CLINICAL LINGUISTICS & PHONETICS
2021, VOL. 35, NO. 3, 237–252
https://doi.org/10.1080/02699206.2020.1777586

Taylor & Francis
Taylor & Francis Group

Characterising spoken responses to an intelligent virtual agent by persons with mild cognitive impairment

Gareth Walker[a], Lee-Anne Morris[b], Heidi Christensen[c], Bahman Mirheidari[c], Markus Reuber[d], and Daniel J. Blackburn[b]

[a]School of English, University of Sheffield, Sheffield, UK; [b]Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK; [c]Department of Computer Science, University of Sheffield, Sheffield, UK; [d]Academic Neurology Unit, Royal Hallamshire Hospital, University of Sheffield, Sheffield, UK

ABSTRACT

The diagnosis of Mild Cognitive Impairment (MCI) characterises patients at risk of dementia and may provide an opportunity for disease-modifying interventions. Identifying persons with MCI (PwMCI) from adults of a similar age without cognitive complaints is a significant challenge. The main aims of this study were to determine whether generic speech differences were evident between PwMCI and healthy controls (HC), whether such differences were identifiable in responses to recent or remote memory questions, and to determine which speech variables showed the clearest between-group differences. This study analysed recordings of 8 PwMCI (5 females, 3 males) and 14 HC of a similar age (8 females, 6 males). Participants were recorded interacting with an intelligent virtual agent: a computer-generated talking head on a computer screen which asks pre-recorded questions when prompted by the interviewee through pressing the next key on a computer keyboard. Responses to recent and remote memory questions were analysed. Mann–Whitney U tests were used to test for statistically significant differences between PwMCI and HC on each of 12 speech variables, relating to temporal characteristics, number of words produced and pitch. It was found that compared to HC, PwMCI produce speech for less time and in shorter chunks, they pause more often and for longer, take longer to begin speaking and produce fewer words in their answers. It was also found that the PwMCI and HC were more alike when responding to remote memory questions than when responding to recent memory questions. These findings show great promise and suggest that detailed speech analysis can make an important contribution to diagnostic and stratification systems in patients with memory complaints.

ARTICLE HISTORY
Received 29 April 2020
Revised 27 May 2020
Accepted 31 May 2020

KEYWORDS
Phonetics; prosody; speech production measurement; mild cognitive impairment; dementia

Background

Formerly known as the 'grey area' between normal aging and dementia, mild cognitive impairment (MCI) has gained acceptance as a clinical entity (Sachdev et al., 2014). The prevalence of MCI is estimated at 10–20% in adults older than 65 years and the risk of developing MCI increases with age (Langa & Levine, 2014). Identifying persons with MCI (PwMCI) from adults of a similar age without cognitive complaints is a greater challenge than identifying persons with dementia from healthy controls (HC). Although a significant

CONTACT Gareth Walker g.walker@sheffield.ac.uk School of English, University of Sheffield, Sheffield S3 7RA, UK
Supplemental data for this article can be accessed here.

proportion of PwMCI will not progress to dementia even 10 years after diagnosis (Mitchell & Shiri-Feshki, 2009), between 21% and 60% of PwMCI will later convert to a dementing illness, the most common of which is Alzheimer's disease (AD; Yaffe et al., 2006). Given the enormity of disease burden and the uncertain trajectory, the development of low-cost, non-invasive tools for early and reliable identification of MCI – and particularly MCI which converts to AD – is of vast clinical, social and economic importance.

Speech analysis has proven relevance to the identification of PwMCI. Subtle changes are evident in the speech and language of PwMCI (Gosztolya et al., 2019) and these changes have the potential for use as disease biomarkers. These changes, which may not be apparent in normal communicative interactions, are evident upon analysis of spoken discourse tasks (Fleming, 2014). It is theorised that various cognitive control mechanisms outside the language system regulate language processing (Caplan, 1992). Collectively referred to as executive functions, these mechanisms include planning, problem-solving, cognitive flexibility, attention shifting and organisation (Caplan, 1992). Breakdown of these skills may be the source of decline in discourse production seen in PwMCI (Fleming & Harris, 2009). A complex speech discourse production task has been shown to be a sensitive tool for the early detection of MCI compared to both HC and persons with AD (Fleming, 2014). Asgari et al. (2017) found they could distinguish PwMCI from HC purely on linguistic analysis of discourse samples using the Linguistic Inquiry and Word Count (LIWC2001 Inc.) tool with 84% classification accuracy. Given the ease with which speech discourse samples can be obtained, their non-invasive nature, and their scope for repeated sampling for longitudinal analyses, this appears promising indeed. Szatloczki et al. (2015) state that computerised analysis of spontaneous speech in the form of a software package may be promising to screen for MCI and early AD. Traditional tests used in the memory clinic for dementia detection (e.g., the Montreal Cognitive Assessment, MoCA; Nasreddine et al., 2005) only briefly screen language function and do not include language tasks complex enough to detect subtle changes. Also, such tests need a clinician to administer them and have learned effects making widespread and repeated use more difficult.

Speech samples used in automated language studies have typically been obtained by the presentation of stimuli such as pictures, short films, paragraphs or stories. In the current study, language samples from responses to questions concerning recent and remote memory will be analysed. The rationale behind this, based on clinical observations as well as Ribot's law (remote memory is spared to a greater extent than recent memory, evident in MCI and early AD: Müller et al., 2016) is that there may be an interesting and clinically meaningful discrepancy between responses to recent and remote memory questions that tasks such as picture descriptions may miss. The relationship between speech and language and cognitive information processing, including memory, is well illustrated by Cohen et al. (2015). They found that language output changed when HC performed tasks with increasingly high internal processing loads with fewer utterances produced, longer pauses evident, and greater silence overall. They suggest the use of vocal expression as a marker of information processing across and also within pathological individuals over time. Since memory is a domain particularly affected in neurodegenerative cognitive disorders, language differences may be present in these tasks to a greater degree than in non-memory tasks due to the higher demands on internal cognitive processing. In early disease stages, recent memory may be more affected than remote memory due to the temporal gradient; hence, language changes may be present to a greater degree in recent memory tasks.

The clinical relevance of using recent and remote memory questions to elicit language samples for analyses is currently unknown. Few studies to date have examined whether this could be a sensitive measure for detection of MCI. Smolík et al. (2016) found that propositional density of speech by persons with amnestic MCI was lower than HC but only when talking about remote memories (childhood) and not when talking about recent events. Han et al. (2014) reported on the vocal expression of emotion by HC and persons with early-stage AD. They found that emotional expression as judged by independent evaluators was higher when talking about remote memories than recent memories.

Data collection for studies of the speech of PwMCI is typically done manually through interviews or tests administered by a researcher with speech samples recorded and later transcribed. Recent studies have made use of intelligent virtual agents (IVAs): a computer-generated talking head on a computer screen which ask pre-recorded questions when prompted by the interviewee. Tanaka et al. (2016) used such a method to administer a range of tests to persons with early-stage dementia and HC and found that data collected in this way were able to inform the detection of early-stage dementia. Mirheidari, Blackburn, Walker et al. (2019) studied interactions between an IVA and persons with neurodegenerative disease (ND) and persons with a functional memory disorder (FMD). They found significant differences in conversational structure, lexis and acoustic properties between the three groups (i.e. ND, FMD, HC). In a development of that work, Mirheidari et al. (2019) found significant differences between ND, FMD, HC and PwMCI in conversational structure, lexis and acoustic properties. There is evidence of ecological validity of responses to an IVA in a memory clinic context. Walker et al. (2018) analysed responses to an IVA from persons with FMD and persons with dementia: between-group differences of diagnostic relevance were similar to those observed in patient–neurologist interactions.

Hoffmann et al. (2010) proposed that temporal parameters serve as a screening method for early AD. Temporal differences in spontaneous speech such as increased number of pauses and increased pause length have been found to be sensitive markers for the detection of early AD (Szatloczki et al., 2015). Satt et al. (2014) studied recordings of HC, PwMCI and persons with AD performing several spoken tasks (counting backwards, picture description, repeating a sentence and naming animals). They used various temporal features extracted from the recordings in a statistical classifier and reported a classification accuracy of approximately 80% for PwMCI versus HC. The analysis of acoustic features in Beltrami et al. (2018) focussed on temporal measures, which were found to be able to distinguish the pathological groups from the control group and in some cases to be able to distinguish between pathological groups. Pauses differ depending on the type of discourse so the choice of spontaneous speech task is important. Pistono et al. (2019) state that anterograde memory function would predict a patient's pause frequency in a memory-based narrative, as pauses are used as compensatory mechanisms in early AD. Memory-based narratives may thus be most sensitive to any change in pause behaviour in PwMCI. Because the length of participant responses can differ greatly, using pause-to-speech ratio may be a consistent way to measure pause differences across different questions. A higher pause-to-speech ratio means that there is a greater amount of total pause in a participant's answer, compared to the total amount of speech.

The main aims of this study were 1) to determine whether generic speech differences were evident between PwMCI and HC, 2) to determine whether such differences were identifiable in responses to recent or remote memory questions, and 3) to determine which speech variables showed the clearest between-group differences.

Method

Participants

Ethical approval was granted for the study prior to commencement. Consenting PwMCI (n = 8; 5 females, 3 males) were recruited from memory and neuropsychology clinics held at a tertiary hospital in the UK. MCI was diagnosed according to Petersen's criteria (Petersen, 2011) by consultant neurologists. All PwMCI had no other neurological disorders. Consenting HC (n = 14; 8 females, 6 males) were recruited through the University of the Third Age, a society for retired community members; participants all scored within the normal range on cognitive testing (Addenbrooke's Cognitive Examination Revised). All participants were below clinical cut-offs for anxiety and depression as measured by the Generalized Anxiety Disorder questionnaire 7 and the Patient Health Questionnaire 9, respectively, and were first language English speakers. Participants were recruited to the study by convenience sampling; all were white, of British descent, and were raised and schooled in English.

Data collection

Participants interacted with an IVA. In these interactions, the interviewee (IE) answered pre-recorded questions posed by the IVA; when IE pressed a button on the laptop keyboard, the IVA asked the next question or repeated the previous question depending on the button pressed. A researcher was present during each session but instructed not to speak unless they were asked direct questions or if other issues arose. Audio and video recordings were made of the interactions. Data from four of the questions asked by the IVA were analysed. Two of the questions related to recent memory ('What did you do over last weekend, giving as much detail as you can?', 'What has been in the news recently?'), and two of the questions related to remote memory ('Tell me about the school you went to and how old you were when you left.', 'Tell me what you did when you left school. What jobs did you do?'). Such questions are typical of those posed by neurologists during memory clinic appointments. Recent memory questions will be referred to as REC-Q and remote memory questions will be referred to as REM-Q; ALL-Q will be used to refer to both question-types combined.

Preparation of data for analysis

This section describes several aspects to preparing the data for analysis: segmentation of the recordings, the creation of pitch traces, preparation of transcriptions, the nature of the selected speech variables, and the selection of statistical tests.

Segmentation

Segmentation involves identifying boundaries in the speech-stream. This was done using a combination of careful listening and inspection of acoustic records (waveforms and spectrograms). All computer-based analysis of speech was conducted using Praat (Boersma & Weenink, 2020).

The start and end of each question by the IVA were identified through careful listening and inspection of acoustic records. The answer to the question was considered to start and end where audible vocal behaviour in response to the question begins and ends. 'Audible vocal behaviour' includes any sound produced by IE's vocal tract, including, for example, breathing, clicks and percussives, as well as speech. Such noises were included as they could be indications of responsive 'gearing up' to speak. Since such gearing up suggests cognitive processing in response to the question, the timing of that gearing up might be useful in characterising the two groups. Out of 88 answers, there were two where IE responds to the question asked by the IVA and is then prompted to say more by the researcher; in these cases, the end of the answer is taken to be the offset of vocal behaviour prior to the researcher prompting more talk. There was one answer where IE responds to the IVA then begins to talk to the researcher; in this case, the end of the answer is taken to be the offset of vocal behaviour prior to the speech directed to the researcher. One answer suffers a brief interruption by a noise from the computer, but since the interruption is very brief, the segmentation was not altered from the normal procedure.

Within the portions labelled as responses from IE there are periods where there is no audible vocal activity. Praat was used to estimate the location and duration of these periods. A silence threshold was determined to do this. The silence threshold is how far below the maximum intensity in a sample the signal must be in order to be considered silence. Since the recordings vary in several ways (e.g., recording quality, ambient noise, distance between IE and the microphone), a silence threshold was established for each recording. The silence threshold was determined by subtracting the mean intensity value of an audible inbreath by IE from the maximum intensity in the recording. An inbreath was used as little, if any, audible vocal behaviour would have a lower intensity than an inbreath. Provided that the duration criteria for silence detection are met, those parts of IE's response with an intensity between the mean intensity of the identified audible inbreath and the maximum intensity in the recording were marked as 'sounding'; the parts of IE's response which do not satisfy those criteria were marked as 'non-sounding'. 'Sounding' intervals are taken as a proxy for speech from IE and the 'non-sounding' intervals as a proxy for silence. Following experimentation with different values applied to several recordings, the minimum non-sounding (silent) interval duration was set at 0.2 s; no intervals shorter than this could be considered silent. The minimum sounding interval duration was left at the default value of 0.1 s.

There are distinct advantages to using silence detection to identify speech and silence within IE's responses. It is fairly quick requiring only the identification of the beginning and end of IE's answers and measuring the intensity of an inbreath in one of those answers. Once the silence threshold has been determined the method is objective and absolutely consistent. In this context, this method compares favourably with labelling by listening and inspection of acoustic records which is time-consuming and always subjective to some extent.

Figure 1 shows a screenshot of a Praat window after segmentation was complete. The label tier 'ques' identifies the questions by the IVA (the screenshot shows the end of the first recent memory question). The label tier 'IEans' identifies the start and end of IE's responses (the screenshot shows the start of the answer to the first recent memory question). The label tier 'IEsp' contains the output of applying the silence threshold method described above,

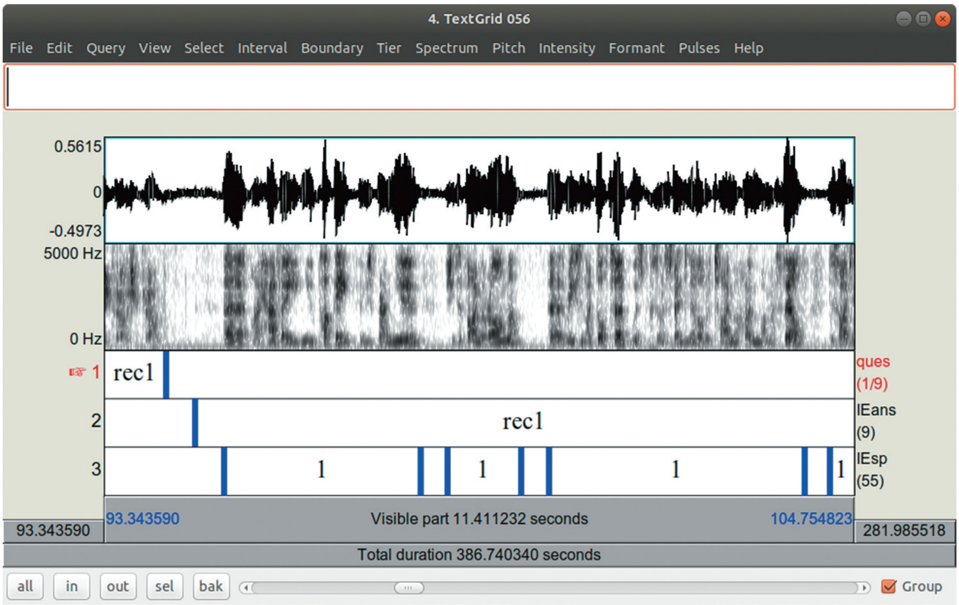
4. TextGrid 056
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
0.5615
0
-0.4973
5000 Hz
0 Hz
1 rec1
2 rec1
3 1 1 1 1
ques (1/9)
IEans (9)
IEsp (55)
93.343590
93.343590
Visible part 11.411232 seconds
104.754823
281.985518
Total duration 386.740340 seconds
all in out sel bak
Group

Figure 1. A screenshot of a Praat editor window showing a waveform (top panel), spectrogram (middle panel), and labels (bottom panel).

with ‘1′ in intervals identified as ‘sounding’ (and treated as speech); intervals identified as ‘non-sounding’ (and treated as silence) are left empty.

Pitch traces

Pitch traces were created with a floor and ceiling set by gender and in accordance with the suggestion in the Praat manual (75–300 Hz for males, 100–500 Hz for females). Pitch traces may be corrected to remove values which do not accurately reflect the rate of vibrations of the speaker’s vocal folds (see Walker, 2017). To reduce the amount of time involved in the preparation of the data for analysis, and to increase the extent to which this study could be replicated, pitch traces have not been corrected. Since these are uncorrected pitch traces they are likely to contain errors, especially at extreme values. For this reason, the measures reported on below only consider values which fall between the 10th and 90th percentiles in the distribution of pitch values.

Transcriptions

Orthographic transcriptions of the recordings prepared by professional transcribers were used to assist in word counts. Fillers (e.g., ‘um’, ‘uh’) were retained as transcribed, as were cut-off words. The number of strings separated by spaces in the transcriptions of speech produced by IEs was taken as a proxy for words.

Speech variables

The speech variables were selected based on several criteria: (a) variables shown to distinguish between PwMCI and HC in the previous research, (b) easily replicable, (c) timeefficient, (d) clinically relevant, namely, potentially discernible by a co-present observer, e.g., in a clinical interview. The variables fall into three broad categories: temporal characteristics (features concerning the duration of speech and/or silence, and measures derived from those features, e.g., speaking and articulation rates), number of words and pitch. Table 1 lists the variables measured for the responses by IEs and which are reported on in this paper, along with descriptions and the units in which measures are reported. Measures of these variables are provided for responses to REC-Q, REM-Q and ALL-Q in later sections and in the supplemental data.

Statistical tests

Statistical tests were performed using *R* (R Core Team, 2020). Mann–Whitney U tests were used to determine whether the measures of each variable had data distributions that were significantly different for PwMCI and HC in responses to ALL-Q, REC-Q and REM-Q. A statistical significance level of 0.05 was used throughout.

Results

The results of Mann–Whitney U tests comparing PwMCI and HC on each variable in responses to ALL-Q, REC-Q and REM-Q are shown in Table 2. Measures of each variable for each IE can be found in the supplemental materials.

Figure 2 shows box and whisker plots for variables where $p < .05$, showing PwMCI and HC for ALL-Q. In the plots, the bottom and top of the box represent the top of the first and third quartiles, respectively; the horizontal line within the box is the median. The whiskers extend up to 1.5 times the interquartile range from the box to reach any values in that range (this is the default in R) and any values lying outside of that range are represented by circles.

Table 1. Speech variables considered; ‘speech’ and ‘silence’ refer to ‘sounding’ and ‘non-sounding’ intervals as identified by silence detection; ‘words’ refers to the number of strings separated by spaces in the orthographic transcriptions.

Variable name	Description	Unit
spDur	Duration of speech	seconds (s)
aveSpDur	Average duration of speech chunks	seconds (s)
aveSilDur	Average duration of silences	seconds (s)
silFreq	Silence frequency, determined by dividing the number of silences by the amount of speech	silence per second (sil/s)
respDur	Duration of speech and silences	seconds (s)
silToSp	Ratio of silence to speech	
delAns	Delay in beginning to answer, determined by measuring the time between the end of the question and the onset of audible vocal behaviour in the answer	seconds (s)
delSp	Delay in beginning to speak, determined by measuring the time between the end of the question and the onset of speech as identified by silence detection	seconds (s)
pRng10to90	Pitch range, determined by calculating the distance between the 10th and 90th percentiles in the distribution of pitch values produced	semitones (ST)
wordCount	Number of words produced	words
speakRate	Speaking rate determined by dividing wordCount by respDur	words per second (words/s)
artRate	Articulation rate determined by dividing wordCount by spDur	words per second (words/s)

Table 2. Results of Mann–Whitney U tests for differences between responses by persons with mild cognitive impairment and healthy controls to all questions (ALL-Q), to questions concerning recent memory (REC-Q), and to questions concerning remote memory questions (REM-Q).

	ALL-Q		REC-Q		REM-Q	
	W	p	W	p	W	p
spDur	90	**.020**	87	**.035**	84	.059
aveSpDur	99	**.002**	102	**.001**	92	**.013**
aveSilDur	15	**.004**	14	**.003**	45	.482
silFreq	19	**.010**	22	**.020**	29	.070
respDur	84	.059	83	.070	80	.110
silToSp	13	**.002**	10	**.001**	34	.145
delAns	18	**.008**	27	.050	25	**.035**
delSp	20	**.013**	23	**.024**	30	.082
pRng10to90	75	.212	71	.330	82	.082
wordCount	88.5	**.029**	88	**.029**	88	**.029**
speakRate	82	.082	82	.082	67	.482
artRate	56	1.000	62	.714	59	.868

Bold indicates *p* <.05.

Discussion

This section discusses the responses to ALL-Q, then to REC-Q and REM-Q. There is then the discussion of some of the limitations to the study and possible avenues for further research.

Responses to all questions

There are significant differences in the distribution of values within responses by PwMCI and HC on eight of the 12 variables in the ALL-Q condition. The statistical differences evident in these data, coupled with inspection of the medians and means on each of these variables, give rise to the following observations:

(1) PwMCI produce speech for less time than HC (spDur)
(2) PwMCI produce speech in shorter chunks than HC (aveSpDur)
(3) The average duration of silences in the responses of PwMCI is longer than HC (aveSilDur)
(4) PwMCI pause more often than HC (silFreq)
(5) PwMCI have a higher pause-to-speech ratio than HC (silToSp)
(6) PwMCI take longer to begin speaking in response to questions than HC (delAns/delSp)
(7) PwMCI produce fewer words in their answers (wordCount)

Several of these results accord with the previous research. Finding (2) accords with Beltrami et al. (2018) who found statistically significant differences in speech segment durations between HC, PwMCI, and persons with early dementia. Finding (3) accords with Szatloczki et al. (2015) and Hoffmann et al. (2010) who found increased pause length to be sensitive markers for the detection of early AD. Finding (4) accords with Szatloczki et al. (2015) who found an increased amount of pauses to be sensitive markers for the detection of early AD.

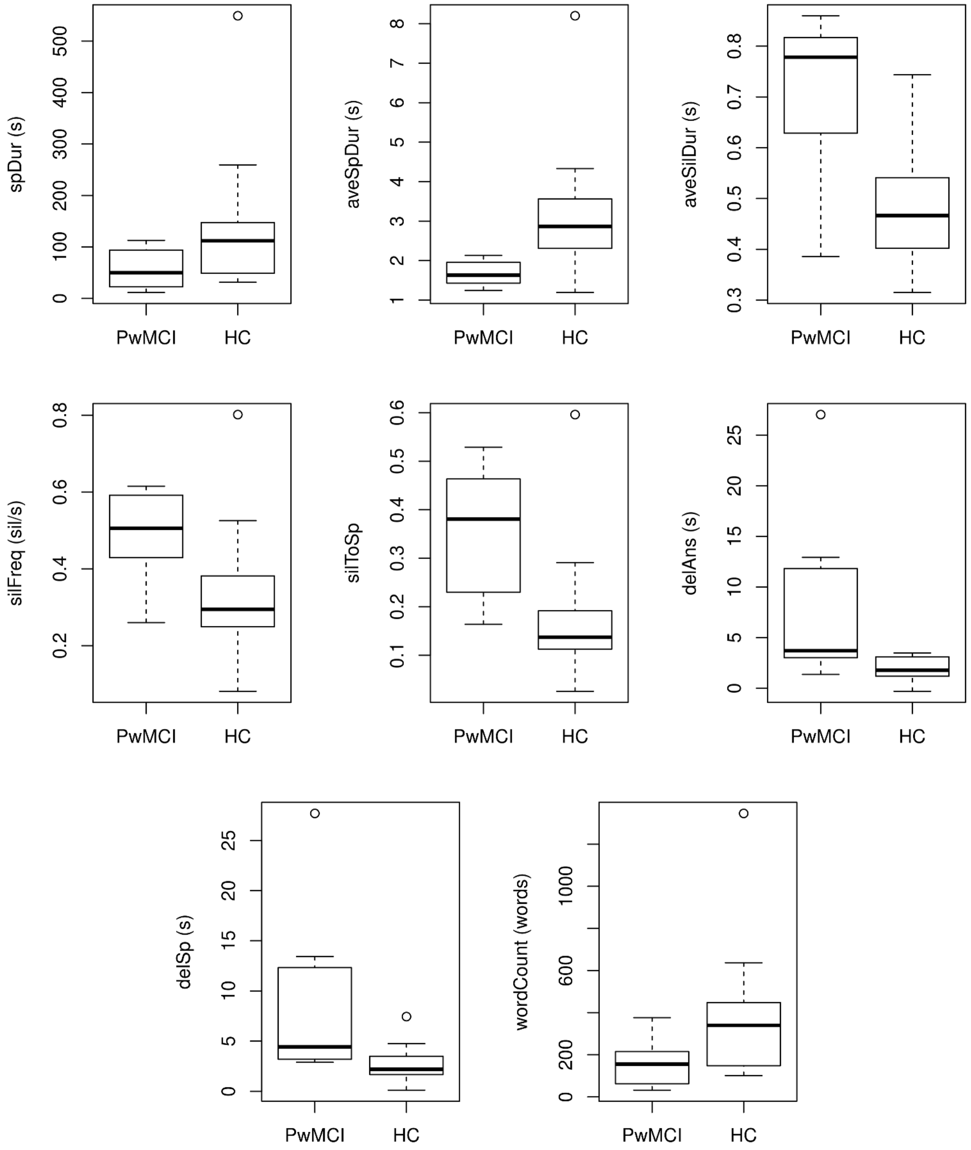

Figure 2. Box and whisker plots for variables which showed a significant difference between responses by persons with Mild Cognitive Impairment (PwMCI) and healthy controls (HC) in Mann–Whitney U tests for recent and remote questions combined (p < .05).

Finding (6) may in part be a reflection of increased information processing speed and reaction times for PwMCI compared with HC (Andriuta et al., 2019; Haworth et al., 2016). Finding (7) accords with Gonzalez-Moreira et al. (2015) who reported that HC produced more syllables than persons with mild dementia in a task consisting of a structured interview and a reading task. (It is extrapolated that increased syllable production results in increased word production, given that words comprise syllables.) In contrast, Roark et al. (2011) and Mueller et al. (2018) found no difference in verbal output between PwMCI and

HC using neuropsychological interviews and picture description tasks, respectively, to elicit speech. Use of memory-related questions to elicit speech may pose a greater cognitive challenge to PwMCI, resulting in reduced verbal output (see studies by Cohen et al., 2014, 2015 for more on cognitive load and verbal output). In contrast, Dodge et al. (2015) found that PwMCI produced a higher proportion of the words spoken in interviews than HC.

There was no significant difference between PwMCI and HC with regard to pitch range (pRng10to90). Previous related research presents a somewhat mixed picture with regard to pitch. Gonzalez-Moreira et al. (2015) reported the mean fundamental frequency for persons with mild dementia (n = 10) to be significantly higher than for HC (n = 10). However, Horley et al. (2010) administered expressive tasks to persons with AD and to HC and found no significant differences between the groups in mean fundamental frequency, but greater pitch modulation was evident for the control group. It is worth noting that the measure of pitch adopted here is quite weak from a technical point of view. The pitch ranges are calculated based on uncorrected pitch traces, created with simple floor and ceiling values which might not have been optimal in all cases. The upper and lower thresholds (10th and 90th percentiles) are somewhat arbitrary based on experience rather than experimental evaluation. Results with greater ecological validity could be arrived at from hand-corrected pitch traces created with floor and ceiling values appropriate for each speaker. However, this would be a time-consuming task and would be more subjective than the method used here.

The non-significant results regarding speaking rate (speakRate) and articulation rate (artRate) are consistent with previous research. Mueller et al. (2018) identified several studies which found no significant differences in speaking rate among groups. Speaking rate and articulation rate are both reflections of motor speech, namely, the physical act of speaking rather than reflections of the 'cognitive' aspect of speaking. Since MCI does not affect motor control, the speaking and articulation rates for PwMCI are expected to be comparable to HC of a similar age.

The lack of significant between-group results for the duration of the response (respDur) was surprising given that there were significant differences in the number of words produced (wordCount) and average silence duration (aveSilDur). However, this variable narrowly missed the threshold for statistically significant difference (p = .059 for all questions). There were clear differences in the distribution of measures on this variable (median = 61.34 s for PwMCI, 155.18 s for HC; mean = 77.35 s for PwMCI, 164.89 s for HC). Differences between the groups were enlarged by one outlier in the HC group (participant 160, 312.22 s), but even with that outlier excluded the median and mean was much higher for HC than PwMCI (if participant 160 is excluded, median = 152.34 s, mean = 153.56 s).

Responses to recent vs remote memory questions

Recent memory

All but one of the eight variables which showed a significant difference between PwMCI and HC in the ALL-Q condition show a significant difference in responses to REC-Q. The variable delAns only narrowly misses out on the threshold for significance (p = .05016). The duration of speech produced by HC is significantly longer than for PwMCI (spDur); the average duration of a speech chunk in the speech of PwMCI is significantly shorter than for HC (aveSpDur); the average silence duration in the speech of PwMCI is significantly longer than for HC (aveSilDur); the silence to speech ratio for PwMCI is significantly higher than for HC (silToSp); and the delay before PwMCI start to speak is significantly longer than for HC (delSp).

Remote memory

There are fewer significant differences between PwMCI and HC in responses to REM-Q than in responses to REC-Q. Measures of four variables (spDur, aveSilDur, SilToSp, delSp) which were significantly different in responses to REC-Q were not significantly different in responses to REM-Q. The reduced difference between PwMCI and HC in responding to REM-Q may reflect that in PwMCI, REM-Q pose less of a cognitive challenge than REC-Q and thus answers are retrieved with greater ease resulting in speech more like that of HC. A higher cognitive processing load may be experienced by PwMCI when answering responding to REC-Q since recent memory is a domain affected early on in the disease course (temporal gradient of memory loss).

However, there is a need for caution in the interpretation of these findings. While fewer variables show significant differences between PwMCI and HC in response REM-Q than in responses to REM-Q, there are still notable differences in the median and mean values. These values are presented in Table 3.

Study limitations

The sample size for the study was small and the participants were not matched for age, gender or level of education. Different methods for measuring the selected variables might have revealed different patterns. For instance, a silence threshold is not perfect at identifying speech. Since Praat cannot easily separate out speech from other kinds of noise, an interval could have been labelled as ‘sounding’ (= speech) on the basis of other background noise. More robust measures of speaking and articulation rate may be possible, albeit more time-consuming, using counts of segments or syllables per second rather than words. Finally, while the variables were selected for reasons described above, other variables could have been selected and may have yielded different insights.

The Addenbrooke’s Cognitive Exam Revised was the only cognitive measure used in the study. Detailed neuropsychology testing on participants could lend depth to the understanding of their cognitive functioning. It would be possible to explore relationships between the speech markers outlined in this study and neuropsychology test scores. Including additional assessments, such as self-report or caregiver-report measures could add additional dimensions to the results of this study. Participants’ awareness of their decline or otherwise, and whether self-report measures and speech performance correlate would be of interest. The inclusion of data from neuroimaging could add valuable information about the underlying neural substrates of the observed behaviours.

Avenues for further research

These preliminary findings show great promise and we recommend further research using memory-related questions to distinguish between PwMCI and HC, as well as determining efficacy to identify those with FMD and dementia. Given that subjective cognitive decline (SCD) may precede a diagnosis of MCI, analysis of speech samples from persons with

Table 3. Comparison of responses to questions concerning recent and remote memory by persons with mild cognitive impairment (PwMCI) and healthy controls (HC).

	Median						Mean					
	Recent			Remote			Recent			Remote		
	PwMCI	HC	diff.	PwMCI	HC	diff.	PwMCI	HC	diff.	PwMCI	HC	diff.
spDur	20.84	56.01	−35.17	23.54	57.84	−34.30	24.31	72.79	−48.48	32.81	68.12	−35.31
aveSilDur	0.93	0.46	0.48	0.58	0.46	0.12	0.87	0.49	0.37	0.57	0.49	0.08
silToSp	0.44	0.14	0.29	0.25	0.14	0.11	0.48	0.19	0.29	0.29	0.17	0.12
delSp	1.94	1.15	0.79	2.15	0.91	1.24	6.52	1.56	4.96	2.29	1.14	1.15

The variables are those which show statistically significant differences in responses to questions concerning recent memory but not in responses to questions concerning remote memory: see Table 2. Columns headed ‘diff.’ show the differences between PwMCI and HC (PwMCI–HC).

subjective cognitive decline would also be of value. Larger samples are needed to help to establish how widespread the identified patterns might be. Future research investigating differing speech patterns in various dementias could be of immense value in a clinical context, either by utilizing automatic speech analysis tools to aid with differential diagnosis or by training clinicians to listen for some of the delineated patterns. Further speech variables might be studied, though these should be considered against the criteria for variable selection set out above to help ensure that the selected variables reflect vocal behaviour in meaningful ways.

There may be qualitative differences between the responses from participants in the two groups. Lunsford and Heeman (2015) compared how a recently told story is retold by PwMCI and HC and found that PwMCI spent significantly more time in verbal hesitations (e.g., ‘uh’, ‘um’, ‘let’s see’) than HC, and that verbal hesitations accounted for a higher proportion of PwMCI’s speaking time than that of HC. Lunsford and Heeman (2015) also found that when retelling a recently told story, a greater proportion of PwMCI used phrases such as ‘I guess’, ‘I think it was’, ‘I can’t remember’ to mark uncertainty than HC.

The focus of this study has been on how participants speak rather than on what they say. There has been no consideration of how much information the participants give, whether answers are accurate or whether all parts of the question are addressed. It is notable, for instance, that in response to the first question relating to remote memory, one participant with MCI describes the school he went to but not how old he was when he left. An approach following the principles of Conversation Analysis (CA) seems a good way forward in this respect. Walker et al. (2018) engage in fine-grained analysis of conversational structure, finding that diagnostically relevant features can be observed when persons with FMD and ND interact with an IVA.

There has been no consideration of visual information (e.g., gaze, posture, gesture) which is captured in the video recordings of the interactions. There is some evidence of the relevance of visible bodily behaviour to the differentiation between PwMCI and HC: Shinkawa et al. (2019) found that when a classification model combined measures of speech (lexis and syntax) with measures of gait, classification accuracy (PwMCI versus HC) improved when compared with models based on one modality.

While there has been some statistical analysis there has not been any attempt at statistical classification of the two groups, though this has been done with some success in other studies (e.g., Kato et al., 2015; König et al., 2015; Mirheidari, Blackburn, Walker et al., 2019;

Roark et al., 2011; Tóth et al., 2018). It seems relevant to such a study that Figure 2 shows that there are outliers on most of the variables shown; this suggests that any classification would require measures of a package of variables.

Conclusions

This promising study has shown that there are clear differences in the speech patterns of PwMCI and HC when responding to memory-related questions asked by an IVA. These differences are reflected in the amount of time respondents speak for (PwMCI<HC), the length of the speech chunks (PwMCI<HC), the average duration of silences (PwMCI>HC), the frequency of silences (PwMCI>HC), the pause-to-speech ratio (PwMCI>HC), the length of time it takes to begin a response (PwMCI>HC) and the number of words produced in answers (PwMCI<HC). There are differences in the way that persons in the two groups respond to questions concerning recent memory and questions concerning remote memory. The highest number of variables exhibiting significant differences between PwMCI and HC occurs when all questions are included, closely followed by recent memory questions, with remote memory questions having the fewest variable which exhibit significant differences between PwMCI and HC. It is proposed that recent memory questions may have particular clinical utility in distinguish between PwMCI and HC. It has also been shown that answers to memory-related questions posed by an IVA can reveal differences in the speech characteristics of PwMCI and HC.

Acknowledgments

This research was supported/funded by the NIHR Sheffield Biomedical Research Centre (BRC)/ NIHR Sheffield Clinical Research Facility (CRF). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care (DHSC). We would like to thank all the patients and their families for participating in this study.

ORCID

Gareth Walker http://orcid.org/0000-0001-5022-4756
Bahman Mirheidari http://orcid.org/0000-0002-7797-2778
Markus Reuber http://orcid.org/0000-0002-4104-6705
Daniel J. Blackburn http://orcid.org/0000-0001-8886-1283

Declaration of interest

The authors report no conflict of interest.

References

Andriuta, D., Diouf, M., Roussel, M., & Godefroy, O. (2019). Is reaction time slowing an early sign of Alzheimer's disease? A meta-analysis. *Dementia and Geriatric Cognitive Disorders*, *47*(4–6), 281–288. https://doi.org/10.1159/000500348

Asgari, M., Kaye, J., & Dodge, H. (2017). Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's and Dementia: Translational Research and Clinical Interventions*, *3* (2), 219–228. https://doi.org/10.1016/j.trci.2017.01.006

Beltrami, D., Gagliardi, G., Favretti, R. R., Ghidoni, E., Tamburini, F., & Calzà, L. (2018). Speech analysis by natural language processing techniques: A possible tool for very early detection of cognitive decline? *Frontiers in Aging Neuroscience*, *10*, 1–13. https://doi.org/10.3389/fnagi.2018.00369

Boersma, P., & Weenink, D. (2020). Praat: Doing phonetics by computer [Computer software]. http://www.praat.org/

Caplan, D. (1992). *Language: Structure, processing, and disorders*. MIT Press.

Cohen, A. S., Dinzeo, T. J., Donovan, N. J., Brown, C. E., & Morrison, S. C. (2015). Vocal acoustic analysis as a biometric indicator of information processing: Implications for neurological and psychiatric disorders. *Psychiatry Research*, *226*(1), 235–241. https://doi.org/10.1016/j.psychres.2014.12.054

Cohen, A. S., McGovern, J. E., Dinzeo, T. J., & Covington, M. A. (2014). Speech deficits in serious mental illness: A cognitive resource issue? *Schizophrenia Research*, *160*(1–3), 173–179. https://doi.org/10.1016/j.schres.2014.10.032

Dodge, H. H., Mattek, N., Gregor, M., Bowman, M., Seelye, A., Ybarra, O., Asgari, M., & Kaye, J. A. (2015). Social markers of mild cognitive impairment: Proportion of word counts in free conversational speech. *Current Alzheimer Research*, *12*(6), 513–519. https://doi.org/10.2174/1567205012666150530201917

Fleming, V. B. (2014). Early detection of cognitive-linguistic change associated with mild cognitive impairment. *Communication Disorders Quarterly*, *35*(3), 146–157. https://doi.org/10.1177/1525740113520322

Fleming, V. B., & Harris, J. L. (2009). Test–retest discourse performance of individuals with mild cognitive impairment. *Aphasiology*, *23*(7–8), 940–950. https://doi.org/10.1080/02687030802586480

Gonzalez-Moreira, E., Torres-Boza, D., Arturo Kairuz, H., Ferrer, C., Garcia-Zamora, M., Espinoza-Cuadros, F., & Alfonso Hernandez-Gomez, L. (2015). Automatic prosodic analysis to identify mild dementia. *BioMed Research International*, *2015*, 1–6. https://doi.org/10.1155/2015/916356

Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., & Hoffmann, I. (2019). Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Computer Speech & Language*, *53*, 181–197. https://doi.org/10.1016/j.csl.2018.07.007

Han, K.-H., Zaytseva, Y., Bao, Y., Chung, E. P. S. Y., Kim, J. W., & Kim, H. T. (2014). Impairment of vocal expression of negative emotions in patients with Alzheimer's disease. *Frontiers in Aging Neuroscience*, *6*, 1–6. https://doi.org/10.3389/fnagi.2014.00101

Haworth, J., Phillips, M., Newson, M., Rogers, P. J., Torrens-Burton, A., & Tales, A. (2016). Measuring information processing speed in mild cognitive impairment: Clinical versus research dichotomy. *Journal of Alzheimer's Disease*, *51*(1), 263–275. https://doi.org/10.3233/JAD-150791

Hoffmann, I., Nemeth, D., Dye, C. D., Pakaski, M., Irinyi, T., & Kalman, J. (2010). Temporal parameters of spontaneous speech in Alzheimer's disease. *International Journal of Speech-language Pathology*, *12*(1), 29–34. https://doi.org/10.3109/17549500903137256

Horley, K., Reid, A., & Burnham, D. (2010). Emotional prosody perception and production in dementia of the Alzheimer's type. *Journal of Speech Language and Hearing Research*, *53*(5), 1132–1146. https://doi.org/10.1044/1092-4388(2010/09-0030)

Kato, S., Homma, A., Sakuma, T., & Nakamura, M. (2015). Detection of mild Alzheimer's disease and Mild Cognitive Impairment from elderly speech: Binary discrimination using logistic regression. *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 5569–5572). https://doi.org/10.1109/EMBC.2015.7319654

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P. H., & David, R. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *1*(1), 112–124. https://doi.org/10.1016/j.dadm.2014.11.012

Langa, K. M., & Levine, D. A. (2014). The diagnosis and management of mild cognitive impairment. *JAMA*, *312*(23), 2551. https://doi.org/10.1001/jama.2014.13806

Lunsford, R., & Heeman, P. A. (2015, January). Using linguistic indicators of difficulty to identify mild cognitive impairment. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (pp. 658–662). https://www.isca-speech.org/archive/interspeech_2015/papers/i15_0658.pdf

Mirheidari, B., Blackburn, D., O'Malley, R., Walker, T., Venneri, A., Reuber, M., & Christensen, H. (2019). Computational cognitive assessment: Investigating the use of an intelligent virtual agent for the detection of early signs of dementia. *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/ICASSP.2019.8682423

Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., & Christensen, H. (2019). Dementia detection using automatic analysis of conversations. *Computer Speech & Language*, *53*, 65–79. https://doi.org/10.1016/j.csl.2018.07.006

Mitchell, A. J., & Shiri-Feshki, M. (2009). Rate of progression of mild cognitive impairment to dementia – Meta-analysis of 41 robust inception cohort studies. *Acta psychiatrica Scandinavica*, *119*(4), 252–265. https://doi.org/10.1111/j.1600-0447.2008.01326.x

Mueller, K. D., Hermann, B., Mecollari, J., & Turkstra, L. S. (2018). Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *Journal of Clinical and Experimental Neuropsychology*, *40*(9), 917–939. https://doi.org/10.1080/13803395.2018.1446513

Müller, S., Mychajliw, C., Reichert, C., Melcher, T., & Leyhe, T. (2016). Autobiographical memory performance in Alzheimer's disease depends on retrieval frequency. *Journal of Alzheimer's Disease*, *52*(4), 1215–1225. https://doi.org/10.3233/jad-151071

Nasreddine, Z. S., Phillips, N. A., Charbonneau, V. S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Petersen, R. C. (2011). Mild Cognitive Impairment. *New England Journal of Medicine*, *364*(23), 2227–2234. https://doi.org/10.1056/NEJMcp0910237

Pistono, A., Pariente, J., Bézy, C., Lemesle, B., Men, J. L., & Jucla, M. (2019). What happens when nothing happens? An investigation of pauses as a compensatory mechanism in early Alzheimer's disease. *Neuropsychologia*, *124*, 133–143. https://doi.org/10.1016/j.neuropsychologia.2018.12.018

R Core Team. (2020). R: A language and environment for statistical computing [Computer software]. https://www.R-project.org/

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. https://doi.org/10.1109/tasl.2011.2112351

Sachdev, P. S., Blacker, D., Blazer, D. G., Ganguli, M., Jeste, D. V., Paulsen, J. S., & Petersen, R. C. (2014). Classifying neurocognitive disorders: The DSM-5 approach. *Nature Reviews. Neurology*, *10*(11), 634–642. https://doi.org/10.1038/nrneurol.2014.181

Satt, A., Hoory, R., König, A., Aalten, P., & Robert, P. H. (2014). Speech-based automatic and robust detection of very early dementia. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (pp. 2538–2542). https://www.isca-speech.org/archive/interspeech_2014/i14_2538.html

Shinkawa, K., Kosugi, A., Nishimura, M., Nemoto, M., Nemoto, K., Takeuchi, T., Numata, Y., Watanabe, R., Tsukada, E., Ota, M., Higashi, S., Arai, T., & Yamada, Y. (2019). Multimodal behavior analysis towards detecting mild cognitive impairment: Preliminary results on gait and speech. *Studies in Health Technology and Informatics*, *264*, 343–347. https://doi.org/10.3233/SHTI190240

Smolík, F., Stepankova, H., Vyhnálek, M., Nikolai, T., Horáková, K., & Matějka, Š. (2016). Propositional density in spoken and written language of czech-speaking patients with mild cognitive impairment. *Journal of Speech, Language, and Hearing Research*, *59*(6), 1461–1470. https://doi.org/10.1044/2016_JSLHR-L-15-0301

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., & Pakaski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Frontiers in Aging Neuroscience*, *7*, 1–7. https://doi.org/10.3389/fnagi.2015.00195

Tanaka, H., Adachi, H., Ukita, N., Kudo, T., & Nakamura, S. (2016). Automatic detection of very early stage of dementia through multimodal interaction with computer avatars. *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*. https://doi.org/10.1145/2993148.2993193

Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatlóczki, G., Bánréti, Z., Pákáski, M., & Kálmán, J. (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research*, *15*(2), 130–138. https://doi.org/10.2174/1567205014666171121114930

Walker, G. (2017). Visual representations of acoustic data: A survey and suggestions. *Research on Language and Social Interaction*, *50*(4), 363–387. https://doi.org/10.1080/08351813.2017.1375802

Walker, T., Christensen, H., Mirheidari, B., Swainston, T., Rutten, C., Mayer, I., Blackburn, D., & Reuber, M. (2018). Developing an intelligent virtual agent to stratify people with cognitive complaints: A comparison of human–patient and intelligent virtual agent–patient interaction. *Dementia, 19*(4), 1173–1188. https://doi.org/10.1177/1471301218795238

Yaffe, K., Petersen, R. C., Lindquist, K., Kramer, J., & Miller, B. (2006). Subtype of mild cognitive impairment and progression to dementia and death. *Dementia and Geriatric Cognitive Disorders*, *22*(4), 312–319. https://doi.org/10.1159/000095427