983.

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

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

[26] S. O. Orimaye, J. S. Wong, K. J. Golden, C. P. Wong, and I. N. Soyiri, "Predicting probable Alzheimers disease using linguistic deficits and biomarkers," *BMC bioinformatics*, vol. 18, no. 1, p. 34, 2017.

[27] 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, p. e0205636, 2018.

[28] K. C. Fraser, K. L. Fors, and D. Kokkinakis, "Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment," *Computer Speech & Language*, vol. 53, pp. 121–139, 2019.

[29] 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.

[30] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1993–2002, 2014.

[31] M. Yancheva and F. Rudzicz, "Vector-space topic models for detecting Alzheimers disease," in *Procs. of ACL*, 2016, pp. 2337–2346.

[32] L. Hernández-Domínguez, S. Ratté, G. Sierra-Martínez, and A. Roche-Bergua, "Computer-based evaluation of Alzheimers disease and mild cognitive impairment patients during a picture description task," *Alzheimer's & Dementia: Diagn., Asses. & Dis. Mon.*, vol. 10, pp. 260–268, 2018.

[33] S. Luz, "Longitudinal monitoring and detection of Alzheimer's type dementia from spontaneous speech data," in *Procs. of the Intl. Symp on Comp. Based Medical Systems (CBMS)*. IEEE, 2017, pp. 45–46.

[34] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. Wiley, 2013.

[35] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, 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 Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016.

[36] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: the Munich versatile and fast open-source audio feature extractor," in *Procs. of ACM-MM*. ACM, 2010, pp. 1459–1462.

[37] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *ACM-MM*. ACM, 2013, pp. 835–838.

[38] F. Haider and S. Luz, "Attitude recognition using multi-resolution cochleagram features," in *Procs. of ICASSP*, 2019, pp. 3737–3741.

[39] 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," *Curr. Alzheimer Res.*, vol. 15, no. 2, pp. 130–138, 2018.

[40] S. Luz, S. D. la Fuente, and P. Albert, "A method for analysis of patient speech in dialogue for dementia detection," in *Procs. of LREC'18*, D. Kokkinakis, Ed. Paris, France: ELRA, may 2018.

[41] K. Lopez-de Ipiña, M. Faundez-Zanuy, J. Solé-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.

[42] K. Lopez-de Ipiña, J. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C. M. Travieso, M. Ecay-Torres, P. Martinez-Lage, and H. Egiraun, "On Automatic Diagnosis of Alzheimer's Disease based on Spontaneous Speech Analysis and Emotional Temperature," *Cognitive Computation*, vol. 7, no. 1, pp. 44–55, 2015.

[43] 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.

[44] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2081–2090, 2011.

[45] 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 *CLPsych*, 2014, pp. 27–37.

[46] S. Luz, "Automatic identification of experts and performance prediction in the multimodal math data corpus through analysis of speech interaction," in *Procs. of ICMI*. ACM, 2013, pp. 575–582.

[47] 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.

[48] 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 *Procs. of INTERSPEECH*. ISCA, 2013, pp. 1692–1696.

[49] A. Satt, R. Hoory, A. König, P. Aalten, P. H. Robert, N. Sophia,

[55] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Detecting signs of dementia using word vector representations." in *INSTERSPEECH*, 2018, pp. 1893–1897.

C. Mémoire, D. Ressources, and C. H. U. D. Nice, "Speech- Based Automatic and Robust Detection of Very Early Dementia," in *Procs. of INTERSPEECH*, Singapore, 2014, pp. 2538–2542.

[50] 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.

[51] 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.

[52] 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.

[53] B. Croisile, B. Ska, M.-J. Brabant, A. Duchene, Y. Lepage, G. Aimard, and M. Trillet, "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.

[54] J. Weiner, M. Engelbart, and T. Schultz, "Manual and automatic transcriptions in dementia detection from speech." in *Procs. of INTERSPEECH*, 2017, pp. 3117–3121.

[56] B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "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.

[57] ——, "An avatar-based system for identifying individuals likely to develop dementia," in *Procs. of INTERSPEECH*, 2017, pp. 3147–3151.

[58] K. Ekberg and M. Reuber, "Can conversation analytic findings help with differential diagnosis in routine seizure clinic interactions?" *Communication & Medicine*, vol. 12, no. 1, p. 13, 2015.

[59] J. Corey Bloom and A. Fleisher, "The natural history of Alzheimer's disease," *Dementia*, vol. 34, pp. 405–15, 2000.

[60] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, 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 Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016.

[61] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Processing Letters*, 2018.

[62] 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.

[63] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1-3, pp. 1–6, 1998.

[64] MATLAB, *version 9.6 (R2019a)*. Natick, Massachusetts: The MathWorks Inc., 2019.

[65] S. H. Ferris and M. Farlow, "Language impairment in Alzheimers disease and benefits of acetylcholinesterase inhibitors," *Clinical Interventions in Aging*, vol. 8, p. 1007, 2013.

[66] 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," *J. of Psychiatric Res.*, vol. 12, no. 3, pp. 189–198, 1975.

[67] 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.