Effects of Age-Related Cognitive Decline on Elderly User Interactions with Voice-Based Dialogue Systems

Masatomo Kobayashi, Akihiro Kosugi, Hironobu Takagi, Miyuki Nemoto, Kiyotaka Nemoto, Tetsuaki Arai, Yasunori Yamada

Effects of Age-Related Cognitive Decline on Elderly User Interactions with Voice-Based Dialogue Systems

Masatomo Kobayashi[1], Akihiro Kosugi[1], Hironobu Takagi[1], Miyuki Nemoto[2], Kiyotaka Nemoto[2], Tetsuaki Arai[2], and Yasunori Yamada[1]

[1] IBM Research, Tokyo 103–8510, Japan
[2] University of Tsukuba, Tsukuba 305–8577, Japan
mstm@jp.ibm.com

Abstract. Cognitive functioning that affects user behaviors is an important factor to consider when designing interactive systems for the elderly, including emerging voice-based dialogue systems such as smart speakers and voice assistants. Previous studies have investigated the interaction behaviors of dementia patients with voice-based dialogue systems, but the extent to which age-related cognitive decline in the non-demented elderly influences the user experiences of modern voice-based dialogue systems remains uninvestigated. In this work, we conducted an empirical study in which 40 healthy elderly participants performed tasks on a voice-based dialogue system. Analysis showed that cognitive scores assessed by neuropsychological tests were significantly related to vocal characteristics, such as pauses and hesitations, as well as to behavioral differences in error-handing situations, such as when the system failed to recognize the user's intent. On the basis of the results, we discuss design implications towards the tailored design of voice-based dialogue systems for ordinary older adults with age-related cognitive decline.

Keywords: Voice-Based Interactions, Smart Speakers, Voice Assistants, Aging, Age-Related Cognitive Decline.

1 Introduction

Voice-based dialogue systems show good potential to help older adults maintain their independent living. Typical examples of such systems include smart speakers and voice assistants, such as Amazon Alexa and Google Home, Apple Siri, and Microsoft Cortana [1]. Older adults can use these systems for a variety of life support services such as asking about the time or weather, accessing healthcare applications [2], and strengthening social connections [3]. Other examples are companion agents or robots for eldercare purposes, which verbally communicate with older adults to provide assistance for daily living through medication reminders and home automations [4][5][6]. The use of voice-based natural interfaces is expected to enable older adults to easily access the system—even those who were excluded from traditional desktop or mobile technologies due to their limited literacy on information technologies or age-related decline of vision and motor abilities. In addition, previous studies have shown that voice input

could be the most preferable input modality for older adults [7], and that the listening ability of older adults was comparable to that of younger adults when they are not visually impaired [8].

At the same time, special consideration of the user's cognitive functioning needs to be taken when designing voice-based dialogue systems for older adults. One study that analyzed human-human conversations revealed a significant difference between people with dementia and healthy controls in the appearance of breakdowns in communication, such as lack of uptake/continuation, where ignorance and interruptions occur [9]. Similar behaviors were found in human-robot conversations between people with Alzheimer's disease (AD) and a companion robot [10]. Studies on speech analysis have revealed associations between people's cognitive abilities and their linguistic and vocal characteristics. For example, we now know that linguistic features such as vocabulary richness may decrease in the conversations of people with AD [11][12]. A difference in vocal features such as pauses and hesitations can be a sign of progress in AD and mild cognitive impairment (MCI) [13][14]. These linguistic and vocal characteristics may reduce the quality of user experiences of voice-based dialogue systems due to resultant failure in automatic speech recognition (ASR) or conversation management engines.

As the studies above have investigated the conversational characteristics of people with MCI, AD, and other types of dementia, little is known about the effects of age-related cognitive decline in non-demented older adults, even though a large volume of older adults who may benefit from voice-based dialogue systems belongs to this cohort. A few exceptions include the studies on the MATCH corpus [15], a rich annotated dataset for the interactions of younger and older adults with a voice-based dialogue system, which also provides cognitive scores for each participant. However, no analysis has been conducted on vocal features, and no effect of cognitive measures on the completion of tasks has been reported in [15]. Further investigation is required to determine how age-related cognitive decline in ordinary older adults affects their interaction behaviors, which may lead to a failure in tasks, and whether there is any need for special considerations when designing voice-based dialogue interfaces for this cohort.

In this work, we conducted an empirical study to investigate the effect of age-related cognitive decline on the user experiences of modern voice-based dialogue systems, in which 40 non-demented older adults aged 60 or above were involved. The participants had cognitive scores assessed by standard neuropsychological tests to examine the relationship between their age-related cognitive decline and behavioral characteristics in interactions with a voice-based dialogue interface. We used a Wizard-of-Oz (WOz) interface [16] to perform three task scenarios that contain typical dialog patterns including error handling situations, which commonly appear in modern voice-based dialogue systems such as smart speakers and voice assistants. We analyzed the relationships between participants' cognitive scores and conversational characteristics from the perspectives of vocal features such as pauses and hesitations as well as rephrasing and correcting behaviors in error handling situations. Then we investigated the implications of our study towards tailored designs of voice-based dialogue systems for older adults who may have age-related cognitive decline.

The contributions of this work include: i) providing the first empirical results investigating how age-related cognitive decline in ordinary older adults influences interaction characteristics on a voice-based dialogue system; ii) identifying significant associations between cognitive scores and vocal features as well as error handling behaviors that may affect the user experience of voice-based dialogue systems; and iii) presenting points for design consideration for voice-based dialogue systems for older adults who may be experiencing age-related cognitive decline.

2 Related Work

2.1 Screen- and Voice-Based Interactive Systems for the Elderly

Aging inevitably involves multiple declines in sensory, perceptual, motor, and cognitive abilities. A combination of accessibility considerations is therefore required, which has led to extensive studies on interface designs for the aged population. For example, elderly interactions on screen-based visual interfaces such as mobile touchscreens have been investigated from the perspectives of target selections, text entry, and gesture-based interactions [17]. Kobayashi et al. [18] studied typical touchscreen operations with ordinary older adults and introduced design implications for the population. Wacharamanotham et al. [19] tested the "swabbing" technique as an assistive input method for people with tremor. For text entry on a touchscreen, Nicolau and Jorge [20] conducted a detailed investigation on the relationship between users' tremor profiles and their text entry performance on mobile and tablet devices. On top of gesture analysis studies, Sato et al. [21] proposed an intelligent help system that automatically provides novice older users with context-aware instructions on gesture interactions. For traditional desktop interfaces, ability-based adaptation techniques have been proposed [22]. Gajos et al. [23] used their SUPPLE system to automatically generate customized interfaces based on users' ability profiles. Trewin et al. [24] introduced the Steady Clicks technique to assist with clicking actions for people with motor impairments, while Wobbrock et al. [25] proposed the Angle Mouse technique to assist them with mouse cursor movement. Sato et al. [26] reported that additional voice-based feedback could improve older users' subjective performance on a visual user interface. These studies on aging and screen-based visual interfaces motivated us to investigate aging and voice-based dialogue interfaces and to discuss prospective design adaptation for older adults.

The voice-based dialogue system is a promising style of interaction for older adults, given their performance on voice-based interactions. Smith and Chaparro [7] showed that voice input is the most effective and preferable input modality for older adults. Bragg et al. [8] reported that the listening speed of sighted older adults is comparable to that of sighted younger adults. Note that studies have also indicated challenges related to ASR for older adults with cognitive disorders. Weiner et al. [27] showed that the accuracy of ASR decreased not only for people with AD but also for those with age-related cognitive decline. Rudzicz [28] indicated that older adults with higher cognitive scores experienced fewer ASR errors, although the trend was not statistically signifi-

cant. Zajicek [29] pointed out that "errors and error recovery represent the primary usability problem for speech systems". Even though there might be a challenge in terms of ASR accuracy, many studies have proposed and investigated voice-based dialogue agents and robots [30], some of them for ordinary older adults and others for those with dementia. Granata et al. [31] tested both voice and graphical input modalities for an eldercare robot for people with cognitive disorders and found there is a need for adaptation of vocabulary and the design of image icons. Wolters et al. [5] conducted focus group studies with people with dementia, caregivers, and older adults without a diagnosis of dementia, suggesting that voice-based dialogue systems should be able to adapt to diverse paths of cognitive aging. As for non-demented older adults, an example is Portet et al.'s work [4], in which a WOz study was conducted to investigate their acceptance of voice command interactions in a smart house environment. Ziman and Walsh [32] studied elderly perception of voice-based and traditional keyboard-based interfaces and reported that the voice-based interface was easier to learn and use, even though the keyboard-based interface was more preferred. Our study aims to provide design implications for these kinds of voice-based dialogue systems for older adults who may have age-related cognitive decline.

2.2 Corpus Analysis on Elderly Interactions with Conversational Systems

There have been some studies that built a corpus for research on the conversational interactions of older adults. The MATCH corpus [15] is a multi-modal dataset that involved both older and younger adults who interacted with nine different spoken dialogue systems. The task scenario used in the data collection phase was "appointment scheduling" as a relevant task for older adults. A unique aspect of this corpus is that it contains information about the users' cognitive abilities and detailed usability assessments of each dialogue system, in addition to utterances and transcripts with annotations. The findings from initial analyses suggested that there was no effect of any of the cognitive measures on task completion. This corpus allows analyses of the conversational characteristics of older adults. For example, it was found that older users more frequently used "social" words and phrases such as "thank you". Vipperla et al. [33] used the MATCH corpus to build language and acoustic models to improve ASR accuracy for older adults' speech. Bost and Moore [34] performed studies using the MATCH corpus as well; they used regression models and showed that users with higher cognitive scores had shorter dialogues while users with shorter dialogues were more satisfied with the dialogue system. Jasmin-CGN [35] is another corpus of multi-generational human-machine conversational interactions. Even though it does not contain information about the users' cognitive abilities, its conversation scenario covers simulated ASR errors. LAST MINUTE [36] is also a multi-modal corpus of younger and older users' interactions with a voice-based dialogue system and contains transcripts, videos, and responses to psychometric questionnaires. A study on the LAST MINUTE corpus [37] reported that discourse particles (a type of hesitation) increased in critical situations in human-computer conversations where, for example, the system's behavior was not understandable for the user. The CADENCE corpus [38], which involved older adults with a diagnosis of dementia or MCI, contains transcribed spoken interactions

between a voice-based dialogue system and older users accompanied with detailed information about users' cognitive abilities; its aim is to support research on inclusive voice interfaces. The conversation data in the corpora above were collected using the WOz method, suggesting that this method is an appropriate way to collect conversation data for controlled empirical analyses.

Studies have also investigated the conversational characteristics of people with dementia. Watson [9] used data from human-human conversations between ten people with AD and ten without to analyze types of breakdowns in conversations (i.e., trouble indicating patterns) and to identify the relevant repair strategies. Rudzicz et al. [10] used a similar approach to analyze human-robot conversations between ten older adults with AD and a voice-based dialogue system and found that older adults with AD were very likely to simply ignore the robot. Rudzicz et al. [39] and Chinaei et al. [40] built a computational model that aimed to exploit linguistic and acoustic features to detect a breakdown in conversations. In contrast to these previous studies that investigated the conversations of older adults with AD, our focus in the present study is interactions between non-demented older adults who may have age-related cognitive decline and a voice-based dialogue system in simulated typical application scenarios with a modern smart speaker or voice assistant.

2.3 Language Dysfunctions Due to Cognitive Impairments

How cognitive functioning changes speech characteristics has mainly been investigated in patients, especially dementia patients. While the most typical symptom of dementia is memory impairment due to shrinkage of the medial temporal lobe [41][42], both retrospective analysis and prospective cohort studies have shown that language dysfunctions prevail even from the presymptomatic period [43][44]. Such clinical symptoms that can precede dementia (including AD) are considered to be a mild cognitive impairment (MCI) [45]. The concept of MCI has been used to identify an intermediate stage of cognitive impairment that is often, but not always, a transitional phase from cognitive changes in normal aging to those typically found in dementia [45]. People with MCI typically exhibit less severe symptomology of cognitive impairment than that seen in dementia. Many computation studies have aimed to automatically capture such gradual changes in cognitive functioning by investigating the difference of speech features, namely, acoustic, prosodic, and linguistic features, among healthy older adults and MCI and AD patients [14][46][47][48]. They mainly investigated speech data during neuropsychological tests and medical interviews. For example, the impairment of short-term memory often makes normal conversation difficult due to language dysfunctions such as difficulties with word-finding and word-retrieving [49][50]. These language dysfunctions have been measured as pauses and fillers (non-words and short phrases like "umm" or "uh") [45][51]. In fact, some studies have shown that patients with AD and MCI use more pauses in spontaneous speech, and on average use longer pauses than healthy controls [14][52]. Such speech changes seem likely to influence the user experience with an interactive system. For this reason, we should investigate whether and how such features occur and influence the user experience even in non-demented older adults with different levels of cognitive functioning.

In addition, many studies have demonstrated the ways in which text features significantly change over the course of cognitive impairment [12][53]. Among them, numerous studies investigated the difference of information content in description tasks [54][55] and found that individuals with MCI and AD tend to produce descriptions with lower information content than healthy controls in both verbal and written picture descriptions tasks [56][57]. Our interest is whether such decline in information content can be observed in older adults with age-related cognitive decline during the use of a voice-based interaction system. If so, such decline might influence whether the system can understand the user requirements because the impact of misrecognized words might be more significant. Therefore, we decided to investigate the speech features described in the above that would change according to the level of cognitive functioning and influence the interaction with a voice-based system.

3 Research Hypotheses

As stated in the introduction, we particularly focused on vocal features such as pauses and hesitations as well as error handling behaviors such as rephrasing and correcting during the analysis. We hypothesized three behavioral characteristics of older adults with age-related cognitive decline, as follows.

H1. *Pauses, hesitations, and other disfluency features increase with cognitive decline*—this is a hypothesis regarding vocal features. We assumed that the disfluency found in vocal feature analysis on people with MCI and dementia could also appear in non-demented older adults with age-related cognitive decline. We chose pauses, hesitations, and delays as commonly used features in the previous vocal feature analysis studies such as [14]. An additional feature, interruptions, was also included, as inspired by [58] and based on our preliminary observations on elderly conversations.

The appearance of these features would cause speech recognition errors due to inappropriate segmentation of speech segments where, for instance, a long pause could be misinterpreted as a sentence delimiter. We also assumed that the occurrence of these vocal features would increase in cognitively demanding contexts such as in error handling situations and when responding to open-ended questions.

H2. *The failure in rephrasing increases with cognitive decline*—this is a hypothesis for error handling features. A voice-based dialogue system often fails to interpret the user's intention in a response, which could happen either because of speech recognition errors or inappropriate wording by the user. In these cases, the user is required to rephrase or simply repeat the response. We assumed that cognitive decline would affect interaction behaviors in this situation because slightly more complex cognitive functions are required, such as lexical access to perform appropriate rephrasing.

H3. *The failure in correcting increases with cognitive decline*—this is also a hypothesis for error handling features. A voice-based dialogue system often incorrectly recognizes the user's input (e.g., "thirteen" vs. "thirty"), which is mainly caused by speech recognition errors. In these cases, the user needs to notice the misrecognition when confirmed by the system and ask for a correction. We assumed that cognitive decline would affect interaction behaviors in this situation because more complex cognitive

functions are required, such as paying attention to notice the misrecognition and then to perform the appropriate correction.

The second and third hypotheses were inspired by [29], which emphasized the importance of error handling in speech systems, and [35], which presented a corpus including simulated errors.

4 Method

4.1 Task Scenarios

Three task scenarios were prepared to simulate typical application scenarios on modern smart speakers and voice assistants, as well as to simulate ASR error conditions. The scenarios consisted of information retrieval (asking for tomorrow's weather), shopping online (booking a movie ticket), and personal schedule management (creating a calendar event) as representative scenarios on a voice-based dialogue system that would help older adults to live active, independent lives. In every scenario, participants started the task by speaking a wake word. The tasks were ordered to start with a simple scenario and then advance to more complicated ones:

1. *Ask for tomorrow's weather*: This is a single round-trip dialogue. The participant simply makes a request once to complete the task (Fig. 1).
2. *Book a movie ticket*: This scenario is a multiple turn dialogue. Once the participant asks the system to purchase tickets for a movie, the system asks what kind of movie the participant wants, the date, the show time, the number of tickets, and payment information.
3. *Create a calendar event*: This scenario is a multiple turn dialogue. The participant adds an event (watching a movie) that has been booked in the previous task. They are asked the date, time, title of the event, and when to set a notification alarm. This scenario purposely includes error handing situations. Specifically, the system verbalizes an error message indicating that it could not catch what the participant said, or it gives the wrong confirmation.

The questions presented by the system during the tasks were categorized as follows (Fig. 2):

- *Open-ended*: Participants respond with a free sentence to answer the question.
- *Multiple options*: Participants choose one from the options stated in the question.
- *Prepared input*: Participants respond with the information (e.g., passcode) specified by the experimenter.
- *Confirmation*: Participants need to accept or reject what the system has stated.

P: Kasuga-san?
S: Hello, how may I help you?
P: What will the weather be like tomorrow in this city?
S: It will be rainy tomorrow in this city.

Fig. 1. An example exchange between a participant (P) and the system (S) in "Ask tomorrow's weather" task. "Kasuga-san?" was the wake word starting the session.

(Open-ended)
S: What movie would you like to watch?
P: I'd like to watch a comedy.

(Multiple options)
S: When would you like to set the alarm? You can set it for 5 minutes, 10 minutes…, or 2 hours before the event.
P: Please set the alarm for 10 minutes before the event.

(Handling with recognition error)
S: How may I help you?
P: Please add a calendar event of going to a movie.
S: *Sorry, I couldn't catch you. Could you repeat the request again?*
P: Well, could you add an event please?

(Handling error confirmation)
S: What time will that event be?
P: It will be at 10:30.
S: *OK. The event is scheduled for 10:00.*
P: No, please set it for 10:30

Fig. 2. Examples of questions by types. System utterance in *italics* is an intentional error.

4.2 Apparatus

We took a Wizard-of-Oz (WOz) approach so that we could conduct quantitative analyses with a limited number of trials by controlling the content of the conversations. In particular, we wanted to control the appearance of error handling situations, where the participants had to perform appropriate responses such as rephrasing and correcting, assuming that such a cognitively demanding situation would lead to more differences in behavioral characteristics. With the WOz method, we could avoid unfavorable results due to environmental disturbances such as room noise.

The system consisted of a tablet (iPad Air2) as the front-end terminal for the participants and a laptop as the controller for the experimenter. Participants sat down in front of the tablet and talked with the system through the tablet to perform the tasks (Fig. 3). The tablet showed a screen indicating whether it was speaking or listening. To record the participants' voices, we used two microphones, a throat microphone (NANZU SH-

12iK) and a lavalier microphone (SONY ECM-CS3), in addition to the embedded microphone in the tablet. All the microphones were set up to record in raw format with the sampling rate of 44.1 kHz.

The iPad's default Siri Female voice was chosen for the vocal type of speech. The speed rate was set to 85% of the normal speed of the voice. The vocal type and speed rate were determined through informal preliminary trials, in which three cognitive-healthy older adults aged 60s–80s (1 male, 2 females) tried earlier versions of the experimental interface and were asked their preferences on the voice type and speed. Even though literature suggests that low-pitch male voices are more preferable for the elderly than female voices [59], we chose the female voice for three reasons: i) a female voice is commonly set as the default in many voice-based dialogue systems; ii) for the language we used, the quality of voice synthesis is much better for the default voice; and iii) the participants in the preliminary trials preferred the female voice.

The experimenter simulated the conversation management engine through a browser-based controller interface that included buttons listing what the tablet would speak along with the scenarios. During the experimental session, the experimenter listened to the participants and determined what and when the system should speak next by clicking one of the buttons. The sentences were scripted in advance and the experimenter tried to mimic the behavior of typical conversation management engines as closely as possible.

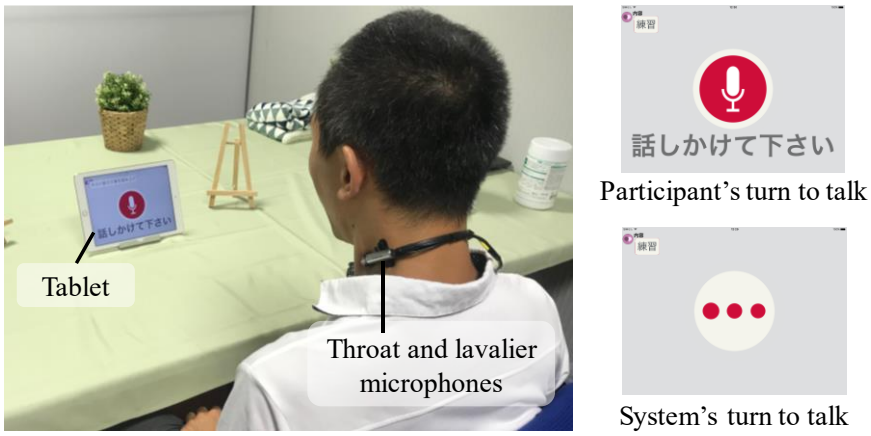

Fig. 3. Overview of experimental setup.

4.3 Participants

Forty older adults (20 female, age: 61–79, average = 69.85, SD = 4.7) in good health were invited. Participants were hired through a local recruiting company, and none of them were diagnosed as having dementia. The criteria for recruiting were "older than 60 without any serious diseases or disabilities including neurodegenerative diseases such as dementia". All participants had a perfect score on the Barthel Index of Activities of Daily Living [60], indicating they did not need any assistance in their everyday lives.

Even though 18 of the 40 participants stated that they had experience with voice-based interfaces such as voice-based text entry and single-turn dialogue, none of them had used multi-turn dialogue systems. This study was conducted under the approval of the local ethics committee.

4.4 Cognitive Measures

We collected cognitive measures of neuropsychological assessments that are typically used for clinically evaluating the cognitive functioning of older adults. Four different scales were used to quantitatively capture different aspects of cognitive functioning. These assessments were administered by clinical psychologists.

MMSE. Min-Mental State Examination (MMSE) [61] is used as a screening measure of global cognitive functioning. This test provides a composite score based on the assessment of multiple cognitive domains: orientation for place and time, memory and attention, language skills, and visuospatial abilities. The highest possible score is 30 points, and lower scores indicate greater degrees of general cognitive dysfunction. A score of 26 or above is typically considered normal.

FAB. Frontal Assessment Battery (FAB) [62] is designed to assess executive functions that are thought to be under the control of the frontal lobes. This test includes a brief battery of six neuropsychological tasks: conceptualization, mental flexibility, motor programming, sensitivity to interference, environmental autonomy, and inhibitory control. FAB is scored from 0 to 18. Lower scores indicate greater degrees of executive dysfunction.

LM. Logical Memory Test (LM) from the Revised Wechsler Memory Scale [63][64] is used to assess cognitive functioning associated with memory and learning. This test involves listening to two short paragraph-length passages with immediate recall (LM1) and 30-minute delayed recall (LM2). A delayed recall trial is administered without warning. Each passage consists of 25 elements, and the score is taken as the mean of the two stories based on the number of correct responses. The highest possible score is 25 points.

TMT. Trail Making Test (TMT) [65][66] is a visuomotor timed task used routinely in clinical evaluations to assess the cognitive domains of cognitive flexibility and executive function, especially as related to attention. The test consists of two parts: TMT-A and TMT-B. TMT-A requires one to draw lines connecting consecutive numbers randomly distributed in space (i.e., 1-2-3…). TMT-B is similar, but instead of just linking numbers, participants are required to draw lines connecting numbers and letters alternately in their respective sequence (i.e., 1-A-2-B-3-C…).

4.5 Procedure

Participants went through an orientation session to help them understand what a voice-based interface is and how conversation through it should proceed. The orientation included: i) explanation of the purpose, where participants were asked to test the voice-based dialogue system and then to share their impressions and suggestions; ii) run-through practice with a simple scenario, where participants went through a practice

session starting with a “wake word”; and iii) confirmation if the volume of the tablet voice was high enough.

Then the participants proceeded to the main tasks, which they worked through in the same order described in Section 4.1. In each task, participants were provided with a printout containing the information required to proceed with the questions, such as the date of the reservation and the number of tickets, for which they would be asked during the session. After the tasks, we conducted a short interview with participants to uncover any difficulties they experienced throughout the experiment and points they felt were useful. Each experimental session took approximately 30 minutes per person.

4.6 Vocal Features

The following categories of vocal features were extracted from the recorded voices of participants by a semi-automatic process without any manual annotations. Each feature was averaged among all questions and four groups categorized by question types. Table 1 shows the full list of features.

Pause. The average length of silent sections in participants’ speech. The silent section was defined as a section with a volume level below a certain threshold and lasting for a certain period, where the thresholds were manually determined by the experimenter before analyzing the collected data, so that the resulting “pause” segments were as consistent with human-perceived pause segments as possible. Specifically, the threshold of the volume level was set to 48 dB, and intervals that last longer than 500 ms were counted. We measured the total length of the pause sections in the responses for each question.

Hesitation. This feature counted how often the “hesitation” attribute applied in the result from ASR (Watson Speech-To-Text). The ASR feeds “hesitation” in the result for the period when participants uttered fillers or spoke unclearly or in a smaller voice. We counted the occurrence of hesitations in a response and divided the count by the length of the recognized text of the response for each question.

Delay. The average length of the silent section before the participants’ speech after the system’s question. The same conditions used for “pause” were used for the “delay” section, but it was only labeled as “delay” when the silent section was detected at the beginning of the speech. We measured the length of the section for each question.

Interruption. The average number of participants’ interruptions while the system was speaking. Segments when the system was speaking were clipped from the sound recorded with the throat microphone. The throat microphone does not record the sound of open-air, so the sections of above a certain volume level in the clipped segment means the period of participants’ interruption. We counted the number of occurrences of the sections for each question.

4.7 Error Handling Features

Participant responses in the simulated error situations were evaluated on the following aspects.

Failure in rephrasing. A label of how the participant responded in the "handling with recognition error" case of the scenario example (Fig. 2). In the normal case, participants asked to create a calendar event in both trials (2nd and 4th line). However, when the request was rejected in the first trial with error, they sometimes rephrased differently, e.g., "I'm going to a movie", which is a request with insufficient information to be determined correctly. The interaction was labelled "failure" unless the contents of both the first and second trials of making a request included all the necessary keywords for the request.

Failure in correcting. A label indicating if the participant accepted a confirmation of the wrong value in the "handling with error confirmation" case in the scenario example (Fig. 2). For example, if the participant replied "well done, thanks" in the 4th line of example dialogue, the value was labelled "failure".

As each error case was executed once in the "create a calendar event" task for each participant, both of the error handing features take a binary value.

4.8 ASR Error Feature

Given that previous studies have repeatedly reported the challenges related to ASR errors in voice-based dialogue systems, we examined it with an up-to-date ASR engine (Watson Speech-To-Text as of September 2018) with neither custom language nor acoustic models.

ASR error rate. This feature counted how often the ASR mis-transcribes the participant's utterance. We compared a sentence automatically transcribed by ASR with one manually transcribed by the experimenter (ground truth) to check how often the utterance of a participant was misrecognized. We counted the occurrence of different categorematic words between the automatic and manual transcriptions. This information was gathered for five questions whose responses were mostly the same among the participants, such as making the initial request or asking for confirmation. Then we divided the total occurrence of different words by the total count of categorematic words appearing in the ground truth text as the average ASR error rate over five questions.

Table 1. List of features used in the analysis.

Name	*Category*	*Source*	*Value*
Pause-All	Pause	All questions	Average length
Pause-O		Open-ended questions	
Pause-M		Multiple options questions	
Pause-P		Prepared input questions	
Pause-C		Confirmation questions	
Hesitation-All	Hesitation	All questions	Average rate
Hesitation-O		Open-ended questions	
Hesitation-M		Multiple options questions	
Hesitation-P		Prepared input questions	
Hesitation-C		Confirmation questions	
Delay-All	Delay	All questions	Average length
Delay-O		Open-ended questions	
Delay-M		Multiple options questions	
Delay-P		Prepared input questions	
Delay-C		Confirmation questions	
Interruption-All	Interruption	All questions	Average rate
Interruption-O		Open-ended questions	
Interruption-M		Multiple options questions	
Interruption-P		Prepared input questions	
Interruption-C		Confirmation questions	
F-Rephrasing	Failure in rephrasing	"Handling with recognition error" question	Binary
F-Correcting	Failure in correcting	"Handling error confirmation" question	Binary
E-ASR	ASR error rate	5 questions normally answered similarly	Average rate

4.9 Statistical Analysis

We investigated the association between cognitive scores and behavioral features by using a linear regression model with age and gender as covariates. We examined deviations of variables from normality with their skewness statistics. A log-transform was applied to variables whose skewness statistics were more than twice the standard error to normalize their distribution. In this study, we set the significance level to 0.05.

5 Results

All 40 participants went through all the task scenarios. We collected 1,040 utterances in total. One participant got upset after the "handling error confirmation" turn in the "create a calendar event" task. We excluded the values that followed this from the results. Another participant could not understand the request to state the title of the event

in “create a calendar event”. We excluded the value of that question and a subsequent one asking for more details of the title. In total, four open-ended questions and one confirmation question were excluded. Figure 4 shows an overview of the extracted vocal features. For error handling features, 12 and eight out of the 40 participants were labeled as “failure” in rephrasing and correcting, respectively. The median of the ASR error rate was 9.7% (interquartile range: 0.7%–17.1%).

The participants did not have any difficulty in hearing synthesized voices, seeing the tablet screen, and talking with the experimenters. The MMSE scores ranged between 25 and 30 (mean = 28, SD = 1.5).

We investigated the relationship between cognitive scores and behavioral features by using a linear regression model controlling for age and gender information, as shown in Table 2. Of the 23 behavioral features, we found statistically significant associations between MMSE scores and the following five features: Pause-O, Hesitation-All, Hesitation-O, F-Rephrasing, and F-Correcting. We also found significant associations between FAB, LM1, or LM2 and the following features: F-Rephrasing for FAB, Interruption-C for LM1, and Interruption-C and E-ASR for LM2. As for the TMT-A and TMT-B assessments, we used the number of errors and the time needed to complete the task. Results showed significant associations between the number of errors with Pause-P and F-Rephrasing for TMT-A and F-Correcting for TMT-B. We also found that time for the tasks was significantly related to Delay-O and F-Correcting for TMT-B, while no significant associations were found for TMT-A.

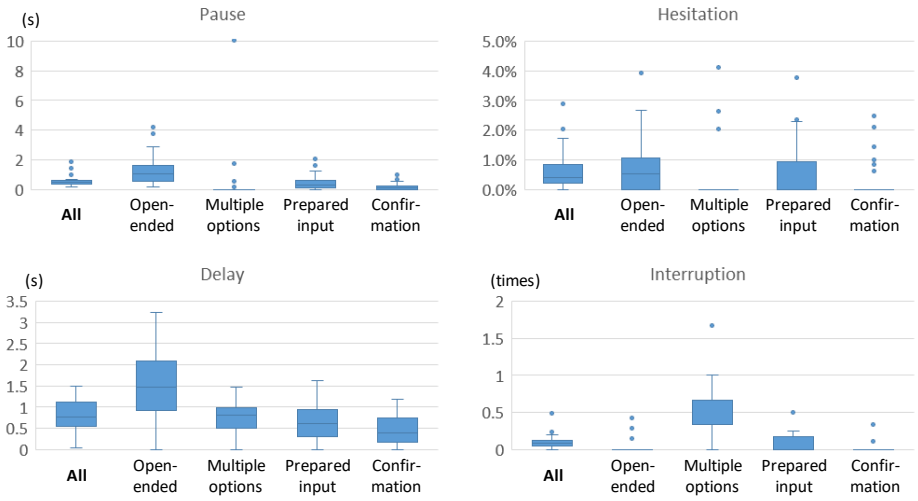

Fig. 4. Overall distribution of vocal features. The lines and boxes represent medians and interquartile ranges (IQR), respectively. The whiskers extend to most extreme data points up to 1.5 times the IQR. The dots represent outliers.

Table 2. Linear regression analysis on cognitive scores.

MMSE

Feature	*β*	*95% CI*		*p Value*
Pause-O	-0.5078	-1.0006	-0.0150	0.0438
Hesitation-All	-0.7045	-1.1807	-0.2284	0.0049
Hesitation-O	-0.6155	-1.0771	-0.1540	0.0104
F-Rephrasing	0.7044	0.2809	1.1280	0.0018
F-Correcting	1.1955	0.0154	2.3756	0.0472

FAB

Feature	*β*	*95% CI*		*p Value*
F-Rephrasing	0.7635	0.0783	1.4486	0.0300

LM1

Feature	*β*	*95% CI*		*p Value*
Interruption-C	1.5022	0.4406	2.5639	0.0068

LM2

Feature	*β*	*95% CI*		*p Value*
Interruption-C	1.5008	0.4417	2.5600	0.0068
E-ASR	-1.2822	-2.3612	-0.2033	0.0212

TMT-A (errors)

Feature	*β*	*95% CI*		*p Value*
Pause-P	-0.2375	-0.3845	-0.0905	0.0023
F-Rephrasing	-0.1710	-0.3280	-0.0140	0.0336

TMT-B (errors)

Feature	*β*	*95% CI*		*p Value*
F-Correcting	-1.7976	-3.5930	-0.0021	0.0497

TMT-A (time)

Feature	*β*	*95% CI*		*p Value*
n.s.				

TMT-B (time)

Feature	*β*	*95% CI*		*p Value*
Start-O	13.4064	4.5527	22.2601	0.0041
F-Correcting	-28.1818	-51.0682	-5.2953	0.0172

β: standardized coefficients, F-Rephrasing and F-Correcting took nominal value

6 Discussion

In this section, we discuss the implications of the experimental results, aiming to extend the general design guidelines for voice-based dialogue systems (e.g., [67][68][69]) by clarifying special consideration points for older users. The analysis of behavioral features should provide useful insights for the design of senior-friendly voice-based interactions. We first examine each hypothesis on the basis of the regression analysis results and then summarize some takeaways.

H1—*Pauses, hesitations, and other disfluency features increase with cognitive decline*—was partially confirmed. The regression analysis identified significant negative associations of MMSE scores with Pause-O and Hesitation-O. This result indicates that participants who had higher cognitive scores tended to exhibit fewer pauses or hesitations, particularly when responding to an open-ended question. As MMSE is known to assess multiple cognitive domains, this seems to indicate a general relationship between cognitive functioning and the appearance of pauses and hesitations. This trend has been repeatedly reported in studies on people with MCI or dementia (e.g., [14]). Our result suggests that the same trend appears for age-related cognitive decline in non-demented older adults when they interact with a voice-based dialogue system.

Most of the pauses and hesitations appeared in response to “open-ended” questions. This is not a surprising result because this type of question requires the participants to articulate their thoughts, confirming the findings in previous studies on spontaneous speech (e.g., [14]). In the median case, the values were roughly 1 second for Pause-O and 0.5% for Hesitation-O, which would not have any serious negative effect on user experience. However, in the worst case, Pause-O was longer than 4 seconds and Hesitation-O was roughly 4%. This would lead to turn-taking errors because, in a typical

voice-based dialogue system, a long pause indicates the end of a response. The adaptation of the sentence segmentation criteria based on the user's cognitive scores and the type of the question could be effective to combat this. Also, the use of the acoustic model of ASR adapted to this trend would alleviate the problem, as suggested in previous work [33].

Interestingly, there was a significant positive correlation between both LM scores and Interruption-C. This result is opposite to our assumption. In short, the participants who had higher cognitive scores made more interruptions. A possible explanation is that healthy older adults with less age-related cognitive decline prefer faster conversations, and the 85% speech rate used in the experimental system was too slow for them. For interruptions, almost all of this type of behavior happened during "multiple options" questions. This seems to be a flaw in the dialogue design. In line with the design guidelines for developing applications on a commercial smart speaker [67], the typical "multiple options" question in our experimental system consisted of two sentences: i) clearly presenting available options and ii) clearly asking the user to make a choice. However, many participants started responding before the system finished the second sentence. It seems that, once they received the list of available options, participants wanted to answer as soon as possible. The system could avoid this type of conversation breakdown simply by accepting pre-emptive responses.

The ASR error rate was significantly higher for those who had lower cognitive (LM-2) scores, which confirmed the finding in [27][28]. The accuracy of ASR is another critical aspect that could strongly affect user experience with a voice-based dialogue system. In the median case, the ASR error rate was lower than 10%. This is a much better value than the ASR accuracy reported in previous studies (e.g., [33]), and it seems to stem from the recent advances in ASR technology. This result indicates that the ASR accuracy itself could be less critical than ever, at least for those with less age-related cognitive decline, unless it comes in conjunction with other issues such as turn-taking errors and inappropriate wording on the part of the user. On the other hand, the ASR accuracy would still be relevant for those with cognitive decline or in confusing situations. Specifically, as stated below, participants tended to exhibit poor performance in an error handling situation. ASR errors in this situation would inhibit recovery from the error state, and have a serious negative impact on the user experience.

H2—*The failure in rephrasing increases with cognitive decline*—was confirmed. The regression analysis found that participants whose request lacked the necessary keywords at least in either the first or second trial during an error handling situation had significantly lower scores for MMSE and FAB. This result suggests that rephrasing requires executive functions. Interestingly, among the 12 participants who were not labeled as "success", seven failed only in the second trial. Even though they provided sufficient keywords at first, they incorrectly paraphrased their request after they received the error message. This change implies that reduction of information content [54] in the user's utterance could be observed in non-demented older adults particularly in an error handling situation. We prepared the error message (Fig. 2) as specified in the design guidelines [67], but it might be confusing to older adults with age-related cognitive decline. The error message should be designed more carefully. A personalized or context-based error message would probably be more effective.

H3—*The failure in correcting increases with cognitive decline*—was also confirmed. The regression analysis found that participants who failed to correct the system's recognition error had significantly lower MMSE scores as well as a larger amount of time and errors for TMT-B. This result suggests that older adults with lower attention ability tend to ignore the system's misrecognition, which makes "confirmation" questions ineffective. To address this challenge, the system could exploit a screen to allow users to visually review the recognition result. The use of social networks might be another solution, where the system would ask the user's close relatives to double-check a response, as long as the conversation does not contain privacy-sensitive contents. There is a trade-off between the potential solutions for H1 and H3. Specifically, open-ended questions in a dialogue design could be replaced with a series of closed questions to reduce the occurrence of pauses and hesitations, but that would increase the number of questions for confirmation and lead to more "failure in correcting" errors.

In summary, our analysis suggests that the language and behavioral dysfunctions reported in previous studies on neuropathological cognitive impairments could also occur in a broader range of older users—particularly in cognitively demanding situations such as when handling errors. Special considerations are needed to provide a better voice-based interaction experience for older users with age-related cognitive decline, which include:

Avoid misrecognitions of the end of a response. Older users with age-related cognitive decline tended to exhibit more pauses and hesitations. The system should allow users to keep talking intermittently. Note that the pauses and hesitations are more likely to appear in a response to an open-ended question, even though existing guidelines recommend the use of open-ended questions. A dynamic adaptation of thresholds [70] might be useful for this purpose.

Accept pre-emptive responses. Older users with a higher cognitive score tended to respond to the system in a pre-emptive manner, at least in the present study. Developers of voice-based dialogue systems should keep in mind that users could face different issues even if they have better cognitive functioning.

Provide personalized, context-based error messages. A general "could not catch it" message seemed to be ineffective for older users with age-related cognitive decline. The system should take into account the details of the situation and the user's cognitive profile and provide an instructive message that clearly tells the user how to recover from the error state. For example, it would be helpful if the system could indicate whether "rephrasing" or "repeating" is needed.

Assist with confirmation. Older users with age-related cognitive decline, particularly those with lower attention ability, tended to incorrectly accept the response from the system. The system should help the user recognize its own mistake, for example, by providing screen-based visual confirmations in conjunction with voice-based ones.

These points have not been structurally emphasized in the design of senior-friendly dialogue interfaces, even though they have been shown to be relevant to older users' experience with voice-based dialogue interactions. Given the increase of aged people and the growing use of voice-based dialogue systems, these considerations will only become more critical.

7 Conclusion and Limitations

In this work, we presented the first empirical results of an investigation into how age-related cognitive decline in non-demented, ordinary older adults influenced behavioral characteristics (i.e., vocal and error handling features) in the use of a voice-based dialogue system. Our analysis on the collected human-machine conversations identified significant associations between the behavioral features and cognitive scores measured by standard assessment tools such as MMSE and LM. The results showed that differences in vocal features such as pauses and hesitations, which have been found in studies on language dysfunction related to MCI and AD, also appeared in typical voice-based dialogue interactions of a broader range of older adults with age-related cognitive decline. We then discussed the potential impact of the identified behavioral characteristics on the user experience of voice-based dialogue systems and presented points for design consideration as workarounds on prospective issues.

The main limitation of this work is the limited size of samples. Even though the number of participants is comparable to those in previous corpus studies such as [15], the collected data cover only a small portion of the diverse nature of cognitive aging. In particular, our investigation on error handling behaviors depends on a small number of question-response pairs observed during two pieces of controlled error handling situations. While this was intentional—i.e., these experimental conversation scenarios were specifically designed to examine realistic error handling behaviors—further samples are needed to quantitatively assess the result. Another potential limitation is the lack of comparison with younger people or people with MCI or dementia, even though this was also intentional as our investigation focused on the variance of age-related cognitive decline among non-demented older adults. To confirm or extend our findings, further controlled studies and large-scale wild studies will be needed. Also, the design consideration points presented in the previous section need to be implemented and evaluated with the target population.

The uniqueness of our study is that it highlighted variations in cognitive scores among healthy older adults, and then showed significant associations between the cognitive scores and interaction characteristics with a voice-based dialogue system, particularly in cognitively demanding error handling situations. We believe that our findings can provide voice-based dialogue interface designers with empirical evidence, which could not be directly supported by previous studies in different contexts.

8 Acknowledgements

We thank all of the participants in the experiment.

References

1. López, G., Quesada, L. and Guerrero, L.A., 2017, July. Alexa vs. Siri vs. Cortana vs. Google Assistant: a comparison of speech-based natural user interfaces. In International Conference on Applied Human Factors and Ergonomics (pp. 241-250). Springer, Cham.

2. Ma, M., Skubic, M., Ai, K. and Hubbard, J., 2017, July. Angel-Echo: a personalized Health care application. In Proceedings of the Second IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (pp. 258-259). IEEE Press.
3. Reis, A., Paulino, D., Paredes, H. and Barroso, J., 2017, July. Using intelligent personal assistants to strengthen the elderlies' social bonds. In International Conference on Universal Access in Human-Computer Interaction (pp. 593-602). Springer, Cham.
4. Portet, F., Vacher, M., Golanski, C., Roux, C. and Meillon, B., 2013. Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. Personal and Ubiquitous Computing, 17(1), pp.127-144.
5. Wolters, M.K., Kelly, F. and Kilgour, J., 2016. Designing a spoken dialogue interface to an intelligent cognitive assistant for people with dementia. Health informatics journal, 22(4), pp.854-866.
6. Russo, A., D'Onofrio, G., Gangemi, A., Giuliani, F., Mongiovi, M., Ricciardi, F., Greco, F., Cavallo, F., Dario, P., Sancarlo, D. and Presutti, V., 2018. Dialogue Systems and Conversational Agents for Patients with Dementia: the human-robot interaction. Rejuvenation research, (ja).
7. Smith, A.L. and Chaparro, B.S., 2015. Smartphone text input method performance, usability, and preference with younger and older adults. Human factors, 57(6), pp.1015-1028.
8. Bragg, D., Bennett, C., Reinecke, K. and Ladner, R., 2018, April. A Large Inclusive Study of Human Listening Rates. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (p. 444). ACM.
9. Watson, C.M., 1999. An analysis of trouble and repair in the natural conversations of people with dementia of the Alzheimer's type. Aphasiology, 13(3), pp.195-218.
10. Rudzicz, F., Wang, R., Begum, M. and Mihailidis, A., 2015. Speech interaction with personal assistive robots supporting aging at home for individuals with Alzheimer's disease. ACM Transactions on Accessible Computing (TACCESS), 7(2), p.6.
11. Bucks, R.S., Singh, S., Cuerden, J.M. and Wilcock, G.K., 2000. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. Aphasiology, 14(1), pp.71-91.
12. Khodabakhsh, A., Yesil, F., Guner, E. and Demiroglu, C., 2015. Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech. EURASIP Journal on Audio, Speech, and Music Processing, 2015(1), p.9.
13. Hoffmann, I., Nemeth, D., Dye, C.D., Pákáski, M., Irinyi, T. and Kálmán, J., 2010. Temporal parameters of spontaneous speech in Alzheimer's disease. International journal of speech-language pathology, 12(1), pp.29-34.
14. König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P.H. and 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), pp.112-124.
15. Georgila, K., Wolters, M., Moore, J.D. and Logie, R.H., 2010. The MATCH corpus: a corpus of older and younger users' interactions with spoken dialogue systems. Language Resources and Evaluation, 44(3), pp.221-261.
16. Salber, D. and Coutaz, J., 1993, April. A wizard of oz platform for the study of multimodal systems. In INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems (pp. 95-96). ACM.
17. Motti, L.G., Vigouroux, N. and Gorce, P., 2013, November. Interaction techniques for older adults using touchscreen devices: a literature review. In Proceedings of the 25th Conference on l'Interaction Homme-Machine (p. 125). ACM.

18. Kobayashi, M., Hiyama, A., Miura, T., Asakawa, C., Hirose, M. and Ifukube, T., 2011, September. Elderly user evaluation of mobile touchscreen interactions. In IFIP Conference on Human-Computer Interaction (pp. 83-99). Springer, Berlin, Heidelberg.
19. Wacharamanotham, C., Hurtmanns, J., Mertens, A., Kronenbuerger, M., Schlick, C. and Borchers, J., 2011, May. Evaluating swabbing: a touchscreen input method for elderly users with tremor. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 623-626). ACM.
20. Nicolau, H. and Jorge, J., 2012, October. Elderly text-entry performance on touchscreens. In Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility (pp. 127-134). ACM.
21. Sato, D., Morimura, T., Katsuki, T., Toyota, Y., Kato, T. and Takagi, H., 2016, December. Automated help system for novice older users from touchscreen gestures. In Pattern Recognition (ICPR), 2016 23rd International Conference on (pp. 3073-3078). IEEE.
22. Wobbrock, J.O., Kane, S.K., Gajos, K.Z., Harada, S. and Froehlich, J., 2011. Ability-based design: Concept, principles and examples. ACM Transactions on Accessible Computing (TACCESS), 3(3), p.9.
23. Gajos, K.Z., Weld, D.S. and Wobbrock, J.O., 2010. Automatically generating personalized user interfaces with Supple.
24. Trewin, S., Keates, S. and Moffatt, K., 2006, October. Developing steady clicks:: a method of cursor assistance for people with motor impairments. In Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility (pp. 26-33). ACM.
25. Wobbrock, J.O., Fogarty, J., Liu, S.Y.S., Kimuro, S. and Harada, S., 2009, April. The angle mouse: target-agnostic dynamic gain adjustment based on angular deviation. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1401-1410). ACM.
26. Sato, D., Kobayashi, M., Takagi, H., Asakawa, C. and Tanaka, J., 2011, October. How voice augmentation supports elderly web users. In The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility (pp. 155-162). ACM.
27. Weiner, J., Engelbart, M. and Schultz, T., 2017. Manual and Automatic Transcriptions in Dementia Detection from Speech. Proc. Interspeech 2017, pp.3117-3121.
28. Rudzicz, F., Wang, R., Begum, M. and Mihailidis, A., 2014. Speech recognition in Alzheimer's disease with personal assistive robots. In Proceedings of the 5th Workshop on Speech and Language Processing for Assistive Technologies (pp. 20-28).
29. Zajicek, M., 2006. Aspects of HCI research for older people. Universal Access in the Information Society, 5(3), pp.279-286.
30. Ienca, M., Fabrice, J., Elger, B., Caon, M., Pappagallo, A.S., Kressig, R.W. and Wangmo, T., 2017. Intelligent assistive technology for Alzheimer's disease and other dementias: a systematic review. Journal of Alzheimer's Disease, 56(4), pp.1301-1340.
31. Granata, C., Chetouani, M., Tapus, A., Bidaud, P. and Dupourqué, V., 2010, September. Voice and graphical-based interfaces for interaction with a robot dedicated to elderly and people with cognitive disorders. In RO-MAN, 2010 IEEE (pp. 785-790). IEEE.
32. Ziman, R. and Walsh, G., 2018. Factors Affecting Seniors' Perceptions of Voice-enabled User Interfaces. In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18), 6 pages.
33. Vipperla, R., Wolters, M., Georgila, K. and Renals, S., 2009, July. Speech input from older users in smart environments: Challenges and perspectives. In International Conference on Universal Access in Human-Computer Interaction (pp. 117-126). Springer.
34. Bost, J. and Moore, J.D., 2014. An Analysis of Older Users' Interactions with Spoken Dialogue Systems. In LREC (pp. 1176-1181).

35. Cucchiarini, C., Hamme, H.V., Herwijnen, O.V. and Smits, F., 2006. Jasmin-CGN: Extension of the spoken dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality.
36. Rösner, D.F., Frommer, J., Friesen, R., Haase, M., Lange, J. and Otto, M., 2012, May. LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions. In LREC (pp. 2559-2566).
37. Rösner, D., Frommer, J., Wendemuth, A., Bauer, T., Günther, S., Haase, M. and Siegert, I., 2017. The LAST MINUTE Corpus as a Research Resource: From Signal Processing to Behavioral Analyses in User-Companion Interactions. In Companion Technology (pp. 277-299). Springer, Cham.
38. Wolters, M.K., Kilgour, J., MacPherson, S.E., Dzikovska, M. and Moore, J.D., 2015, April. The CADENCE corpus: a new resource for inclusive voice interface design. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 3963-3966). ACM.
39. Rudzicz, F., Chan Currie, L., Danks, A., Mehta, T. and Zhao, S., 2014, October. Automatically identifying trouble-indicating speech behaviors in Alzheimer's disease. In Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility (pp. 241-242). ACM.
40. Chinaei, H., Currie, L.C., Danks, A., Lin, H., Mehta, T. and Rudzicz, F., 2017. Identifying and avoiding confusion in dialogue with people with alzheimer's disease. Computational Linguistics, 43(2), pp.377-406.
41. Kirshner, H.S., 2012. Primary progressive aphasia and Alzheimer's disease: brief history, recent evidence. Current neurology and neuroscience reports, 12(6), pp.709-714.
42. MacKay, D.G., James, L.E. and Hadley, C.B., 2008. Amnesic HM's performance on the language competence test: Parallel deficits in memory and sentence production. Journal of Clinical and Experimental Neuropsychology, 30(3), pp.280-300.
43. Van Velzen, M. and Garrard, P., 2008. From hindsight to insight–retrospective analysis of language written by a renowned Alzheimer's patient. Interdisciplinary Science Reviews, 33(4), pp.278-286.
44. Oulhaj, A., Wilcock, G.K., Smith, A.D. and de Jager, C.A., 2009. Predicting the time of conversion to mci in the elderly role of verbal expression and learning. Neurology, 73(18), pp.1436-1442.
45. Petersen, R.C., Caracciolo, B., Brayne, C., Gauthier, S., Jelic, V. and Fratiglioni, L., 2014. Mild cognitive impairment: a concept in evolution. J Intern Med, 275(3), pp.214-228.
46. Mueller, K.D., Koscik, R.L., LaRue, A., Clark, L.R., Hermann, B., Johnson, S.C. and Sager, M.A., 2015. Verbal fluency and early memory decline: results from the Wisconsin Registry for Alzheimer's prevention. Archives of Clinical Neuropsychology, 30(5), pp.448-457.
47. Bertola, L., Mota, N.B., Copelli, M., Rivero, T., Diniz, B.S., Romano-Silva, M.A., Ribeiro, S. and Malloy-Diniz, L.F., 2014. Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, mild cognitive impairment and normal elderly controls. Frontiers in aging neuroscience, 6, p.185.
48. Lundholm, K.F., Fraser, K. and Kokkinakis, D., 2018. Automated Syntactic Analysis of Language Abilities in Persons with Mild and Subjective Cognitive Impairment. Studies in health technology and informatics, 247, pp.705-709.
49. Henry, J.D., Crawford, J.R. and Phillips, L.H., 2004. Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. Neuropsychologia, 42(9), pp.1212-1222.
50. Kavé, G. and Goral, M., 2018. Word retrieval in connected speech in Alzheimer's disease: a review with meta-analyses. Aphasiology, 32(1), pp.4-26.

51. Lunsford, R. and Heeman, P.A., 2015. Using linguistic indicators of difficulty to identify mild cognitive impairment. In Sixteenth Annual Conference of the International Speech Communication Association.
52. Toth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatloczki, G., Banreti, Z., Pákáski, M. and Kalman, J., 2018. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. Current Alzheimer Research, 15(2), pp.130-138.
53. Ahmed, S., de Jager, C.A., Haigh, A.M. and Garrard, P., 2013. Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease. Neuropsychology, 27(1), p.79.
54. Fraser, K.C., Meltzer, J.A. and Rudzicz, F., 2016. Linguistic features identify Alzheimer's disease in narrative speech. Journal of Alzheimer's Disease, 49(2), pp.407-422.
55. Sajjadi, S. A., Patterson, K., Tomek, M., and Nestor, P. J. (2012). Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. Aphasiology, 26(6):847–866.
56. Croisile, B., Ska, B., Brabant, M.J., Duchene, A., Lepage, Y., Aimard, G. and Trillet, M., 1996. Comparative study of oral and written picture description in patients with Alzheimer's disease. Brain and language, 53(1), pp.1-19.
57. Ahmed, S., Haigh, A.M.F., de Jager, C.A. and Garrard, P., 2013. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. Brain, 136(12), pp.3727-3737.
58. Natale, M., Entin, E. and Jaffe, J., 1979. Vocal interruptions in dyadic communication as a function of speech and social anxiety. J Pers Soc Psychol, 37(6), p.865.
59. Brewer, R., Garcia, R.C., Schwaba, T., Gergle, D. and Piper, A.M., 2016. Exploring traditional phones as an e-mail interface for older adults. TACCESS, 8(2), p.6.
60. Barthel Activities of Daily Living (ADL) Index. (1993). Occasional paper (Royal College of General Practitioners), (59), 24.
61. Folstein, M.F., Folstein, S.E. and McHugh, P.R., 1975. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. Journal of psychiatric research, 12(3), pp.189-198.
62. Dubois, B., Slachevsky, A., Litvan, I. and Pillon, B.F.A.B., 2000. The FAB: a frontal assessment battery at bedside. Neurology, 55(11), pp.1621-1626.
63. Wechsler, D., 1945. A standardized memory scale for clinical use. The Journal of Psychology, 19(1), pp.87-95.
64. Wechsler, D., 1984. WMS-R: Wechsler memory scale-revised: manual. Psychological Corporation.
65. Reitan, R.M., 1958. Validity of the Trail Making Test as an indicator of organic brain damage. Perceptual and motor skills, 8(3), pp.271-276.
66. Stuss, D.T. and Levine, B., 2002. Adult clinical neuropsychology: lessons from studies of the frontal lobes. Annual review of psychology, 53(1), pp.401-433.
67. Alexa Design Guide, https://developer.amazon.com/docs/alexa-design/intro.html, last accessed 2019-01-25.
68. Conversation Design, https://designguidelines.withgoogle.com/conversation/conversation-design/, last accessed 2019-01-25
69. Cortana design guidelines, https://docs.microsoft.com/en-us/cortana/voice-commands/voicecommand-design-guidelines, last accessed 2019-01-25
70. Raux, A. and Eskenazi, M., 2008. Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. In Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue (SIGdial '08), pp.1-10.