Dementia detection using automatic analysis of conversations

Bahman Mirheidari[a,*], Daniel Blackburn[b], Traci Walker[c], Markus Reuber[d], and Heidi Christensen[a]

[a] *Department of Computer Science, University of Sheffield, Sheffield, UK*
[b] *Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK*
[c] *Department of Human Communication Sciences, University of Sheffield, Sheffield, UK*
[d] *Academic Neurology Unit, University of Sheffield, Royal Hallamshire Hospital, Sheffield, UK*

Abstract

Neurogenerative disorders, like dementia, can affect a person's speech, language and as a consequence, conversational interaction capabilities. A recent study, aimed at improving dementia detection accuracy, investigated the use of conversation analysis (CA) of interviews between patients and neurologists as a means to differentiate between patients with progressive neurodegenerative memory disorder (ND) and those with (non-progressive) functional memory disorders (FMD). However, manual CA is expensive and difficult to scale up for routine clinical use. In this paper, we present an automatic classification using an intelligent virtual agent (IVA). In particular, using two parallel corpora of respectively neurologist- and IVA-led interactions, we show that using acoustic, lexical and CA-inspired features enables ND/FMD classification rates of 90.0% for the neurologist-patient conversations, and an encouraging 90.9% for the IVA-patient conversations. Analysis of the significance of individual features show that some differences exist between the IVA and human-led conversations for example in average turn length of patients.

Keywords: Dementia detection, conversational analysis, speech recognition and segmentation, processing of pathological speech

1. Introduction

Dementia is a neurodegenerative disorder of the brain, which is caused by a number of conditions including Alzheimer's disease (AD). AD presents most

☆

[*] B Mirheidari
Email address: bmirheidari2@sheffield.ac.uk (Bahman Mirheidari[a,*], Daniel Blackburn[b], Traci Walker[c], Markus Reuber[d], and Heidi Christensen[a])

commonly with an episodic memory disorder, but language impoverishment is also frequently present; manifesting itself as problems with e.g., object naming, noun production and verb usage (Bayles and Boone, 1982; Bayles, 2003; Tomoeda et al., 1996; Hamilton, 1994; Forbes-McKay and Venneri, 2005).

The observation of a patient's language is therefore included in routine dementia assessments. A typical part of any consultation with a doctor is the process of *history taking*, where an examiner (the doctor) asks a number of specific questions of a patient, observes and interprets the responses. This whole process is carried out in an informal qualitative manner (leaving the clinician with an impression that a particular patient is or is not likely to be developing a form of dementia). In addition, the majority if not all patients with cognitive complaints will initially present to primary care. Clinicians working in primary care have limited time, training or skills in diagnosing dementia, in particular in the early stages. They utilise several brief general pen-and-paper cognitive screening tests. These tests lack sensitivity and specificity (Hessler et al., 2014) and they only represent a snapshot of the patient's capability at that particular time.

There is currently no automatic tool to aid the processing or scoring of the spoken communication ability of the patient. For conditions like dementia, where detecting the more subtle diagnostic clues is likely to require expert neuropsychological knowledge, this can mean that conversational clues are missed and clinicians in primary care, may over rely on scores of brief cognitive testing and miss early changes in language that would suggest presence of early stage underlying neurodegenerative pathology.This may result in incorrect referral decisions, so that patients who are highly functioning (e.g. present as very articulate) are not referred to specialist services for the further investigation of possible dementia, and as a result will be not receive a correct diagnosis until their disorder has progressed (false negatives). Conversely, other patients, presenting with memory problems due to excessive worry or anxiety may be referred for further assessment in memory clinics despite *not* being at increased risk of having neurodegeneration (false positives). Therefore, it would be highly desirable to develop an automatic screening aid based on processing of the patient's conversational and communication capabilities. Such a tool could be particularly useful for non-experts involved in screening (GPs, community nurse, etc.) but it could also be helpful for experts, especially if used longitudinally, for instance to monitor signs of progression in people with Mild Cognitive Impairment or response to treatment. Developing reliable methods for automatically detecting dementia could therefore open up for more ubiquitous and frequent testing in more relaxed settings, such as people’s own homes. This would improve test reliability by averaging out the effect of patients having a *bad day* or feeling anxious when tested in a more formal, clinical setting.

A recent study, by Elsey *et al.*, applied conversation analysis (CA) to neurologist-patient interactions and found that a set of interactional patterns of patients' (and accompanying others') conversational behaviour could be used to distinguish between patients developing a neurodegenerative (ND) disorder and patients with functional memory disorder (FMD; other non-dementia related problems with memory) (Elsey et al., 2015; Jones et al., 2015). The study showed promising results in terms of diagnostic accuracy, but relied on manual CA for the detection of the interaction patterns in the conversation. This involves a number of steps including audio recording, manually transcribing the encounters and carrying out a qualitative analysis by a trained expert. It is thus prohibitively expensive and time consuming, and not suitable for routine clinical use. This paper presents an automatic dementia detection test developed around a classifier pipeline where CA-based, acoustic and lexical features are extracted from conversations. We report on the outcome of a feasibility study where this interaction analysis was carried out on conversations elicited by an interaction virtual agent (IVA). To our knowledge, this is the first test of such a system in a real clinical setting.

Automatic CA is an emerging and challenging area of research with some promising results, e.g. (Shriberg, 2005; Moore, 2015). It typically involves a number of technologies to automate the required steps including automatic speech recognition (ASR), speaker diarisation (*"who's speaking when"*) and some automatic speech understanding.

Previously, we reported that we were able to replicate the findings of Elsey et al. (2015) in a proof-of-principle study (Mirheidari et al., 2016, 2017b,a). The linear SVM machine learning classifier could classify between the two different patient group with 95% accuracy if the manually produced transcript of conversation was used. By replacing the manual transcripts with the ASR outputs, the classification accuracy rate dropped to 79%, which improved to 90% when only the top 10 features were selected. In addition, all of the overlapping and short segments were removed from the data and it was assumed that the speaker's turn information was produced by a perfect diarisation tool.

The work outlined in this paper, presents further steps towards an automatic cognitive impairment screening or stratification system. The use of innovative screening device to improve detection of cognitive impairment is a research priority due to the very high demands on both primary and secondary care clinicians as dementia becomes increases with and ageing population and is widely feared by the public. We envisage an automated speech analysis system being of use in the early steps of the cognitive impairment diagnostic pathway. This is likely to be in between the primary care physician and secondary care consultation as has been suggested in the 2015 review by (Laske et al., 2015). The initial analysis of the IVA-based system in a memory clinic setting (Mirheidari et al., 2017a) showed promising results. This paper presents a more detailed analysis of the evaluation including the effects of automatic speech recognition and diarisation. We show that patients, in general, liked interacting with the IVA, and that, despite dealing with very challenging data with an inherently large prevalence of hesitations, false starts and overlapping speech, a reliably high classification accuracy can be obtained.

This paper is structured as follows: Section 2 summarises the background for the current study. Section 3 includes details of the proposed dementia detection system and challenges of dealing with conversations including spontaneous speech recognition, speaker diarisation and spoken language understanding. The experimental setup and results are described in sections 4 and 5 respectively. Finally, section 6 contains the conclusions and a discussion of future directions for this work.

2. Background

Conversation Analysis (CA) was primarily developed as a sociological method but has incorporated contributions from other disciplines including linguistics, communication and political science, anthropology and psychology (Sidnell and Stivers, 2012). CA is a qualitative research approach designed to investigate the structural organisation of everyday social interaction. It is based on the observation that conversations are built on structures known as adjacency pairs (such as question and answer, greeting and greeting, compliment and down player, request and grant, etc. Jurafsky and Martin (2008), and that take place as a joint activity between two or more interlocutors who exchange discourses in a consecutive manner (turns). CA provides a rigorous framework which identifies the normative underpinnings of talk and employs a proof procedure reliant on evidence found within the sequential and linguistic features of the talk itself in order to discover recurrent patterns in the conversation (Lerner, 2004).

Whilst the automatic analysis of interaction is quite a new field of study (Moore, 2015) and a method, which has not been applied to the differential diagnosis of memory problems, a significant amount of work has been carried out using machine learning techniques to identify signs of dementia in patient's speech and language. For instance, researchers have attempted to extract acoustic features from recorded speech of people with dementia. Lopez de Ipina et al. (2013) investigated a number of acoustic features including durations (e.g. voice/unvoiced segments), time domain (short time energy), frequency domain (spectral centroid), and the fractal dimension from the AZTIAHO database of multilingual recordings of the spontaneous speech of 50 healthy adults and 20 Alzheimer patients. They used a perceptron classifier to distinguish between the patients with dementia and healthy controls. In their recent work Lopez de Ipina et al. (2015) used an alternative classifier, the Gaussian support vector machine (SVM), which produced better results.

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

Toth et al. (2015) have found other acoustic and lexical features (such as the number of phonemes per second, length of utterance and pauses) very useful in identifying patients with mild cognitive impairment (MCI), a condition characterised by more modest cognitive difficulties, and not resulting in significant functional impairment as opposed to those seen in dementia. It is important to detect those cases of MCI with higher likelihood of progressing to dementia in order to test new treatments to be trailed earlier in the disease process. They trained their ASR using the BEA Hungarian Spoken Language Database (spontaneous speech of people with MCI) focusing only on phoneme recognition. In their recent work, Gosztolya et al. (2016) expanded the initial feature set by including a set of 'extended' features including descriptors for silence pauses, filled pauses and some particular phonemes. They also added 708 'overcomplete' features (these are redundant versions of features with different descriptors for 57 phonemes, pauses, breathing noises, laughter and coughs). They applied a number of different feature selection algorithms to identify the most informative features for classification. The results showed that training a classifier with fewer features obtained by an efficient feature selection algorithm can outperform classifiers trained on all the features of the initial feature set - including extended or overcomplete features. They also suggested a new technique for feature selection ('correlation-based' method), - which can be as accurate as a forward feature selection algorithm, yet, somewhat more efficient and faster.

Most studies have focused on distinguishing healthy controls from patients with different types of dementia, most commonly AD. Jarrold et al. (2014) combined half of their ASR outputs with half of their human transcriptions of spontaneous speech to extract acoustic and lexical features. They classified 48 participants into groups with different types of dementia and a healthy control group. The classification accuracy amongst all subjects was 61%, while the binary classification accuracy between AD and healthy controls rose to 88%.

Likewise, Thomas et al. (2005) saw a drop in accuracy when attempting to distinguish between more than one type of dementia. They extracted several lexical and semantic features to achieve 95% accuracy in a binary classification task differentiating between patients with severe dementia and normal controls but saw a drop to around 75% when attempting differentiate between patients with mild dementia and healthy elders – two groups with far greater features overlap. When four classes of cognitive performance were introduced (severe dementia, moderate dementia, mild dementia, and normal group), the classification accuracy dropped further to around 50%. Satt et al. (2013) carried out a study on 89 subjects (43 with MCI, 27 with AD and 19 healthy adults). The subjects were asked to complete tasks such as verbally describing a picture while looking at it, looking once at a picture and describing it from memory and repeating a sentence given by the interviewer. The gained 80% accuracy rate for the classification between MCI and AD group.

In a study on data collected from the interdisciplinary longitudinal study on adult development and ageing (ILSE, a German collection of 1000 participants' spontaneous speech in their middle adulthood and later life, spanning 20 years), Weiner et al. (2016) extracted a number of acoustic and linguistic features (e.g. silence duration, silence-to-speech ratio, word rate, phoneme rate) to train a classifier to distinguish between three categories: AD (5 patients), ageing-associated cognitive decline (AACD) (13 patients) and healthy adults (80). For this work, they did not apply ASR but instead used manual transcriptions for the lexical features and trained a voice activity detection (VAD) to calculate the acoustic features. Using a linear discriminant analysis (LDA) classifier, they obtained 85.7% classification accuracy for the three participant groups. While differentiating between healthy participants and those with AD was successful, the classifier was not capable of categorising the healthy group from the AACD patients.

Recent research Fraser et al. (2015); Yancheva et al. (2015) has used the DementiaBank corpus (containing speech of patients with AD, vascular dementia, MCI and healthy controls describing the 'Cookie Theft' picture) to predict changes in patients' Mini Mental State Examination (MMSE) scores over time. The researchers extracted a wide range of features (over 477 lexico-syntactic, acoustic, and semantic) and selected the 40 most informative, reporting an accuracy of over 92% in terms of the distinction of AD patients from healthy controls. This relatively high classification accuracy was based on manually transcribed audio files. However, in their very latest study Zhou et al. (2016) used state-of-the-art ASR to produce automatic transcriptions. Their best ASR had a 38.2% word error rate (WER). Ignoring the prosodic and the acoustic features, this time they only extracted the lexical features to train an SVM classifier in order to differentiate between the healthy group and AD patients. The accuracy of the classifier dropped significantly as the ASR WER increased. However, they found a weak correlation between these two. The poor quality of the recordings and problems recognising the participants' speech (because of high levels of breathiness, jitter, shimmer, and a slower speaking rate) were reported as the main challenges for the ASR.

Recently, Tanaka et al. (2016) described an avatar/IVA-based system for detecting dementia. Although, that system was based on standard neuropsychological tests, it demonstrated encouraging results for the use and acceptability of an IVA-based, automatic and interactional system for patients with memory concerns.

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

Many health-related software tools and apps are currently using, or exploring the use of IVAs for interacting with people with mental health problems (Rus-Calafell et al., 2014; Leff et al., 2014; Hayward et al., 2017), healthy elderly (Cyarto et al., 2016) and people with AD (Carrasco et al., 2008; Tran et al., 2016). These technologies are preferred over other modes of interaction (keyboards or touch screens). Disclosing information to a computer, rather than a person, may reflect a more honest detailed history as social embarrassment is avoided especially if the talking head is perceived to be run by AI(Rizzo et al., 2016).

In general, the distinction between AD and healthy controls represents much less of a diagnostic challenge than the differentiation of MCI and age-matched adults without cognitive complaints, or even age-matched adults with non-progressive memory complaints.

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

3. Automatic dementia detection system

Figure 1 shows a block diagram of our proposed automatic dementia detection system. First, an audio file containing a recording of the conversation is passed to a diarisation tool to identify the speech portions of the input audio stream as well as the speaker of each speech segment. This information is then passed to an automatic speech recognition (ASR) system. ASR is given both the input audio file and the output produced by the diarisation tool to generate a string of words spoken by each speaker.

We employ the SHoUT (Huijbregts, 2008) diarisation toolkit. The Kaldi toolkit (Povey et al., 2011) was used following a standard recipe for training the speech recognition acoustic model (16 mixture model, sat-trained HMM-GMM) and language model (tri-gram; based on training dataset using Kneser smoothing).

Next, the output of the diarisation tool and the ASR are given to the feature extraction unit to extract a number of features. For instance, using the start time and end time of each turn of the conversation, the average length of the turn for a specific speaker can be calculated. Some features may require further techniques such as text processing, natural language processing (NLP) and spoken language understanding (SLU). A number of acoustic features can be extracted directly from the audio recording using toolkits such as 'Praat'. The extracted feature types are further described in Section 4.2.

Finally, the extracted features are sent to a machine learning classifier (SVM) to decide which category the whole conversation belongs to; in this study the two groups are patients diagnosed with neurodegenerative dementia (ND) and patients with functional memory disorder (FMD) (Schmidtke et al., 2008).

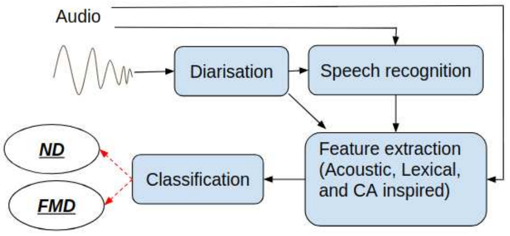


Figure 1: Automatic dementia detection system.

4. Experimental setup

4.1. Data

This study has used data from two parallel corpora both of which were recorded at the Royal Hallamshire Hospital's memory clinic in Sheffield, UK. All participants received written information about the study and gave informed consent prior to taking part. Ethical permission for the study was granted by the NRES Committee South West-Central Bristol (Rec number 16/LO/0737) in May 2016.

4.1.1. Neurologist-patient conversations

The first dataset is that previously used in (Elsey et al., 2015). It consists of audio recordings and associated manual CA annotations of neurologist-patient conversations. For this study, a total of 30 conversations have been used (15 FMD and 15 ND). Patients were encouraged to bring an accompanying person with them for their outpatient appointment in the memory clinic to provide more information about the patients if required, and to support the patient, if needed (21 people brought an accompanying person including 15 in ND group and 6 in FMD group). As a result, many conversations include three participants. In the following, we will use 'Neu' (neurologist)), 'Pat' (patient) and 'Aps' (accompanying person(s)) to distinguish them. The neurologists were instructed to attempt to follow a predefined set of questions constructed to reveal the typical signs of impairments in the conversation. Several categories of questions were included:

- Closed questions needing long-term memory recall of personal details the person is meant to know (e.g, "How old were you when you left school?").
- Compound questions (e.g., "Why have you come here today, and what are your expectations?"). People with dementia tend to find it difficult to remember to answer both parts.
- Open-ended questions like "What did you do after you left school?".
- Questions related to the memory concern, like "Who is most worried about your memory?" (for ND patients it tends to be other members of the family who are worried about the patient) and "Tell me about the last time you had a problem with your memory", which FMD patients find easier to answer, in particular providing detailed examples with accurate temporal information.

As the data were recorded in ordinary clinical settings and were not initially recorded with the aim of applying speech recognition, little effort was made to reduce background noise and acoustic interference, and for many of the recordings, the microphone placement was relatively *ad hoc* (often being placed closer to the neurologist than the patient). In addition, the speech itself was very challenging with a high percentage of overlapping speech segments – on occasion even the professional transcribers have not been able to transcribe the material. In preliminary work, we tested the effect of various noise reduction techniques like reducing noise by taking a profile of the background noise and subtracting it from the entire recording. The approach worked to some extent, however, due to the high speech-to-noise ratio, further reductions resulted in a loss of speech quality. Therefore, we ended up a moderate noise reduction by the deduction technique. Automatically identifying and removing all segments containing overlapping segments from the conversations is a challenging and non-trivial task, and runs the risk of introducing a significant loss of information around the borders of overlapping segments which may affect the classification accuracy. Therefore, we decided to handle overlapping speech segments with a very light-tough approach essentially by only removing very short segments (less than 0.4 sec as output by the diarisation moduel), which automatically reduced the amount of overlap while preserving some of the border information.

For this study, we had access to an additional dataset of 24 conversations for the training of the speech recogniser. These were recordings of patients whose diagnosis is either uncertain or not belonging to the ND or FMD groups. We have used this data to boost the acoustic model though.

4.1.2. IVA-patient conversations

The second dataset was recorded with the purpose of investigating the feasibility using a computer-based agent, an IVA to take a memory clinic-style history. Figure 2 shows a screen-shot of the IVA as well as the tool in use.

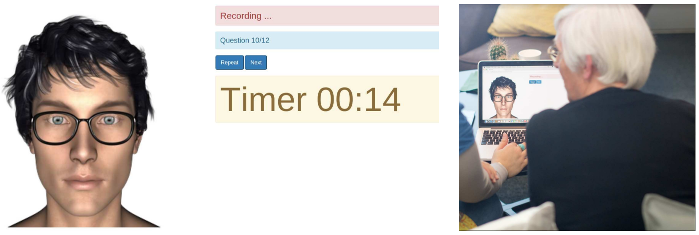


Figure 2: Prototype IVA (using https://www.botlibre.com) and tool in use.

Questions similar in style to the ones used in the neurologist-patient conversations were put together in collaboration with a clinical linguist. For this initial study, these questions were recorded by an American male but for future work we plan to replace this with a synthetic voice. For this feasibility study we chose a male character to match the male voice recordings we had of the questions. We know from subsequent co-creation work that giving a choice of character characteristics (gender, age, accent etc) is seen as advantageous by many end-users. To minimise any problems with understanding the voice of the IVA, we chose to use recordings of a real person as opposed to synthesised speech. This meant that the IVA's mouth was only synched to the voice in as far as the timing of the full question is concerned. In addition, the laptops loudspeaker was on at a relatively high volume and the participants were sitting in close proximity, directly in front of the computer. All participants could re-play a question, by pushing one of only two buttons from the keyboard to control the avatar.

Participants were given instructions about how to use the software on a laptop; to minimise confusion, most of the keys on the keyboard were covered leaving only labelled ‘next’ and ‘play’ keys visible. 24 participants took part and out of these, a total of 12 recordings of patients interacting with the IVA were further analysed (6 ND, 6 FMD), as we excluded 4 with depressive pseudo dementia, 6 with MCI and 2 in whom the diagnosis was not clear. Audio was recorded using the laptop's built-in microphone. We also recorded two video streams from the built-in webcam and from a webcam positioned to the side of the participants. In the current study, only the audio has been used. As with the neurologist-led conversations, the patient were allowed to bring an accompanying person, if they wanted to.

Table 1 shows an example of the transcript of the recorded responses for two patients (1 ND and 1 FMD) as they attempt to answer the question “Why have you come here today and what are your expectations?” followed by “Tell me what problems have you had with your memory”. Each of the responses show a number of characteristic differences between the two groups. Firstly, the ND patient take a very long pause before starting to answer (15 seconds vs. 1 second). They also use far simpler and shorter answers, and when asked for specific details of their memory problems, the FMD patient is able to give a concrete example describing how words disappear from their vocabulary when in conversation.

Table 1: Samples of conversation between the IVA and the patients from two groups ND (number 07) and FMD (number 10). Pat:Patient, (2 seconds):2 seconds silence.

Group (Patient Code)	**Conversation**
ND (07)	**IVA:** Why have you come in today and what are your expectations? **Pat:** (15 seconds) For a medical trial (2 seconds) um helping diagnosing it better. **IVA:** Tell me what problems have you had with your memory? **Pat:** (18 seconds) I lost a year or more of my memory. (4 seconds) Er, I still have (2 seconds) er gaps in my memory (5 seconds) er, and it just, I can't remember a lot of things.
FMD (10)	**IVA:** Why have you come in today and what are your expectations? **Pat:** (1 second) I've come in er today as a response er for my assistance er in taking this, or assisting with this Avatar er program. **IVA:** Tell me what problems have you had with your memory? **Pat:** Um, I lose words um (2 seconds) they disappear from my vocabulary er when I'm in conversation with people and I find it very difficult er when I speak to people and I lose er, I lose the word, to find, to find another word er to cover for my response.

4.2. Feature extraction

Table 2 lists the extracted features. A total of 44 features (20 CA-based, 12 acoustic, and 12 lexical) were extracted individually for each conversation. For the neurologist-led conversations, they were extracted for participants and named accordingly using prefixes:'Neu', 'Pat' and 'Aps'. For the IVA-led conversations, only the 'Pat' features were used. All the features were extracted automatically using different software/packages. The CA-inspired features were calculated by the NLTK package, lexical features were extracted by the Penn Treebank parser and the acoustic feature by the Praat.

CA inspired features

The primary objective in Elsey et al. (2015) was to define a set of characteristics[1] that would enable the generation of a *diagnostic profile of conversational features* to be drawn up for each patient with the aim of informing the differentiation between ND and FMD; a total of six such characteristics were defined:

- (Role of) "accompanying person" (F1),
- "responding to neurologists' questions about memory problems" (F2),
- "Patient recall of recent memory failure" (F3),
- "responding to compound questions" (F4),
- "inability to answer" (F5),
- "and patient's elaborations and length of turn" (F6).

Similar to the previous studies (Mirheidari et al. (2016, 2017a,b)), a total number of 20 features were extracted from the outputs of the diarisation and the ASR modules, to replicate these qualitative features as closely as possible. The linguistic features were collected using Bag-of-Words (BoW) (Salton, 1983) and the NLTK python library (Bird et al., 2009). For the *conceptual* features, a simple approach of searching for predefined keywords were used. For a full description of the features please refer to Mirheidari et al. (2017b). In addition to the CA inspired features, two more groups of features were extracted from the patients' turns only: lexical (part of speech) and more acoustic features.

Lexical features Penn Treebank part of speech tags (Taylor et al., 2003) were assigned to the words uttered by the patients in the conversations. The number of the Penn Treebank' tags are originally 36, however, similar tags (e.g. different types of verbs) were joined together to make more general tags. The tags were gathered under 12 different groups (Table 2; second row).

Acoustic features Using the 'Praat vocal toolkit' (Boersma et al., 2002), a total of 12 acoustic features were extracted from the audio recordings of the patients in the conversations. We were interested in features that are usually

[1] Elsey *et al.* call these 'features', however to avoid confusion, we will refer to them here as 'profiling characteristics' and use the term 'features' as is conventional in speech technology.

Table 2: List of extracted features. Prefixes:‘Neu’ (neurologist), ‘Pat’ (patient), ‘Aps’ (accompanying person(s).

Category	**Feature**
CA inspired	number of turns (**APsNoOfTurns**, **PatNoOfTurns**, **NeuNoOfTurns**); average length of turn (**APsAVTurnLength**, **PatAVTurnLength**, **NeuAVTurnLength**); number of unique words in a turn (**APsAVUniqueWords**, **PatAVUniqueWords**, **NeuAVUniqueWords**); patient answers “me” for question “who's most concerned” (**PatMeForWhoConcerns**); patient recalls memory failure features (**PatFailureExampleEmptyWords**, **PatFailureExampleAVPauses**, **PatFailureExampleAllTime**); patient replies ‘dunno for the expectation question (**PatDunnoForExpectations**); average number of filler, empty, unique and low-frequency words (**PatAVFillers**), **PatAVEmptyWords**, **PatAVUniqueWords PatAVAllWords**); average number of repeated questions (**AVNoOfRepeatedQuestions**); average number of topics discussed (**AVNoOfTopics**)
Lexical	average number of verbs, nouns, adjectives, adverbs, pronouns, wh_words(e.g, who), determiner conjunctions, cardinals, existential(e.g., there is), prepositions etc(**PatAvgVerb**, **PatAvgNoun**, **PatAvgAdjective**, **PatAvgAdverb**, **PatAvgPronoun**, **PatAvgWh_word** **PatAvgDeterminer**, **PatAvgConjunction**, **PatAvgCardinal**, **PatAvgExistential**, **PatAvgPreposition**, **PatAvgOtherPOS**)
Acoustic	average overall intonation, pitch, duration and silence(**PatAvgIntonation**, **PatAvgPitch**, **PatAvgDuration PatAvgSil**); difference between the first harmonic and the harmonic close to the first, second and third formants(**PatAvgH1-A1**, **PatAvgH1-A2**, **PatAvgH1-A3**); difference between the two first harmonics (**PatAvgH1-H2**); local jitter and shimmer(**PatAvgGitterLocal**, **PatAvgShimmerLocal**); harmonics-to-noise and noise-to-harmonics ratios(**PatAvgMeanHNR**, **PatAvgMeanNHR**)

marked in formal CA transcripts, including the prosodic features (duration, pitch and intonation), creakiness and breathiness (H1-H2 (Gordon and Ladefoged, 2001), H1-A1 (Khan et al., 2015), H1-A2, H1-A3) and vocal stability (jitter, shimmer, harmonics-to-noise and noise-to-harmonics ratios).

5. Results

5.1. Speech recognition and diarisation

As both the neurologist-patient and IVA-patient interactions are largely “natural” and unstructured conversations, ASR and diarisation will be challenging because of additional complexities including having to deal with turn-taking, overlapping speech, prosody, sentence boundaries, coping with dysfluencies or hesitations, other extra non-linguistic information such as emotional content, accelerated speaking rate and sloppy pronunciation (Nakamura et al., 2008; Shriberg, 2005; Moore, 2015).

As shown in Figure 1, diarisation is the first step in the dementia detection system. Table 3 shows the diarisation error (DER). Although the DER is the most common error measure for diarisation procedures, it is only based on the

Table 3: Diarisation error (including missing speaker:E_{MISS}, false alarm:E_{FA}, and speaker error:E_{SPKR})

$Data$	E_{MISS}	E_{FA}	E_{SPKR}	DER	$WDER$
HUM_dia	2.7%	14.9%	12.8%	30.4%	5.7%
IVA_dia	11.6%	6.9%	11.1%	29.6%	16.8%

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

This measure is based on the ratio of the number of words not assigned to the right speakers, to the total number of words (we refer to this measure as word diarisation error, $WDER$). This can be calculated by adding two errors (equation 1): the ratio of the number of missing words to the total words (missing words ratio, E_{MWR}) and the ratio of the number of words assigned to the wrong speakers (words assigned to the wrong speakers ratio, E_{WAWSR}):

$$WDER = E_{MWR} + E_{WAWSR} \tag{1}$$

Table 3 shows very similar $DERs$ for the neurology-led (HUM) and IVA-led (IVA) datasets but with some discrepancy seen in the $WDER$ where the IVA data has a higher error rate than HUM (16.8% vs. 5.7%).

The speaker-turn split data output by the diarisation module is then sent to the ASR. Table 4 presents the results of training various ASR models on the HUM and IVA conversations using the Kaldi toolkit (Povey et al., 2011) (acoustic model: SAT trained HMM-GMM; language model: based on training dataset using Kneser smoothing). In both cases, because of the limited amount of data available, we have maximised the use of the data by using a leave-one-out approach whilst at all times making sure that models are tested and adapted using only the respective test sets contained the conversations with appropriate diagnoses as described in Section 4.1. The HUM baseline system achieves a WER of 55.7% (first row in Table 4). In comparison, the straightforward IVA baseline system (trained only on IVA data) achieves a much higher WER of 77.0% (second row). This is mainly due to there being relatively little data and, in fact, the third row shows that the HUM only model is a better match for the IVA data. The remaining rows shows how MAP adaptation brings down the results to 58.7% and combining the data achieves WER = 46.2%. As can be seen, the WER is relatively high, which reflects the very challenging nature of the recordings (background noise, quality of the microphones, distance to the microphones, etc.) as well as the nature of spontaneous speech (very unique and person-specific language, a lot of repair, emotional speech, overlap and disfluencies).

Table 4: Speech recognition results.

System	Train	Test	WER
Baseline_HUM	HUM	HUM	55.7%
Baseline_IVA	IVA	IVA	77.0%
Cross domain	HUM	IVA	65.0%
MAP adaptation	Map on IVA	IVA	58.7%
Combining data	HUM+IVA	IVA	**46.2%**

5.2. Effect of feature type and selection

Feature selection is a common method for futher improving upon standard classifier pipelines. We investigated the significance of each feature using the standard T-test. Table 5 shows the list of features with p-values < 0.05. Of the 44 features in total, 34 were found to be normally distributed (using the Kolmogorov-Smirnov test) and of those 34, 12 had significant differences between the two groups of patients (FMD and ND). Although the majority of features are based on the patient ('Pat') as expected, there is one features that is based on the AP, namely the number of turns of the accompanying person, indirectly indicating how dominating they have had to be in the conversation.

Table 5: Features with significant difference between two groups of patients:FMD and ND (T test; p-value < 0.05)

Feature	Type	p-value
PatAvgDuration	Acoustic	0.016759
PatAvgSil	Acoustic	0.021879
PatAVEmptyWords	CA Inspired	0.020121
PatAVFillers	CA Inspired	0.001158
PatAVAllWords	CA Inspired	0.000062
APsNoOfTurns	CA Inspired	0.006843
PatAVUniqueWords	CA Inspired	0.000007
PatAVTurnLength	CA Inspired	0.000395
PatAvgNoun	Lexical	0.045126
PatAvgWh_word	Lexical	0.025126
PatAvgDeterminer	Lexical	0.000567
PatAvgConjunction	Lexical	0.01631

Selecting based on p-value is common, but when having access to the full classifier pipeline, basing the selection on the classifier outcome is doable. Table 6 shows the effect of feature selection on the classification using the top

10 features as chosen by the recursive feature elimination (RFE) approach (Pedregosa and Varoquaux, 2011) on the HUM data with manual transcripts. These features are then used for all other datasets. RFE finds the most important features by examining the effect on the classification accuracy of eliminating one feature at a time until all features have been eliminated. The features making the smallest contributions are eliminated recursively until the desired number of features are left.

Like for the T-set based features, APsNoOfTurns has been picked, and is actualy ranked first as the most significant. In addition, and ranked third, is the number of unique words used by the neurologist. This is interesting, as it indicates that, during the conversation, and despite being asked to stick to the given list of questions, the way the neurologist asks the questions (here measured through choice of unique words) is indicating that they speak differently to the patient depending on the ND/FMD group, despite not knowing at the beginning of the interviewe, what the diagnosis is. This gives further motivation to the validity of using an IVA-based system as it will be un-biased.

Table 6: Top 10 important features for the SVM classifier.

Rank	**Feature**	**Type**
1	APsNoOfTurns	CA Inspired
2	PatMeForWhoConcerns	CA Inspired
3	NeuAVUniqueWords	CA Inspired
4	PatFailureExampleAVPauses	CA Inspired
5	PatAvgSil	Acoustic
6	PatAvgWh_word	Lexical
7	PatAvgExistential	Lexical
8	PatAvgAdverb	Lexical
9	PatAvgH1-H2	Acoustic
10	PatAvgPitch	Acoustic

For both feature selection methods, a mix of CA, acoustic and lexical features are picked, with slightly more CA-based features for both cases.

5.3. Classification results

Table 7 shows the classification results for the HUM and IVA conversations for the individual groups of features, for all features, when applying the RFE top selected features, and when using the top selected features based on the T-test. All the classifications tasks were carried out using the "leave-one-out" cross validation method. _man indicates that the data partition (train or test) was using manual transcripts (as opposed to those from an ASR). Each row indicates different levels of automisation of the system - with and without making use of the manual transcripts (indicated as e.g., HUM_man for the neurologist-based conversations with manual transcripts). Looking at the effect of using the different types of features individually (columns 3-5) shows that, depending on the degree of automatising, different types of features are more useful. For example, when using the manual transcripts for IVA (2nd row in Table 7), we get the highest results using the lexical features but their usefulness drops as the ASR transcripts (with associated recognition errors) are introduced. The best result for the IVA conversations (90.9%) is obtained when using all features, which is in line with (or even slightly above) what we achieved for the HUM conversations. It is clear that replacing the neurologist with an IVA can still elicit conversations in which signs of dementia are present and detectable. The reported results are carried out on a relatively small set of conversations which warrants some caution when generalising.

Looking at the feature selection, shows that selecting based on the HUM_man does not always ensure the highest classification rate for each system. In future evaluation, where more data has been collected, we plan to replicate these experiments but using held-out data on which to select the features.

Table 7: Classification accuracy; ”man” = gold-standard transcript instead of ASR-produced transcripts for the respective train or test partition. CA = CA-style features; AC = acoustic features and LX = lexical features.

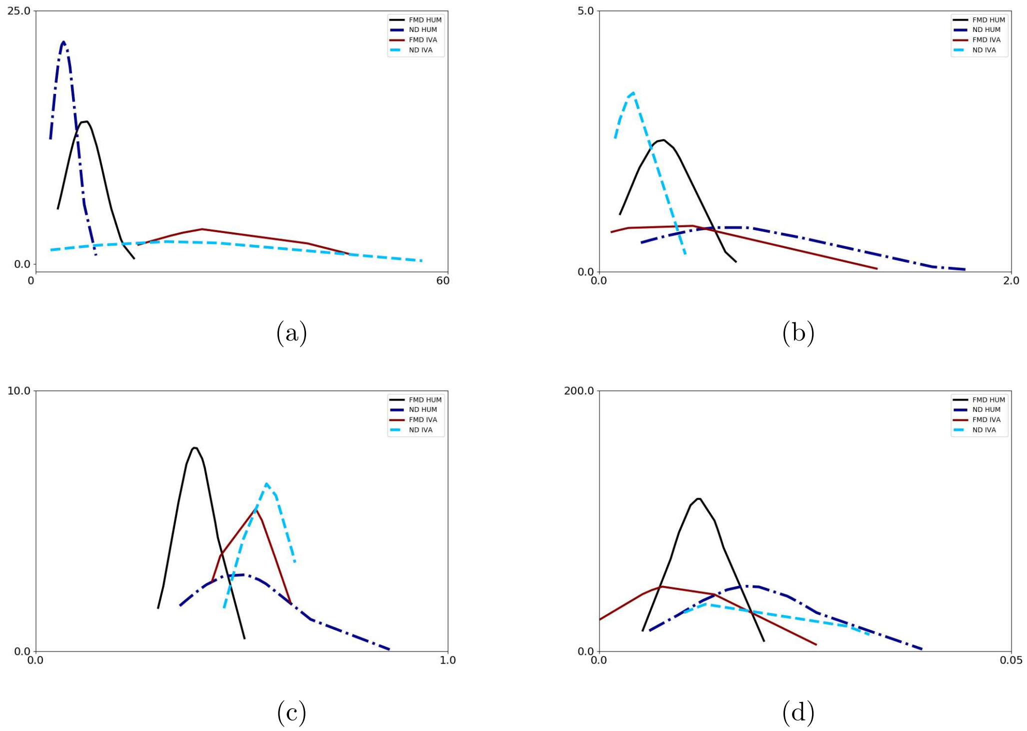
Train	Test	CA(20)	AC(12)	LX(12)	ALL(44)	RFE(10)	T-test(12)
HUM_man	HUM_man	**96.7**%	83.3%	66.7%	76.7%	**100**%	90.0%
HUM	HUM	76.7%	60.0%	50.0%	76.7%	**90.0**%	80.0%
IVA_ man+HUM_man	IVA_man	58.3%	66.7%	**83.3**%	66.7%	75.0%	73.8%
IVA_man+HUM_man	IVA	72.7%	63.6%	63.6%	**81.8**%	72.7%	76.2%
IVA+HUM_man	IVA	63.6%	54.5%	63.6%	**90.9**%	72.7%	81.4%

5.4. Comparing neurology to IVA-led conversations

Some differences were observed between the IVA and HUM data. Figure 3 shows a plot of four measures plotted for the ND and FMD groups and for the IVA and HUM data respectively.

Looking at the distribution of the average length of the turns (Figure 3(a)), in both datasets the patients speaking to the neurologist had shorter turns than when speaking to the IVA. However, overall the IVA conversations had much longer turns, which is likely to be because this initial IVA provides no feedback to the patients in the form of nods, clarifying questions or back-channel noises to steer the conversation. As a result, some patients chose to give very lengthy responses to some of the questions.

The average silence plotted in Figure 3(b) shows a different picture. The least silence is observed for the ND-IVA group and the most for the ND-HUM group. This may be a result of the neurologists being instructed to wait much



(a)

(b)

(c)

(d)

Figure 3: (a) Distribution of the average turn length (in seconds). (b) Distribution of the average silence (in seconds). (c) Distribution of the average overall duration (in seconds). (d) Distribution of the average number of conjunctions.

longer than would normally be expected for the patients to provide an answer. When working with the IVA, the patients always had the option of clicking 'next' and moving the IVA on to the next question, which it looks like many may have chosen to do quite readily, when feeling unable to give a satisfactory answer.

Figure 3(c) shows the average overall duration of the conversation. Despite the average turn length (a), appearing to be quite a discriminative measure, this is less clearly so. Finally, Figure 3(d) shows the use of conjunctions. Here, the FMD groups show signs of using slightly fewer. Perhaps a sign of them speaking more coherently and in full sentences.

6. Conclusions

Spoken communication of people with dementia can be affected in the early stages of neurodegenerative memory disorders, and it is evident that expert analysis using the qualitative methodology of conversation analysis of neurologist-patient interactions can provide diagnostic clues for clinicians. However, the

CA process depends on highly skilled experts in interaction, takes time and is costly. This paper introduces a novel, easy-to-use automatic screening method based on computer-aided processing coupled with an intelligent virtual agent (IVA) front-end for asking memory-probing questions of the patient. The aim is to assess the conversational behaviour of patients with memory problems. We have suggested an automatic dementia detection system including a diarisation unit, an automatic speech recogniser, a CA-based acoustic and lexical feature extraction module and a machine learning classifier that can facilitate and improve screening procedures for dementia.

Parallel corpora of audio recordings of neurologist or IVA-led conversation with a patient with/without an accompanying person, are given to the system which determines whether the conversation is with a patient with functional memory disorder (FMD) or a patient with neurodegenerative dementia (ND).

We have demonstrated the feasibility of using an automated IVA to screen or stratify patients with cognitive complaints. We explored the effect of using different types of features as well as different feature selection methods. Overall, classification accuracies of 90.0% for the neurology-led and 90.9% for the IVA-led conversations were obtained. Analysing various measures in details, such as speaker turn length and amount of silence in the conversations revealed some differences between the kind of interactions the patients were having with neurologists and with the IVA.

The subjective survey at the end of the experiment showed that the overall participants' feedback about the IVA was very positive with a high level of satisfaction - some patients even indicated that they would prefer an automated test like this as it would feel less intimidating.

This is an essential first step for developing a low-cost tool for the early detection of dementia as well as for identifying deterioration over time in people with pre-clinical or prodromal dementia.

Future work will deploy a more interactional IVA to improve on the naturalness of the conversation. We will also deploy the system in more memory clinic settings as part of a larger proof-of-concept study. Future studies will also include additional participants within both the FMD and ND diagnostic categories as well as MCI and healthy controls.

References

Bayles, K. A., 2003. Effects of working memory deficits on the communicative functioning of alzheimers dementia patients. Journal of Communication disorders 36 (3), 209–219.

Bayles, K. A., Boone, D. R., 1982. The potential of language tasks for identifying senile dementia. Journal of Speech and Hearing Disorders 47 (2), 210–217.

Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python. OReilly Media Inc.

Boersma, P. P. G., et al., 2002. Praat, a system for doing phonetics by computer. Glot international 5.

Carrasco, E., Epelde, G., Moreno, A., Ortiz, A., Garcia, I., Buiza, C., Urdaneta, E., Etxaniz, A., González, M. F., Arruti, A., 2008. Natural interaction between avatars and persons with alzheimers disease. In: International Conference on Computers for Handicapped Persons. Springer, pp. 38–45.

Cyarto, E. V., Batchelor, F., Baker, S., Dow, B., 2016. Active ageing with avatars: a virtual exercise class for older adults. In: Proceedings of the 28th Australian Conference on Computer-Human Interaction. ACM, pp. 302–309.

Elsey, C., Drew, P., Jones, D., Blackburn, D., Wakefield, S., Harkness, K., Venneri, A., Reuber, M., 2015. Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics. Patient Education and Counseling 98, 1071–1077.

Forbes-McKay, K. E., Venneri, A., 2005. Detecting subtle spontaneous language decline in early alzheimers disease with a picture description task. Neurological sciences 26 (4), 243–254.

Fraser, K. C., Meltzer, J. A., Rudzicz, F., 2015. Linguistic Features Identify Alzheimer s Disease in Narrative Speech. Journal of Alzheimer's Disease 49, 407–22.

Gordon, M., Ladefoged, P., 2001. Phonation types: a cross-linguistic overview. Journal of Phonetics 29 (4), 383–406.

Gosztolya, G., Tóth, L., Grósz, T., Vincze, V., Hoffmann, I., Szatlóczki, G., Pákáski, M., Kálmán, J., 2016. Detecting Mild Cognitive Impairment from Spontaneous Speech by Correlation-Based Phonetic Feature Selection. Interspeech 2016, 107–111.

Hamilton, H. E., 1994. Conversations with an Alzheimers patient: An interactional sociolinguistic study. Cambridge, England: Cambridge University Press.

Hayward, M., Jones, A.-M., Bogen-Johnston, L., Thomas, N., Strauss, C., 2017. Relating therapy for distressing auditory hallucinations: A pilot randomized controlled trial. Schizophrenia research 183, 137–142.

Hessler, J., Brönner, M., Etgen, T., Ander, K.-H., Förstl, H., Poppert, H., Sander, D., Bickel, H., 2014. Suitability of the 6cit as a screening test for dementia in primary care patients. Aging & mental health 18 (4), 515–520.

Huijbregts, M., 2008. Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled. Ph.D. thesis, University of Twente, The Netherlands.

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

Jones, D., Drew, P., Elsey, C., Blackburn, D., Wakefield, S., Harkness, K., Reuber, M., 2015. Conversational assessment in memory clinic encounters: interactional profiling for differentiating dementia from functional memory disorders. Aging & Mental Health 7863, 1–10.

Jurafsky, D., Martin, J. H., 2008. Speech and language processing, 2nd Edition. Prentice Hall.

Karunanayaka, P., Martinez, B., Eslinger, P. J., Yang, Q. X., 2017. Olfactory processing is highly cognitively demanding: Sensitive functional marker for cognitive deficits and dementia in ad. Alzheimer's & Dementia: The Journal of the Alzheimer's Association 13 (7), P1118–P1119.

Khan, S. u. D., Becker, K., Zimman, L., 2015. The acoustics of perceived creaky voice in american english. The Journal of the Acoustical Society of America 138 (3), 1809–1809.

Lafaille-Magnan, M.-E., Madjar, C., Hoge, R., Breitner, J. C., 2015. Olfactory identification correlates with cerebral blood flow in cognitively normal adults at risk of alzheimers dementia. Alzheimer's & Dementia: The Journal of the Alzheimer's Association 11 (7), P160–P161.

Laske, C., Sohrabi, H. R., Frost, S. M., López-de Ipiña, K., Garrard, P., Buscema, M., Dauwels, J., Soekadar, S. R., Mueller, S., Linnemann, C., et al., 2015. Innovative diagnostic tools for early detection of alzheimer's disease. Alzheimer's & dementia: the journal of the Alzheimer's Association 11 (5), 561–578.

Leff, J., Williams, G., Huckvale, M., Arbuthnot, M., Leff, A. P., 2014. Avatar therapy for persecutory auditory hallucinations: what is it and how does it work? Psychosis 6 (2), 166–176.

Lerner, G. H., 2004. Conversation Analysis: studies from the first generation. Amsterdam John Benjamins Pub.

Lopez de Ipina, K., Alonso, J.-B., Travieso, C. M., Sole-Casals, J., Egiraun, H., Faundez-Zanuy, M., Ezeiza, A., Barroso, N., Ecay-Torres, M., Martinez-Lage, P., Martinez de Lizardui, U., 2013. On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. Sensors 13, 6730–45.

Lopez de Ipina, K., Sole-Casals, J., Eguiraun, H., Alonsod, J. B., Travieso, C. M., Ezeiza, A., Barrosoa, N., Ecay Torres, M., Martinez Lage, P., Beitia, B., 2015. Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach. Computer Speech and Language 30, 43–60.

Mirheidari, B., Blackburn, D., Harkness, K., Walker, T., Venneri, A., Reuber, M., Christensen, H., 2017a. An avatar-based system for identifying individuals likely to develop dementia. Proc. Interspeech 2017, 3147–3151.

Mirheidari, B., Blackburn, D., Harkness, K., Walker, T., Venneri, A., Reuber, M., Christensen, H., 2017b. Toward the automation of diagnostic conversation analysis in patients with memory complaints. Journal of Alzheimer's Disease (Preprint), 1–15.

Mirheidari, B., Blackburn, D., Reuber, M., Walker, T., Christensen, H., 2016. Diagnosing people with dementia using automatic conversation analysis. In: Proceedings of Interspeech. ISCA, pp. 1220–1224.

Moore, R. J., 2015. Automated Transcription and Conversation Analysis. Research on Language and Social Interaction 48 (3), 253–270.

Nakamura, M., Iwano, K., Furui, S., 2008. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. Computer Speech and Language 22 (2), 171–184.

Parsons, S., Rego, D. M., Shawe-Taylor, J., Firth, N. C., Primativo, S., Crutch, S. J., Shakespeare, T. J., Slattery, C. F., Macpherson, K., Carton, A. M., et al., 2017. Modelling eye-tracking data to discriminate between alzheimer's patients and healthy controls. Alzheimer's & Dementia: The Journal of the Alzheimer's Association 13 (7), P597–P598.

Pedregosa, F., Varoquaux, G., 2011. Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12, 2825–2830.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., Dec. 2011. The kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, iEEE Catalog No.: CFP11SRW-USB.

Rizzo, A. A., Lucas, G. M., Gratch, J., Stratou, G., Morency, L.-P., Chavez, K., Shilling, R., Scherer, S., 2016. Automatic behavior analysis during a clinical interview with a virtual human. In: MMVR. pp. 316–322.

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., Kaye, J., 2011. Spoken language derived measures for detecting mild cognitive impairment. IEEE transactions on audio, speech, and language processing 19 (7), 2081–2090.

Rus-Calafell, M., Gutiérrez-Maldonado, J., Ribas-Sabaté, J., 2014. A virtual reality-integrated program for improving social skills in patients with schizophrenia: a pilot study. Journal of behavior therapy and experimental psychiatry 45 (1), 81–89.

Salton, G., 1983. Introduction to modern information retrieval. New York, London, McGraw-Hill.

Satt, A., Sorin, A., Toledo-Ronen, O., Barkan, O., Kompatsiaris, I., Kokonozi, A., Tsolaki, M., 2013. Evaluation of speech-based protocol for detection of early-stage dementia. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech, 1692–1696.

Schmidtke, K., Pohlmann, S., Metternich, B., 2008. The syndrome of functional memory disorder: definition, etiology, and natural course. Am J Geriatr Psychiatry 16, 981–8.

Shriberg, E., 2005. Spontaneous speech: How people really talk and why engineers should care. Interspeech, 1781–1784.

Sidnell, J., Stivers, T., 2012. The Handbook of Conversation Analysis. Wiley-Blackwell.

Stringer, G., Couth, S., Brown, L., Montaldi, D., Gledson, A., Mellor, J., Sutcliffe, A., Sawyer, P., Keane, J., Bull, C., et al., 2018. Can you detect early dementia from an email? a proof of principle study of daily computer use to detect cognitive and functional decline. International journal of geriatric psychiatry.

Tanaka, H., Adachi, H., Ukita, N., Kudo, T., Satoshi, N., 2016. Automatic detection of very early stage of dementia through spoken dialog with computer avatars. In: IEEE Engineering in Medicine and Biology Society.

Taylor, A., Marcus, M., Santorini, B., 2003. The penn treebank: an overview. In: Treebanks. Springer, pp. 5–22.

Thomas, C., Keselj, V., Cercone, Rockwood, K., Asp, E., 2005. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. Proceedings of the IEEE International Conference on Mechatronics & Automation, 1569–1574.

Tomoeda, C. K., Bayles, K. A., Trosset, M. W., Azuma, T., McGeagh, A., 1996. Cross-sectional analysis of alzheimer disease effects on oral discourse in a picture description task. Alzheimer Disease & Associated Disorders 10 (4), 204–215.

Toth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatloczki, G., Biro, E., Zsura, F., Pakaski, M., Kalman, J., 2015. Automatic detection of mild cognitive impairment from spontaneous speech using ASR. Interspeech.

Tran, M. K. P., Robert, P., Bremond, F., 2016. A virtual agent for enhancing performance and engagement of older people with dementia in serious games. In: Workshop Artificial Compagnon-Affect-Interaction 2016.

Weiner, J., Herff, C., Schultz, T., 2016. Speech-Based Detection of Alzheimerś Disease in Conversational German. Interspeech 2016, 1938–1942.

Yancheva, M., Fraser, K., Rudzicz, F., 2015. Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. 6th Workshop on Speech and Language Processing for Assistive Technologies.

Zhang, Y., Wilcockson, T., Kim, K. I., Crawford, T., Gellersen, H., Sawyer, P., 2016. Monitoring dementia with automatic eye movements analysis. In: Intelligent Decision Technologies 2016. Springer, pp. 299–309.

Zhou, L., Fraser, K. C., Rudzicz, F., 2016. Speech recognition in Alzheimerś disease and in its assessment. Proceedings of the 17th Annual Meeting of the International Speech Communication Association (Interspeech), 1948–1952.