HHS Public Access

Author manuscript

Comput Speech Lang. Author manuscript; available in PMC 2023 March 01.

Published in final edited form as:

Comput Speech Lang. 2022 March ; 72: . doi:10.1016/j.csl.2021.101297.

Evaluating Voice-Assistant Commands for Dementia Detection

Xiaohui Liang[a,*], **John A. Batsis**[b], **Youxiang Zhu**[a], **Tiffany M. Driesse**[b], **Robert M. Roth**[c], **David Kotz**[d], **Brian MacWhinney**[e]

[a]Department of Computer Science, University of Massachusetts Boston, 100 Morrissey Blvd., Boston, MA 02125-3393 USA

[b]Division of Geriatric Medicine, University of North Carolina at Chapel Hill, 5017 Old Clinic Building, Chapel Hill, NC 27599 USA

[c]Department of Psychiatry, Geisel School of Medicine at Dartmouth/DHMC, Lebanon, NH 03756 USA

[d]Department of Computer Science, Dartmouth College, Hanover, NH 03755 USA

[e]Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue Pittsburgh, PA 15213 US

Abstract

Early detection of cognitive decline involved in Alzheimer's Disease and Related Dementias (ADRD) in older adults living alone is essential for developing, planning, and initiating interventions and support systems to improve users' everyday function and quality of life. In this paper, we explore the voice commands using a Voice-Assistant System (VAS), i.e., Amazon Alexa, from 40 older adults who were either Healthy Control (HC) participants or Mild Cognitive Impairment (MCI) participants, age 65 or older. We evaluated the data collected from voice commands, cognitive assessments, and interviews and surveys using a structured protocol. We extracted 163 unique command-relevant features from each participant's use of the VAS. We then built machine-learning models including 1-layer/2-layer neural networks, support vector machines, decision tree, and random forest, for classification and comparison with standard cognitive assessment scores, e.g., Montreal Cognitive Assessment (MoCA). Our classification models using fusion features achieved an accuracy of 68%, and our regression model resulted in a Root-Mean-Square Error (RMSE) score of 3.53. Our Decision Tree (DT) and Random Forest (RF) models using selected features achieved higher classification accuracy 80–90%. Finally, we analyzed the contribution of each feature set to the model output, thus revealing the commands and features most useful in inferring the participants' cognitive status. We found that features of overall performance, features of music-related commands, features of call-related commands, and

*Corresponding author.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

features from Automatic Speech Recognition (ASR) were the top-four feature sets most impactful on inference accuracy. The results from this controlled study demonstrate the promise of future home-based cognitive assessments using Voice-Assistant Systems.

Keywords

Voice Assistant; Cognitive Decline; Speech Analysis; Machine Learning; Alzheimer's Disease

1. Introduction

In 2020, an estimated 5.8 million Americans age 65 and older were living with Alzheimer's Disease (AD). Family members and friends provided nearly $244 billion in unpaid care to people with AD [2, 3]. While pharmacotherapy is available, it has not been proven to alter the underlying pathophysiological processes leading to dementia. Until effective disease-modifying therapies become available, early detection of cognitive decline is important to permit long-term planning and initiate non-pharmacological interventions that can significantly impact a participant's functional trajectory, their quality of life, and importantly the family and caregiver's social support system.

Speech deficits in AD have been well documented [45, 49, 53], and as the disease progresses, communication skills degrade further with deficits in both production [32] and comprehension of language [15, 26, 39, 41, 51, 36]. Such challenges are reflected in communication breakdowns in everyday interactions [52] and increased frustration, which may result in challenging behaviors [55]. Speech is a rich and ubiquitous source of cognitive data, where computational speech analysis has the potential to aid clinicians in early and accurate diagnosis of dementia [31, 42, 16]. A large body of research aims to study the language samples elicited in more structured ways, such as through open-ended questions or semi-structured interviews [22]. While open-ended elicitation methods may provide a larger quantity of output, they can be highly variable within and across individuals and contexts and thus cannot be easily standardized for between- and within-group comparisons.

One well-known speech-dementia study is the Pitt Corpus, collected on 104 Healthy Control (HC) and 208 participants with AD, longitudinally, on a yearly basis from 1984 to 2006 [13]. The Cookie Theft Picture (CTP) description can be evaluated with standardized measures, and if the picture is visible throughout the task, it relies less on episodic memory, which is a core deficit in AD and common in Mild Cognitive Impairment (MCI) and other dementias. The CTP dataset in Pitt Corpus contains 243 audio files from HC participants and 309 audio files from participants with AD. The Pitt Corpus is publicly available to the research community [8] and widely used to validate speech-dementia methods. In 2020, the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) Challenge was the first shared-task event focused on AD detection using Pitt Corpus [38]. Other data collection efforts include Wisconsin Longitudinal Study [7, 29], conversation with neurologists [20], and conversations with Intelligent Virtual Agent [43].

Researchers have exploited classification and regression models to infer cognitive status. A classification model of cognitive assessment is to classify participants in three groups of

HC, MCI, AD participants. The grouping can be done with the use of scores from cognitive screening measures, such as the Mini-Mental State Examination (MMSE) or Montreal Cognitive Assessment (MoCA). It is commonly believed that the ability of a classifier to distinguish between HC and AD is higher than the ability to distinguish between HC and MCI, as the speech deficits in AD should be more severe than those in MCI. Nonetheless, the accurate classification between HC and MCI would suggest that features of voice samples could be used for early detection of cognitive decline, which would be critical for the implementation of early intervention to prevent or slow down further cognitive decline. In addition, a more complicated task is to create a regression model to infer the cognitive test scores, such as from the MMSE or MoCA [57, 38].

In this paper, we explore a newly collected speech dataset from 40 older adults (age ≥ 65) using a Voice-Assistant System (VAS), Amazon Alexa. In general, a VAS enables users to speak voice commands to interact with a large number of in-home and third-party services over their smartphones, tablets, computers, and smart speakers [18]. VAS has become increasingly popular in recent years and can improve quality of life in older adults [6, 47, 56, 50, 10, 46, 9]. The total audience of smart speakers in the US reached 54.4 million in 2018 [33]; 22% of owners are age > 55 [35]; and more than 60% of owners use smart speakers every day, with an average of 2.79 uses per day (0.33 for smartphone VAS) [34]. We aim to leverage VAS to develop a low-cost, passive, and practical home-based cognitive assessment method. The voice commands used with a VAS are a special type of speech initiated by users and sparsely distributed over time; their content, language, and pattern may be drastically different across users. Furthermore, VAS data has a unique data structure compared to the previous speech dataset [8, 7, 29, 20, 43]; it consists of daily-used spontaneous commands initiated by users and intended for seeking assistance on daily tasks from a computer with artificial intelligence. Finally, we note that VAS data has been previously exploited for the determination of physical and emotional characteristics [30]. Here, we report our preliminary results from a controlled study. We analyze the transcript and audio data collected by an Alexa device in a controlled setting. Our goal is to explore novel features extracted from the transcript and audio of the voice commands and investigate whether voice commands and their unique features are sensitive to the difference in cognitive functioning in older adults. In this paper, we make three contributions, as follows.

First, we collected Alexa transcript and audio data from 40 participants over 30 selected commands. The 40 participants include 18 HC and 22 MCI, grouped using their MoCA scores. We have extracted 163 unique command-relevant features from the Alexa transcript and audio and analyzed the features of participants' performance.

Second, we implemented Machine-Learning (ML) models of classification and regression, including 1-hidden-layer Neural Network (1NN), 2-hidden-layer Neural Network (2NN), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF), and we identified the features that are most impactful to the model output.

Third, we discussed our plans for future research direction, including features, longitudinal home-based evaluation, as well as identifying the scope and limitations of the present work.

2. Method

In this section, we introduce the evaluation environment, participants, commands, data, features, and models.

2.1. Environment and participants

In a controlled study, participants were invited to perform a set of specific tasks under the instructions given by the research investigators. Our evaluation was conducted remotely due to the COVID-19 pandemic, as shown in Figure 1. The study and evaluation activities were approved by the local Institutional Review Board (IRB). A Research Assistant (RA) set up a laptop, an Alexa device, and smart-home devices physically in an office. The RA then used the laptop to set up a Zoom session with participants who used their own computers. The participant's computer received the audio of the commands from the participant and transmitted the audio to the RA's laptop over the Zoom session. The RA's laptop then played the command audio at a speaker device close to the Alexa device. The Alexa device received the command audio and played the response audio at its speaker. The RA's laptop received the response audio and transmitted it to the participant's computer over the Zoom session. In such a way, the participant successfully interacted with the remote Alexa on the RA's side. During the session, the participant's computer and the RA's laptop share a screen of the 30 commands, a camera view of participant's face, and a camera view of the smart home devices. The remote setting enables participation by remote participants and requires no control of the participants' computers and networks. The RA closely monitored the participant-to-Alexa interaction to ensure that it was effective. As a result, the average number of accomplished commands (see Section 2.4.1) per participant is 27.45 out of 30, and the average number of commands per participant is 42.125.

2.2. 30 commands on voice assistant systems

We selected 30 commands in five groups and displayed them on a shared screen over the Zoom session, as shown in Table 1. We selected the 30 commands using the following steps.

1. We studied the smart speaker consumer adoption report by voicebot.ai in March 2018, which includes the smart speaker use case frequency [1].
2. We then filtered the popular Alexa commands for older adults by surveying relevant news and reports [4, 5].
3. We finally selected 30 commands that can be reliably performed in our lab environment.

Each participant is required to perform the 30 commands in a fixed order at the remote Alexa device. If a participant failed at one command, the participant might make further attempts right after the failed attempt or after finishing the whole round at the RA's request. In each participant's session, the RA first introduced the environment, protocol, and the devices, then demonstrated sample commands at the Alexa device, and provided assistance and explanations as needed. Participants were instructed to wait for the Alexa device to finish the response of the current command before moving to the next command.

Rationale for 30-command instruction.—We thought about letting participants freely select Alexa commands in the evaluation. However, we decided not to do so because we considered VAS technology is relatively new, and most users (especially older adults) use a very limited number of commands. Without pre-defined commands, participants with Alexa usage experience may generate effective but limited commands; and participants without such experience may generate unsuccessful commands and have frustrating moments. Furthermore, the Automatic Speech Recognition (ASR) and Natural Language Unit (NLU) behind the Alexa might not be able to process commands that have utterances not related to the Alexa skills. As the first stage of such a new project, clear instruction of 30 commands would i) help participants understand what Alexa is capable of; ii) produce data effectively, especially for those with cognitive problems; iii) enable us to receive feedback on the usefulness of these commands; and iv) enable us to compare their performance and understand the difficulty levels of different commands. We also understand the advantage of spontaneous speech and the real scenario, so in our in-home evaluation, we will collect longitudinal data from participants (who were determined to have positive Alexa usage experience in the in-lab evaluation) and let them freely select their own commands in our in-home evaluation.

2.3. Transcript and audio datasets

We focus on analyzing the Alexa audio and the Alexa transcript, which are directly downloaded from the Alexa servers. These two types of data are available for analysis of cognitive decline in a real-world home setting. On the one hand, these data represent the ability of the Alexa device to collect and interpret the participants' commands. On the other hand, given that the Alexa device employs the same algorithms and mechanism, these data, in our vision, may objectively reflect the speech ability and cognitive ability of the participants. While we have collected Zoom recordings of the entire session, we do not analyze the Zoom recordings because they will not be available for a home-based setting.

The Alexa server provides the data for each detected command, i.e., a short period of the recording after a wake-up phrase is detected. In other words, the Alexa server provides the audio of the command, the transcript of the command converted using Automatic Speech Recognition (ASR), and the transcript of the response by the Alexa. We organized the transcript of both the Alexa commands and the Alexa responses in the CHAT format, introduced by MacWhinney et al. [40]. An example is shown in Figure 2, where *PAR represents the participant's Alexa command and %xvas represents Alexa's response. Each participant's command is indexed and associated with audio downloaded from the Alexa server. Based on the five groups of the 30 commands, we wrote a program to auto-group the commands into five groups. We used colorful labels to distinguish different groups and manually confirmed the grouping process. If a command is partially recognized or unrecognized, e.g., 39 and 45 in Figure 2, we manually grouped the commands based on the partial transcript, corresponding Alexa audio, the context, and the Zoom recordings.

We denote two datasets of 40 participants as $D_t = \{d_{t,1}, d_{t,2}, \ldots, d_{t,40}\}$ and $D_a = \{d_{a,1}, d_{a,2}, \ldots, d_{a,40}\}$ to represent the Alexa transcripts and audio, respectively. Both D_t and D_a contain the data of participants' commands but do not contain the Alexa responses. The Alexa

server provides the transcript of the responses but not the audio of responses (the audio was generated via a standard text-to-speech algorithm). We do not include the transcripts of Alexa responses in D_t because the Alexa responses are similar given the same command as inputs, which are not related to participants' speech or cognitive abilities. Any data related to the RA's demonstration and assistance has been removed from the datasets D_t and D_a.

2.4. Baseline collection

We collected the Alexa audio and transcript from 40 older adults (18 HC and 22 MCI). They provided information on their demographic characteristics and medical history. In addition, they were administered the Callahan assessment, MoCA, Older Americans Resources and Services (OARS), Geriatric Anxiety Inventory (GAI), and Geriatric Depression Scale (GDS). As shown in Table 2, based on the MoCA scores, 40 participants are grouped into a HC group (total score on the MoCA ≥ 26) and a MCI group (total score on the MoCA < 26). In the HC group, the number of participants is even in males and females. However, in the MCI group, female participants are significantly more than male participants. While there was a greater percentage of women within the MCI (68%) than HC (50%) sample, this was not statistically significant [$\chi^2(1) = 1.36$, $p = .24$].

2.5. Extracting features

In this section, we study the features of overall performance, grouped commands, specific commands, and ASR.

2.5.1. Features of overall performance A-1 to A-5—

We obtained the features of the overall performance by checking each participant's commands in D_t and D_a.

Feature A-1. Number, duration, mean, and standard deviation of participant's commands.: We examined the Alexa transcript D_t and counted the number of participant's commands that Alexa recorded (excluding the RA's) as n_p. We observed that for the commands partially recognized or unrecognized by Alexa, the provided transcript contained some texts or a special Alexa message, such as "audio could not be understood." We still counted such commands as they might be relevant to the participant's speech ability. We then cross-checked the Alexa transcript with the Alexa audio. We added the duration of all audio data corresponding to the participant's commands to a duration t_p. We further calculated the mean and deviation $\bar{t}_p = t_p/n_p$ and $\sigma_p = \left(\frac{\sum (t_i - \bar{t}_p)^2}{n_p}\right)^{1/2}$ where t_i is the duration of one command from the participant.

Feature A-2. Number, duration, mean, and standard deviation of matched commands.: We cross-checked the Alexa transcript with the 30 commands (Table 1). We define a matched command as a command in the transcript D_t that accurately matches with any of the 30 commands. For example, "Alexa, what is the weather outside" shown in the transcript is considered as a matched command because the command string is accurately matched with the first of the 30 commands. If a matched command is repeated in the transcript, it has been counted multiple times. We added the duration of all audio

data corresponding to the participant's matched commands to a duration t_m. Similarly, we calculated mean $\bar{t}_m$ and standard deviation σ_m.

Feature A-3. Number, duration, mean, and standard deviation of unmatched but recognized commands.: We cross-checked the Alexa transcript with the 30 commands. We define a recognized command as a command that is recognized by the Alexa ASR, i.e., the transcript of the command is not empty in D_t. Recognized commands can be either matched or unmatched commands. We counted unmatched and recognized commands to a number $n_{um,r}$. For example, "Alexa, use volume eight" differs from the standard "Alexa, volume eight", "Alexa, how do you make chocolate chip cookies?" from "Alexa, how do you bake chocolate chip cookies," or "Alexa, call" from "Alexa, call 603-XXX-XXXX." We added the duration of all audio data corresponding to the participant's unmatched but recognized commands to a duration $t_{um,r}$. We further calculated mean $\bar{t}_{um,r}$ and standard deviation $\sigma_{um,r}$.

Feature A-4. Number, duration, mean, and standard deviation of unrecognized commands.: We cross-checked the Alexa transcript with the 30 commands. We define an unrecognized command as a command that cannot be recognized by Alexa ASR and the corresponding transcript is empty. We denote this number as n_u. For example, Alexa shows "[]", "[audio could not be understood]", "[audio was not intended for Alexa]", and "[no text stored]". We added the duration of all audio data corresponding to the participant's unrecognized commands to a duration t_u. We further calculated mean $\bar{t}_u$ and standard deviation σ_u.

Feature A-5. Number, duration, mean, and standard deviation of accomplished commands.: We cross-checked the Alexa transcript with the 30 commands. We searched the 30 commands, and counted the number of commands that appear in the transcript. Note that if a matched command is repeated, only the first one has been counted. The maximum value of n_a is 30. n_a is an indicator of the accomplishment level of the standard 30 commands. We then cross-checked the Alexa transcript with the Alexa audio. We added the duration of all audio data corresponding to the participant's accomplished commands to a duration t_a. We further calculated mean $\bar{t}_a$ and standard deviation σ_a.

For each participant, we have $n_p = n_m + n_{um,r} + n_u$ and $t_p = t_m + t_{um,r} + t_u$. In addition, $0 \leq n_a \leq 30$ represents the percentage of the 30 commands that have been accomplished by the participants.

2.5.2. Features of grouped commands—

In this section, we further extracted the features of grouped commands. We first divided the 30 commands into five groups. Group 1 consists of commands 1–8, related to questions and answers. Group 2 consists of commands 9–12; participants play music at Alexa and adjust the volume of the sound. Group 3 consists of commands 13–20, related to reminders, alarm, timer, and list. These commands assist users in their daily living tasks. Group 4 consists of commands 21–22; participants can call a phone number or call a phone's name to make the phone ring. Group 5 consists of commands 23–30 related to the control of smart-home devices.

Through an auto-grouping program and a manual grouping process, we separated the Alexa transcript D_t and the Alexa audio D_a into five groups, denoted by $(D_{t,i},D_{a,i})_{1\le i\le 5}$, where each group contains data of commands in the corresponding group. Then, we can apply features A-1 to A-5 on each group data $(D_{t,i},D_{a,i})_{1\le i\le 5}$:

Feature i-1. $(n_{p,i},t_{p,i},\bar{t}_{p,i},\sigma_{p,i})$ are the number, duration, mean, and standard deviation of commands in ith-group.

Feature i-2. $(n_{m,i},t_{m,i},\bar{t}_{m,i},\sigma_{m,i})$ are the corresponding features of matched commands.

Feature i-3. $(n_{um,r,i},t_{um,r,i},\bar{t}_{um,r,i},\sigma_{um,r,i})$ are the corresponding features of unmatched but recognized commands.

Feature i-4. $(n_{u,i},t_{u,i},\bar{t}_{u,i},\sigma_{u,i})$) are the corresponding features of unrecognized commands.

Feature i-5. $(n_{a,i},t_{a,i},\bar{t}_{a,i},\sigma_{a,i})$) are the corresponding features of accomplished commands.

Since the five groups are mutually exclusive, the sum of the number and duration feature values of the five groups are the feature values of the overall performance, e.g., $n_p=\sum_{1\le i\le 5}n_{p,i}$, $t_p=\sum_{1\le i\le 5}t_{p,i}$.

2.5.3. Features of specific commands—

We observed the performance on commands 16 and 21 are highly different across participants, and we further extracted features of these two specific commands.

Command 16 “Alexa, remember my daughter’s birthday is June first” triggers Alexa to generate follow-up questions and engages participants in a multi-round conversation. The first response from Alexa is, “you want me to note my daughter’s birthday is June first, right?” If the participant replies, “right”, Alexa further responds, “okay, noted. I can also remind you on June one at nine a.m.. do you want me to do that?” As the follow-up questions are not shown on the 30-command list, participants may generate different performances regarding command 16.

Command 21 requires participants to call a 10-digit phone number. However, we observed a certain number of participants had made multiple attempts to finish this command. One main reason is that when participants have large silent pauses between numbers, Alexa may stop listening and respond with an error message. We further applied A-1 to A-5 on command x = 16, 21 and obtained the following.

Feature x-1. $(n_{p,x},t_{p,x},\bar{t}_{p,x},\sigma_{p,x})$ are the number, duration, mean, and standard deviation of command x.

Feature x-2. $(n_{m,x},t_{m,x},\bar{t}_{m,x},\sigma_{m,x})$ are the corresponding features of matched commands.

Featurex-3. $(n_{um,r,x},t_{um,r,x},\bar{t}_{um,r,x},\sigma_{um,r,x})$arethecorrespondingfeaturesofunmatchedbutrecognizedcommands.

Feature x-4. ($n_{u,x}$,$t_{u,x}$, $\bar{t}_{u,x}$, $\sigma_{u,x}$) are the corresponding features of unrecognized commands.

Feature x-5. ($n_{a,x}$,$t_{a,x}$, $\bar{t}_{a,x}$, $\sigma_{a,x}$) are the corresponding features of accomplished commands.

2.5.4. Features from automatic speech recognition—We obtained the Alexa transcripts that were auto-generated by the Alexa system. We have no access and no knowledge of the Alexa ASR, which could be updated by the Alexa team continuously. By comparing the Alexa transcripts with the 30 pre-defined commands, we observed that the average number of accomplished commands (exact match) per participant is 27.45 out of 30, and the average number of total commands per participant is 42.125 (if failed at some commands, participants were required to repeat them for additional attempts). This demonstrated that participants followed the pre-defined commands, and the Alexa ASR was effective in the transcription process. In our project, we used the Alexa transcript as the baseline. We further used the state-of-the-art open-source Wav2vec ASR to generate another transcript from the Alexa audio recording. Then, we analyzed the difference between the Alexa transcript and the Wav2vec transcript. We consider if the voice quality is high, both Alexa ASR and Wav2vec ASR will produce similar transcripts; if the voice quality is low, the Alexa ASR will produce transcript much more accurate than the transcript from the Wav2vec ASR because the Wav2vec ASR is a general ASR algorithm, not specifically designed for Alexa commands. Our intuition is if the two transcripts are more different, the voice quality is lower; if the two transcripts are less different, the voice quality is higher. In the following, we studied three features by co-analyzing D_t' and D_t [44]. Let H, S,D and I denote the total number of word hits, substitutions, deletions, and insertions when comparing aligned D_t and D_t'. Let N_1,N_2 and N denote the total number of words in D_t, the total number of words in D_t', and total number of pairs between D_t and D_t'. We thus have $N = H + S + D + I$, $N_1 = H + S + D$, and $N_2 = H + S + I$. Word Error Rate (WER) is defined as the proportion of word errors to words processed, i.e., $\frac{S+D+I}{N_1} = \frac{S+D+I}{H+S+D}$, normalized WER as $\frac{S+D+I}{max(N_1,N_2)}$, and Match Error Rate (MER) as $\frac{S+D+I}{H+S+D+I} = 1 - \frac{H}{N}$. We envision that this proposed ASR analysis method is the first by comparing two different ASR results and can be adopted for a large-scale longitudinal evaluation.

2.5.5. Classification and regression models—Our goal of classification is to infer the participant's cognitive group, either HC or MCI, based on the feature vectors. The classification model takes a feature vector x as input and outputs an inference y, where y is a score between 0 and 1, representing a prediction estimate that the feature vector came from a HC participant (coded as 0) or a MCI participant (coded as 1). The final classification is obtained by thresholding this score (default threshold is 0.5). As shown in the top part of Figure 3, we first adopt an *early fusion strategy* to concatenate the feature values from overall performance, grouped commands, and specific command, and ASR into a single feature vector for each participant and trained the ML model to infer the output class. We implemented the SVM, DT, 1NN, 2NN, and RF models. We then adopted a *late fusion strategy*, as shown in the bottom part of Figure 3. We trained nine classifiers separately over feature vectors of overall performance, the performance of each group, specific commands,

and ASR. The nine classifiers have exclusive groups of feature vectors as inputs. We aggregate the outputs by either voting (for binary output 0,1) or averaging (for probability or score outputs) to produce the final inference result. In the *early fusion strategy*, we selected different combination of feature vectors and compare the ML results. The higher the ML accuracy is, the more impactful the features are to the cognitive assessment. In the *late fusion strategy*, we assessed the accuracy of each individual classifier to observe which single feature is the most impactful on inferring the cognitive statuses.

In addition to the classifiers, we built regression models to infer the MoCA scores, where both the early fusion strategy and the late fusion strategy were adopted. Similarly, the regression models were implemented with SVM, DT, 1NN, 2NN, and RF models.

All ML models were implemented with Python and the scikit-learn library. For 1NN, we used a hidden layer with 5 neurons; For 2NN, we used two hidden layers with 10 and 5 neurons, respectively. We trained the 1NN and 2NN with a maximum of 10000 epoch and an initial learning rate of 0.0001. For RF, we used a 10-tree setting. All other parameters followed the default values of the scikit-learn library.

3. Results

In this section, we first report the feature values and then show the ML results of classification and regression.

3.1. Feature values

We collected 1685 Alexa commands from the 40 participants, 955 commands from 22 MCI participants and 730 from 18 HC. On average, one MCI participant generated 43.41 commands, and one HC participant generated 40.56 commands. In total, MCI participants generated 63 unrecognized commands and HC participants generated 11 unrecognized commands. The average WER of 22 MCI participants is 0.27, while the average WER of 18 HC participants is 0.22.

Figure 4a shows the numbers of accomplished commands and total commands. The average number of accomplished commands per participant is 27.45 out of 30. And the average number of total commands per participant is 42.125. Thus, it appears most participants were able to interact with the remote Alexa device effectively. Participants (8, 20, 24, 38, 40) generated the most commands, while their MoCA scores are (23, 21, 17, 24, 25), grouped in the MCI group. We consider that HC participants generated fewer commands than MCI participants because they would more clearly and carefully speak the commands and incur less command failure. The number of accomplished commands of participants (8, 36, 39) were the lowest and their MoCA scores are (23, 24, 25), grouped in the MCI group.

Figure 4b shows the numbers of matched, unmatched but recognized, and unrecognized commands. Participants (8, 14, 24, 40) have the most unrecognized commands and their MoCA scores are (23, 22, 17, 25), grouped into the MCI group. Participants (1, 3, 7, 8, 30, 39) have the least matched commands and their MoCA scores are (20, 10, 23, 23, 26, 25), five of which were grouped into the MCI group. We found that participants with more

unrecognized commands or fewer matched commands were more likely to belong to the MCI group. Unmatched but recognized commands can be caused by various reasons, e.g., the participants may misread some words, leave a long pause between words causing Alexa to stop recording, or attempt to try some new commands. Due to mixed reasons, the number of this type of commands may produce a positive or negative impact on our analysis.

Figure 5a and Figure 5b show the numbers of the matched commands, unmatched but recognized, and unrecognized commands in group 3 and command 16, respectively. Command 16 is one of many commands in group 3.

From Figure 5a, participants (8, 24, 35) generated the most commands on group 3; all were in the MCI group. In addition, we found participant 10 (HC group) had few matched commands. By checking at the transcript of participant 10, we found “Alexa set my alarm for 7 am tomorrow” was split to “Alexa set my alarm” and “Alexa set my alarm” and “seven am”; and “oranges” was interpreted as “orange juice”. These two cases lead to the decrease of the number of matched commands and the increase of the number of unmatched but recognized commands. While observing some similar cases on other participants, we were not able to find a direct relation between the number of matched commands in group 3 and the cognitive status.

From Figure 5b, participants (9, 14, 36) were the only ones who did not generate the matched command, and their MoCA scores are (29, 22, 24), two MCI and one HC. We found participant 9’s command 16 is “Alexa remember my daughter’s birthday is June the first ” where “the” is an additional word compared to our standard commands. As matched commands need to accurately match one of the 30 commands, this command is not a matched command but an unmatched but recognized command. Participant 17 generated a significant more commands than other participants, and has an MoCA score 30, the highest in the HC group. We found participant 17 actually listened to the Alexa response “you want me to note my daughter’s birthday is June first, right?” and replied, “not your daughter but my daughter.” Though the conversation is not as expected, this can be considered as a positive sign of cognitive status. We further looked at other participants’ performance; most generated one matched command and two unmatched but recognized commands, which are not helpful in differentiating their performance.

Figure 6a and Figure 6b show the numbers of the matched commands, unmatched but recognized, and unrecognized commands in group 4 and command 21, respectively. Command 21 is one of two commands in group 4. We observed that the performance across participants on command 21 is consistent with the performance across participants on group 4. By checking the matched commands, 15 participants did not generate the matched command, and 10 of 15 are grouped in the MCI group. By checking the total commands, participants (8, 16, 20, 38) generated the most commands in both group 4 and command 21, and their MoCA scores are (23, 24, 21, 26), three MCI and one HC. The feature values on command 21 appeared more consistent with the MoCA scores, compared to command 16. We thus expected that the features of command 21 and group 4 are more effective for inferring the cognitive status than those of command 16 and group 3.

Figure 7a shows the WER by comparing the transcripts from Wav2Vec ASR algorithm with the Alexa transcript. Participants (8, 10, 24, 38) have the highest WER and their MoCA scores are (23, 28, 17, 24). Three participants belong to the MCI group and one participant belongs to the HC group. We hypothesized the WER is associated to the quality of the command audio and the higher the WER, the more likely the participant belongs to the MCI group. Lastly, Figure 7b shows the MoCA scores of all 40 participants, and score 26 was the threshold to group HC and MCI.

For the classification task, we evaluated the accuracy, that is, $\frac{TN+TP}{N}$ where N is the number of participants, TP and TN are the numbers of true positives and true negatives, respectively. We also evaluated the precision $\pi = \frac{TP}{TP+FP}$, recall $\rho = \frac{TP}{TP+FN}$, and F1 score $\frac{2\pi\rho}{\pi+\rho}$, where N is the number of participants, FP and FN are the numbers of true positives, false positives and false negatives, respectively. For the regression task, we employed the Root-Mean-Square Error (RMSE), which is a frequently used measure of the differences between values produced by a model or an estimator and the values observed [28, 38]. Due to the limited size of the dataset, we adopted a Leave-One-Subject-Out (LOSO) cross-validation setting where the training data do not contain any information from validation subjects. We reported the accuracy of classification on the left of Figure 8 and the RMSE scores on the right of Figure 8.

We first built nine classifiers, each using a single set of features, including overall, group 1–5, commands 16&21, and ASR. We mainly studied the mean of the classification results from five different ML models. We observed the classification accuracy of a single set of features reached the highest 59% when using either the features of overall performance or the features of group 4. The features of overall performance are related to all commands and represent a comprehensive evaluation of participants' performance; the features of group 4 are related to the call commands. The classification accuracy reached 55% when using features of command 21. As command 21 belongs to group 4, we concluded that the participants' performance on "call" command is impactful in inferring the cognitive status. Then, we studied the early fusion strategy and late fusion strategy that utilized all 163 features. While the late fusion strategy reaches 56%, the early fusion strategy achieves classification accuracy 64%, higher than any of the nine classifiers and the late fusion strategy. The performance gain of the early fusion strategy confirms the complementary information among these sets of features. We further exploited the combination of features for a higher classification accuracy where the ML models might be over-fitted. We showed the top-2 performances for each ML model and the associated features. We observed that the top-2 results of DT were 90% and 88%, and the top-2 results of RF were 80% and 80%. Features of command 21 and ASR appear in these four combinations. In addition, features of overall performance and group 2 appear in 7 out of total 10 combinations. Thus, we concluded that features of overall performance, features of groups 2, features of call-related command, and features from ASR were the four feature sets, most impactful on classification accuracy.

We studied the regression results from a single set of features and observed that the best regression results were from features of overall performance, features of groups 2 and 4, features of command 21, features of ASR. Both the early fusion and late fusion strategies do not outperform the results from ASR alone. The regression models were trained with MoCA scores, which may lead to more variability in the models. We envision the regression results will improve and the model will output more stable results with additional data. Then, we studied the feature combinations and observed the top-3 of features are overall for 5 times, group 2 for 6 times, and ASR for 5 times.

In sum, by jointly analyzing the ML results on classification and regression, we concluded that the features of overall performance, features of groups 2, features of call-related command, and features from ASR were the four feature sets that are most impactful to the cognitive assessments.

We further showed the precision, recall, and F1 score of five models using the early fusion strategy in Table 4. Most of the classifiers achieved balanced precision and recall (except for 1NN), which demonstrated the effectiveness of the extracted features. For 1NN, its performance may be limited due to a relatively-high dimension of features (163) and the small size of the neuron (5). When we placed one more hidden layer to extend 1NN to 2NN, the classification results were significantly improved from 53% to 68%. Also, RF achieved more balanced F1 scores than the DT by ensembling more trees in high dimension feature space, while their accuracy remained the same at 68%. We believe the ensembling processes reinforced the robustness of the classifier in high dimension feature space with limited samples.

4. Discussion

In this section, we discussed the features, home-based evaluation, and other challenges in our evaluation.

4.1. Linguistic and acoustic feature

Existing speech-dementia studies have sought to extract linguistic and acoustic features from speech samples to explore any possible early indication of cognitive issues. In the Cookie Theft picture description [14], topic-based features refer to a concept in the image, e.g., woman, sink, overflowing, and more. Participants were given credit for mentioning a given topic [19, 17, 12, 11, 23, 24]. In our controlled study, to enable a fair and effective data collection, we selected the 30 commands and aim to examine participants' performance over the same set of commands. On the one hand, the collected transcript and audio data across participants are highly comparable; on the other hand, the transcript data across participants are highly similar. Thus, topic-based features do not apply to our evaluation as the topics in the transcript were kept the same. Another feature, perplexity, indicates how well an utterance spoken by a participant can be understood by a Language Model (LM). This feature has been studied in recent works [37, 21, 27], but it does not apply to our evaluation as the standard commands are pre-defined, and participants are required to perform the standard commands and have limited freedom in constructing their own commands. In addition, one important set of features are silent pauses and filled pauses.

In 2015, Lunsford et al. found that the speakers with impairment, as compared to those who are cognitively intact, spent more time engaged in verbalized hesitations (e.g., “and um …”) prior to speaking story content and that these verbalized hesitations accounted for a larger ratio of the time spent retelling. In the same year, Toth et al. [54] considered four descriptors for the silent pauses, the filled pauses, the silent and filled pauses together, and the phonemes. In ADReSS 2020, Yuan et al. [58] found that “um” was used much less frequently in Alzheimer’s speech, compared to “uh.” However, pauses features do not apply to our data for two observed reasons, i) the commands are usually short, simple, and familiar to participants. Pauses occur rarely; ii) the pauses will incur command failure at the Alexa device because the Alexa device will stop recording when a certain duration of the pause is detected. For example, in command 21, participants encountered command failure by leaving pauses between digits of a phone number. As such, the occurrence of pauses in the commands will result in unmatched but recognized commands or even unrecognized commands. In such a way, the pause information will not be available in the Alexa transcript and audio data.

4.2. Longitudinal home-based evaluation

In a home-based evaluation, the participants’ interaction with the VAS device at home will be spontaneous. Such interaction between participants and the VAS device will be direct in close physical proximity, initiated by the participants, and have minimum interference and interruption from external factors. In the home settings, the categories, topics, and commands will be different across participants. Topic-based features and perplexity features may be useful in a home-based VAS dataset. In future work, we plan to study repetition features: some commands may be repeatedly performed on a weekly basis, a daily basis, or even an hourly basis. Studies show that a user with dementia may be unaware of the time elapsed and thus forget about having posed questions [25, 48]. The considered repetition features include: (high-level) changes of the frequency, interval, content, and utterance of the repeated task; (mid-level) the changes of duration and pausing interval of repeated sentences; and (low-level) the changes of voice pitch, duration, and the recall time of repeated words. With the knowledge and experience from this controlled study and our upcoming study (aimed at 90 participants), we will have a better understanding of the data about commands and their association to the cognitive status across participants.

4.3. Other challenges.

Our classification and regression models are not expected to be as accurate as clinical assessment results. Our study is limited in that participants may have varying factors affecting their cognitive status such as cerebrovascular disease, and the participants in our study had limited representation across race, education, and acculturation. In addition, the detection accuracy of our system in this stage will be limited by the accuracy of the cognitive assessment results, e.g., MoCA scores, as they are used as labels. This is a preliminary study, and further testing on a larger number of human subjects would be needed. Ultimately, our system is not meant to supplant traditional means of diagnosis; we envision that our system would, upon a repeated classification of a formerly HC person as MCI, suggest follow-up assessments by referral to a clinical provider; the system can also collect valuable voice evidence for clinicians to augment their existing diagnostic

approaches. Our method could be deployed in a large-scale community-based setting with minimal usability issues to assist in those who are already suspected of experiencing cognitive decline and want to gather further information for verification, but do not want a higher-cost or more invasive method, such as MRI and a spinal tap, at the time. Our method could also be deployed in senior-living housing for longitudinal awareness of the onset of cognitive decline among residents. If successful, such deployments could have a significant impact on mitigating the dementia problem in the older adult community.

Acknowledgments

This document is the results of the research project funded by the US National Institutes of Health Grant No. 1R01AG067416.

List of Acronyms

5.

ADRD	Alzheimer’s Disease and Related Dementias
AD	Alzheimer’s Disease
MCI	Mild Cognitive Impairment
HC	Healthy Control
CTP	Cookie Theft Picture
ASR	Automatic Speech Recognition
ML	Machine-Learning
MMSE	Mini-Mental State Examination
MoCA	Montreal Cognitive Assessment
GDS	Geriatric Depression Scale
GAI	Geriatric Anxiety Inventory
SVM	Support Vector Machine
OARS	Older Americans Resources and Services
ADReSS	Alzheimer’s Dementia Recognition through Spontaneous Speech
VAS	Voice-Assistant System
RA	Research Assistant
WER	Word Error Rate
MER	Match Error Rate
1NN	1-hidden-layer Neural Network

2NN	2-hidden-layer Neural Network
DT	Decision Tree
RF	Random Forest
RMSE	Root-Mean-Square Error
LOSO	Leave-One-Subject-Out
IRB	Institutional Review Board

References

[1]. Smart speaker consumer adoption report, URL https://voicebot.ai/wp-content/uploads/2018/10/voicebot-smart-speaker-consumer-adoption-report.pdf.

[2]. 2019 Alzheimer's disease facts and figures. Alzheimer's Association.

[3]. 2020 Alzheimer's disease facts and figures. Alzheimer's Association.

[4]. Alexa guide for seniors: 14 ways older adults can use amazon echo devices, URL https://www.vivint.com/resources/article/alexa-guide-for-seniors.

[5]. Amazon echo and alexa for the elderly, URL https://www.techenhancedlife.com/explorers/amazon-echo-and-alexa-elderly.

[6]. Amazon echo for dementia: Technology for seniors. http://dailycaring.com/amazon-echo-for-dementia-technology-for-seniors/.

[7]. Wisconsin longitudinal study. URL https://www.ssc.wisc.edu/wlsresearch/.

[8]. Dementia bank. https://dementia.talkbank.org/. [Supported by NIH-NIDCD grant R01-DC008524 for 2007–2017].

[9]. Introducing lisa by cuida health. https://cuidahealth.com/lisa/#1538938947069-a80a08ef-5f2a, 2018.

[10]. Amazon echo and alexa for the elderly. https://www.techenhancedlife.com/explorers/amazon-echo-and-alexa-elderly, 2018.

[11]. Ahmed S, de Jager CA, Haigh A-M, and Garrard P. Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease. Neuropsychology, 27(1):79, 2013. [PubMed: 23356598]

[12]. Ahmed S, Haigh A-MF, de Jager CA, and Garrard P. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. Brain, 136(12):3727–3737, 2013. [PubMed: 24142144]

[13]. Becker JT, Boiler F, Lopez OL, Saxton J, and McGonigle KL. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. Archives of Neurology, 51(6):585–594, 1994. [PubMed: 8198470]

[14]. Becker JT, Boller F, Lopez OL, Saxton J, and McGonigle KL. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. Archives of Neurology, 51(6):585–594, 1994. [PubMed: 8198470]

[15]. Bickel C, Pantel J, Eysenbach K, and Schröder J. Syntactic comprehension deficits in Alzheimer's disease. Brain and Language, 71(3): 432–448, 2000. [PubMed: 10716871]

[16]. Braaten AJ, Parsons TD, McCUE R, Sellers A, and Burns WJ. Neurocognitive differential diagnosis of dementing diseases: Alzheimer's dementia, vascular dementia, frontotemporal dementia, and major depressive disorder. International Journal of Neuroscience, 116(11):1271–1293, 2006.

[17]. Bschor T, Kühl K-P, and Reischies FM. Spontaneous speech of patients with dementia of the Alzheimer type and mild cognitive impairment. International psychogeriatrics, 13(3):289–298, 2001. [PubMed: 11768376]

[18]. Canalys RF. 56 million smart speaker sales in 2018 says canalys. https://www.voicebot.ai/2018/01/07/56-million-smart-speaker-sales-2018-says-canalys/.

[19]. Croisile B, Ska B, Brabant M-J, Duchene A, Lepage Y, Aimard G, and Trillet M. Comparative study of oral and written picture description in patients with Alzheimer's disease. Brain and language, 53(1):1–19, 1996. [PubMed: 8722896]

[20]. Elsey C, Drew P, Jones D, Blackburn D, Wakefield S, Harkness K, Venneri A, and Reuber M. Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics. Patient Education and Counseling, 98(9):1071–1077, 2015. [PubMed: 26116418]

[21]. Frankenberg C, Weiner J, Schultz T, Knebel M, Degen C, Wahl H-W, and Schroeder J. Perplexity–a new predictor of cognitive changes in spoken language?–results of the interdisciplinary longitudinal study on adult development and aging (ilse). Linguistics Vanguard, 5(s2), 2019.

[22]. Frankenberg C, Weiner J, Knebel M, Abulimiti A, Toro P, Herold CJ, Schultz T, and Schröder J. Verbal fluency in normal aging and cognitive decline: Results of a longitudinal study. Computer Speech & Language, page 101195, 2021.

[23]. Fraser KC, Meltzer JA, and Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. Journal of Alzheimer's Disease, 49(2):407–422, 2016.

[24]. Fraser KC, Linz N, Li B, Fors KL, Rudzicz F, König A, Alexandersson J, Robert P, and Kokkinakis D. Multilingual prediction of Alzheimer's disease through domain adaptation and concept-based language modelling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3659–3670, 2019.

[25]. Grewal RP. Awareness of time in dementia of the Alzheimer type. Psychological reports, 76(3):717–718, 1995. [PubMed: 7568581]

[26]. Grossman M, D'Esposito M, Hughes E, Onishi K, Biassou N, White-Devine T, and Robinson KM. Language comprehension profiles in Alzheimer's disease, multi-infarct dementia, and frontotemporal degeneration. Neurology, 47(1):183–189, 1996. [PubMed: 8710075]

[27]. Guo Z, Ling Z, and Li Y. Detecting Alzheimer's disease from continuous speech using language models. Journal of Alzheimer's Disease, 70(4):1163–1174, 2019.

[28]. Haider F, De La Fuente S, and Luz S. An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech. IEEE Journal of Selected Topics in Signal Processing, 14(2):272–281, 2019.

[29]. Herd P, Carr D, and Roan C. Cohort profile: Wisconsin longitudinal study (wls). International journal of epidemiology, 43(1):34–41, 2014. [PubMed: 24585852]

[30]. Jin H and Wang S. Voice-based determination of physical and emotional characteristics of users, 2018. US Patent 15/457,846.

[31]. Johnson S. Speech changes, pauses may be first signs of Alzheimer's. URL https://www.alzheimers.net/speech-changes-may-be-first-signs-of-{Alzheimer}s/.

[32]. Kemper S, Thompson M, and Marquis J. Longitudinal change in language production: effects of aging and dementia on grammatical complexity and propositional content. Psychology and aging, 16(4):600, 2001. [PubMed: 11766915]

[33]. Kinsella B. Smart speaker users pass 50 million in u.s. for the first time. https://voicebot.ai/2018/06/28/smart-speaker-users-pass-50-million-in-u-s-for-the-first-time/, 2018.

[34]. Kinsella B. Smart speaker owners use voice assistants nearly 3 times per day. https://voicebot.ai/2018/04/02/smart-speaker-owners-use-voice-assistants-nearly-3-times-per-day/, 2018.

[35]. Kinsella B. Smart speaker owner demographics are getting younger as market nearly tripled in 12 months. https://voicebot.ai/2018/07/18/smart-speaker-owner-demographics-are-getting-younger-as-market-nearly-tripled-in-12-months/, 2018.

[36]. Kirshner HS. Primary progressive aphasia and Alzheimer's disease: brief history, recent evidence. Current neurology and neuroscience reports, 12(6):709–714, 2012. [PubMed: 22932755]

[37]. Linz N, Tröger J, Lindsay H, Konig A, Robert P, Peter J, and Alexandersson J. Language modelling for the clinical semantic verbal fluency task. 2018.

[38]. Luz S, Haider F, de la Fuente S, Fromm D, and MacWhinney B. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. arXiv preprint arXiv:2004.06833, 2020.

[39]. MacDonald MC, Almor A, Henderson VW, Kempler D, and Andersen ES. Assessing working memory and language comprehension in Alzheimer's disease. Brain and language, 78(1):17–42, 2001. [PubMed: 11412013]

[40]. MacWhinney B. The CHILDES project: The database, volume 2. Psychology Press, 2000.

[41]. Martin A and Fedio P. Word production and comprehension in Alzheimer's disease: The breakdown of semantic knowledge. Brain and language, 19(1):124–141, 1983. [PubMed: 6860932]

[42]. McCullough KC, Bayles KA, and Bouldin ED. Language performance of individuals at risk for mild cognitive impairment. Journal of Speech, Language, and Hearing Research, 62(3):706–722, 2019.

[43]. Mirheidari B, Blackburn D, Walker T, Reuber M, and Christensen H. Dementia detection using automatic analysis of conversations. Computer Speech & Language, 53:65–79, 2019.

[44]. Morris AC, Maier V, and Green P. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In Eighth International Conference on Spoken Language Processing, 2004.

[45]. Mueller KD, Hermann B, Mecollari J, and Turkstra LS. 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, 2018. [PubMed: 29669461]

[46]. MUTCHLER A. Cuida health launches lisa, a voice-based social wellness app for seniors. https://voicebot.ai/2018/10/24/cuida-health-launches-lisa-a-voice-based-social-wellness-app-for-seniors/, 2018.

[47]. O'Brien E. Older adults buddy up with amazon's alexa. www.marketwatch.com, 2016.

[48]. Papagno C, Allegra A, and Cardaci M. Time estimation in Alzheimer's disease and the role of the central executive. Brain and Cognition, 54(1):18–23, 2004. [PubMed: 14733896]

[49]. Price BH, Gurvit H, Weintraub S, Geula C, Leimkuhler E, and Mesulam M. Neuropsychological patterns and language deficits in 20 consecutive cases of autopsy-confirmed Alzheimer's disease. Archives of neurology, 50(9):931–937, 1993. [PubMed: 8363447]

[50]. Rieland R. Alexa? how voice-first technology helps older adults. www.forbes.com, 2018.

[51]. Ross GW, Cummings JL, and Benson DF. Speech and language alterations in dementia syndromes: Characteristics and treatment. Aphasiology, 4(4):339–352, 1990.

[52]. Savundranayagam MY and Orange JB. Matched and mismatched appraisals of the effectiveness of communication strategies by family caregivers of persons with Alzheimer's disease. International Journal of Language & Communication Disorders, 49(1):49–59, 2014. [PubMed: 24372885]

[53]. Taler V and Phillips NA. Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. Journal of clinical and experimental neuropsychology, 30(5):501–556, 2008. [PubMed: 18569251]

[54]. Tóth L, Gosztolya G, Vincze V, Hoffmann I, Szatlóczki G, Biró E, Zsura F, Pákáski M, and Kálmán J. Automatic detection of mild cognitive impairment from spontaneous speech using asr. In Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[55]. Woodward M. Aspects of communication in Alzheimer's disease: clinical features and treatment options. International psychogeriatrics, 25 (6):877–885, 2013. [PubMed: 23522497]

[56]. Woyke E. The octogenarians who love amazon's alexa. MIT Technology Review, 2017.

[57]. Yancheva M, Fraser K, and Rudzicz F. Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. In Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 134–139, 2015.

[58]. Yuan J, Bian Y, Cai X, Huang J, Ye Z, and Church K. Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease. Proc. Interspeech 2020, pages 2162–2166, 2020.

Highlights

- We explored the voice commands using a Voice-Assistant System (VAS), such as Amazon Alexa, from 40 older adults who were either Healthy Control (HC) patents or Mild Cognitive Impairment (MCI) patents, age 65 or older.
- We collected 1685 Alexa commands from the 40 patents, 955 commands from 22 MCI, and 730 from 18 HC. On average, one MCI patent generated 43.41 commands, and one HC patent generated 40.56 commands. In total, MCI patents generated 63 unrecognized commands and HC patents generated 11 unrecognized commands.
- We extracted 163 unique command-relevant features from each patient's use of the VAS, including overall performance, grouped commands, specific commands, and Automatic Speech Recognition (ASR).
- Our classification models using fusion features achieved an accuracy of 68%, and our regression model resulted in a Root-Mean-Square Error (RMSE) score of 3.53. Our Decision Tree and Random Forest models using selected features achieved higher classification accuracy 80–90%.
- We analyzed the contribution of each feature set to the model output, thus revealing the commands and features most useful in inferring the patient's cognitive status. We found that features of overall performance, features of music-related commands, features of call-related commands, and features from ASR were the top-four feature sets most impactful on inference accuracy.
- We discussed our plans for future research direction, including features, longitudinal home-based evaluation, as well as identifying the scope and limitations of the present work.

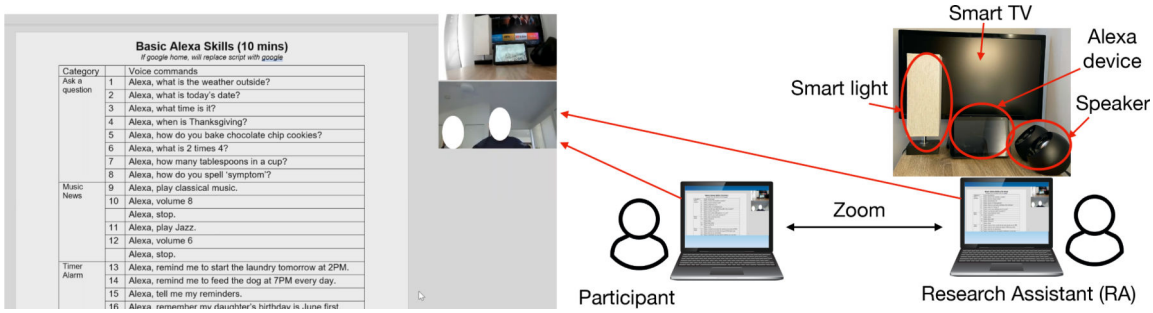


Figure 1:

Remote setting to evaluate participants' performance of the Alexa interaction. Participants follow a 30-command instruction to interact with an Alexa set up in the RA's location.

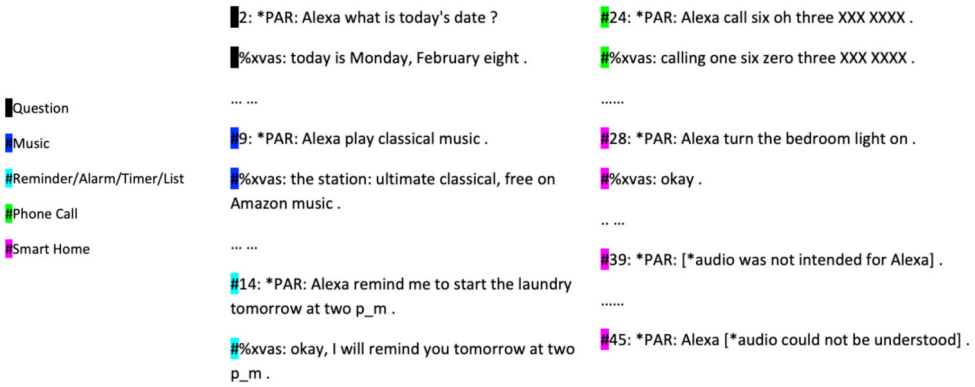


Figure 2:

An example of transcript after auto grouping and manual grouping. *PAR represents the participant's Alexa command; %xvas represents the Alexa response that has no corresponding audio.

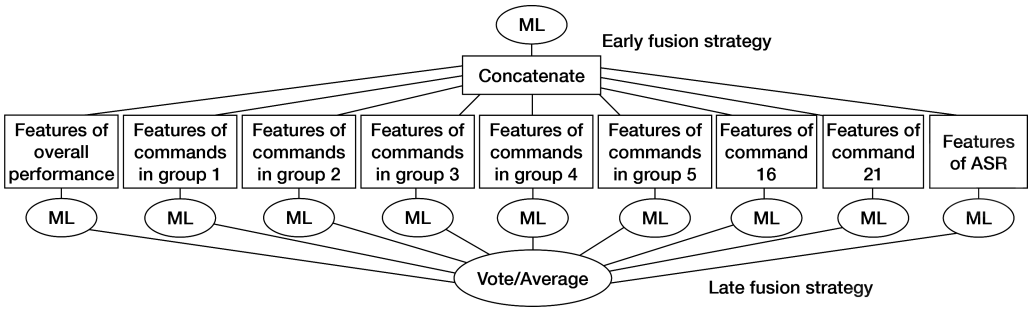
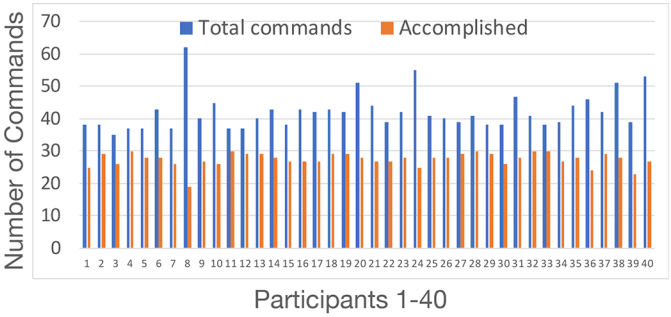
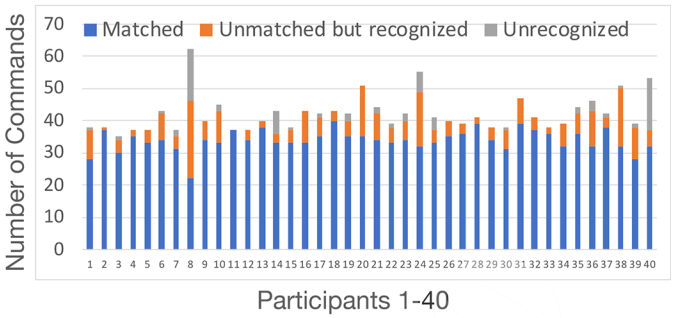


Figure 3:

Early fusion strategy and late fusion strategy of machine-learning models: an early fusion strategy concatenates all feature values into a single feature vector, and a late fusion strategy trains classifiers separately and aggregates the classifiers' outputs.



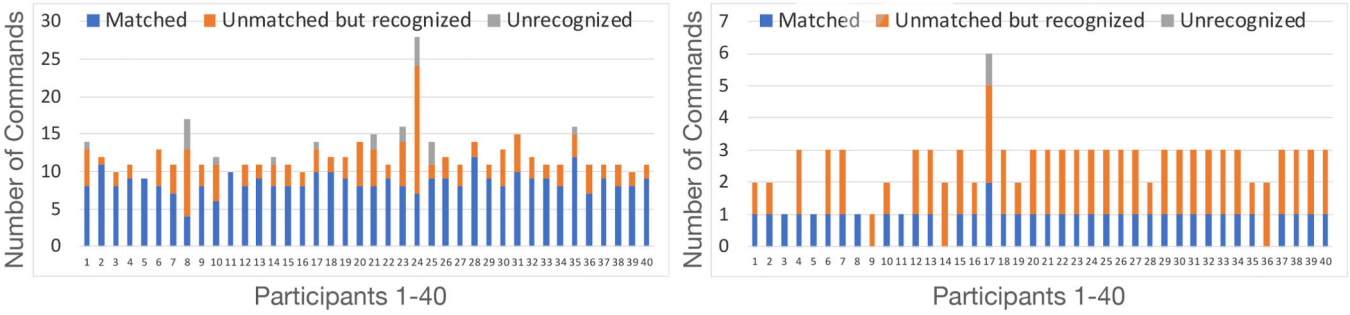
(a) Numbers of accomplished and total commands.



(b) Numbers of three types of commands

Figure 4:

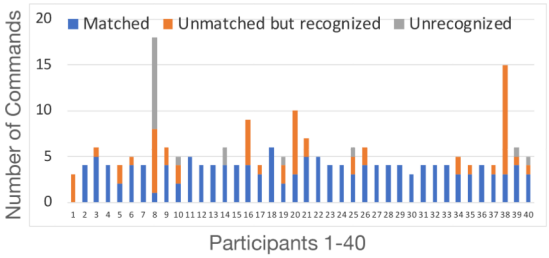
Overall performance of 30 commands from the 40 participants. Total commands mean all commands of one participant that Alexa recorded; accomplished commands mean those among the 30 commands that appear in the Alexa transcript (count ≤ 30); matched commands mean those appear in transcript and match with any of the 30 commands; unmatched but recognized commands mean those appear in the transcript but do not match with any of the 30 commands; unrecognized commands mean those have audio only but do not appear in the transcript.



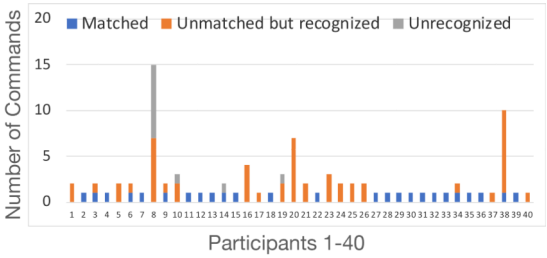
(a) Numbers of three types of commands in group 3

(b) Numbers of three types of commands of command 16

Figure 5:
Performance on Group 3 and Command 16 from the 40 participants

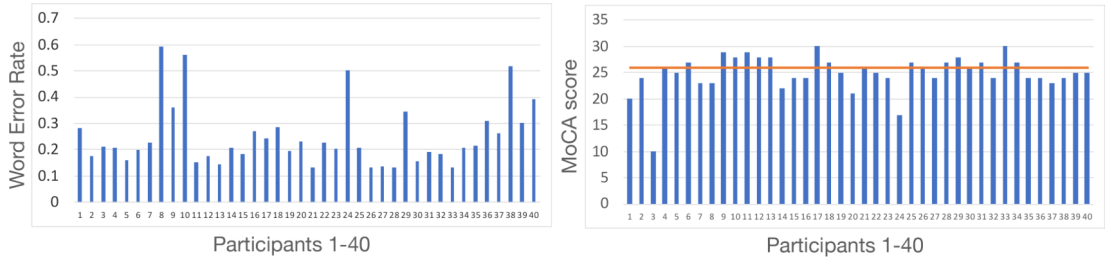


(a) Numbers of three types of commands in group 4



(b) Numbers of three types of commands of command 21

Figure 6:
Performance on Group 4 and Command 21 from the 40 participants



(a) Word error rate (b) MoCA scores

Figure 7:
Word error rate and MoCA scores from the 40 participants

Classification						
Features	SVM	DT	1NN	2NN	RF	Mean
Overall (20)	0.62	0.55	0.60	0.55	0.63	0.59
Group 1 (20)	0.58	0.45	0.53	0.43	0.53	0.50
Group 2 (20)	0.45	0.63	0.50	0.53	0.45	0.51
Group 3 (20)	0.45	0.48	0.45	0.35	0.30	0.41
Group 4 (20)	0.63	0.55	0.55	0.63	0.58	0.59
Group 5 (20)	0.50	0.55	0.48	0.48	0.58	0.52
Command 16 (20)	0.45	0.40	0.43	0.58	0.43	0.46
Command 21 (20)	0.53	0.65	0.48	0.55	0.55	0.55
ASR (3)	0.63	0.48	0.55	0.43	0.53	0.52
Early (163)	0.63	0.68	0.53	0.68	0.68	0.64
Late (163)	0.63	0.63	0.48	0.53	0.55	0.56
Feature Combination						
O/G2/C16	**0.70**	0.55	0.70	0.65	0.68	0.66
O/G1,2,4	**0.70**	0.48	0.60	0.70	0.70	0.64
G2/C21/ASR	0.63	**0.90**	0.53	0.58	0.73	0.67
G3/C21/ASR	0.38	**0.88**	0.38	0.43	0.65	0.54
O/G2,3,4,5/C16	0.65	0.48	**0.78**	0.58	0.60	0.62
O/G5	0.63	0.45	**0.73**	0.50	0.58	0.58
G2,4/C16	0.68	0.45	0.68	**0.78**	0.58	0.63
O/G2,4/C21	0.65	0.65	0.60	**0.78**	0.60	0.66
O/G1,2,4/C21/ASR	0.68	0.75	0.63	0.68	**0.80**	0.71
O/G1,2,5/C21/ASR	0.65	0.65	0.58	0.55	**0.80**	0.65

Regression						
Feature	SVM	DT	1NN	2NN	RF	Mean
Overall (20)	3.58	4.76	7.70	6.81	3.83	5.34
Group 1 (20)	3.76	6.20	10.00	11.70	4.42	7.22
Group 2 (20)	3.60	3.77	8.89	10.17	3.36	5.96
Group 3 (20)	3.73	6.25	9.83	10.27	5.12	7.04
Group 4 (20)	3.62	3.98	7.03	7.59	4.32	5.31
Group 5 (20)	3.59	5.18	12.17	12.37	3.98	7.46
Command 16 (20)	3.68	7.62	13.60	7.96	5.37	7.65
Command 21 (20)	3.74	5.58	6.07	12.18	4.14	6.34
ASR (3)	3.49	4.85	3.82	5.10	4.02	4.25
Early (163)	3.53	4.26	13.49	15.55	3.99	8.16
Late (163)	3.60	4.17	4.72	5.43	3.78	4.34
Feature Combination						
O/G2	**3.36**	4.63	9.82	17.11	3.50	7.68
O/G2,4	**3.35**	4.75	16.40	14.12	3.76	8.48
O/G1,2/C16/ASR	3.58	**3.22**	14.25	11.31	3.44	7.16
O/G2/C16/ASR	3.52	**3.27**	18.09	7.95	3.48	7.26
ASR	3.49	4.85	**3.82**	5.10	4.02	4.25
C16/ASR	3.61	6.77	**6.02**	15.96	4.80	7.43
ASR	3.49	4.85	3.82	**5.10**	4.02	4.25
O	3.58	4.76	7.70	**6.81**	3.83	5.34
G2/C21	3.52	3.97	13.30	22.27	**3.22**	9.26
G2,3/C21/ASR	3.48	4.04	19.43	15.11	**3.26**	9.06

Figure 8:

Machine learning results. The left part shows the classification results and the right part shows the regression results. **SVM**: Support Vector Machine; **DT**: Decision Tree; **1NN**: 1-layer Neural Network; **2NN**: 2-layer Neural Network; **RF**: Random Forest; **O**: Overall; **G**: Group; **C**: Command; **Early**: early fusion; **Late**: late fusion; **(*)**: Number of features.

Table 1

30 commands of 5 categories shared with participants in the evaluation

	Question		Music
1	Alexa, what is the weather outside?	9	Alexa, play classical music.
2	Alexa, what is today's date?	10	Alexa, volume 8.
3	Alexa, what time is it?		Alexa, stop.
4	Alexa, when is Thanksgiving?	11	Alexa, play Jazz.
5	Alexa, how do you bake chocolate chip cookies?	12	Alexa, volume 6.
6	Alexa, what is 2 times 4?		Alexa, stop.
7	Alexa, how many tablespoons in a cup?		
8	Alexa, how do you spell 'symptom'?		
	Reminder/Alarm/Timer/List		**Phone Call**
13	Alexa, remind me to start the laundry tomorrow at 2pm.	21	Alexa, call (XXX)-XXX-XXXX.
14	Alexa, remind me to feed the dog at 7pm everyday.		Alexa, hang up.
15	Alexa, tell me my reminders.	22	Alexa, find my phone.
16	Alexa, remember my daughter's birthday is June first.		Alexa, quit.
17	Alexa, set a timer in 5 seconds.		
	Alexa, stop.		
18	Alexa, set my alarm for 7am tomorrow.		
19	Alexa, add oranges and grapes to my shopping list.		
20	Alexa, what is in my shopping list?		
	Smart Home		
23	Alexa, turn the bedroom light on.	27	Alexa, open the kitchen camera.
24	Alexa, turn the bedroom light red.	28	Alex, hide the kitchen camera.
25	Alexa, change brightness to 10.	29	Alexa, play White Collar on Fire TV
26	Alexa, turn off the bedroom light.	30	Alexa, pause

Table 2

Participants demographic characteristics. NA: Not Available.

	Healthy Control			Mild Cognitive Impairment		
Age	**Male**	**Female**	**MoCA**	**Male**	**Female**	**MoCA**
[65, 70)	2	4	27.33	3	4	23.14
[70, 75)	4	4	27.75	3	8	23.36
[75, 80)	3	1	27.5	1	1	23
≥80	0	0	NA	0	2	17.5
Total	9	9	27.56	7	15	22.73

Table 3

Features of overall performance of the 30 commands.

Index	Feature description
A-1	Number, duration, mean, and deviation of participant's commands (n_p, t_p, $\bar{t}_p$, δ_p)
A-2	Number, duration, mean, and deviation of matched commands (n_m, t_m, $\bar{t}_p$, δ_p)
A-3	Number, duration, mean, and deviation of unmatched and recognized commands ($n_{um,r}$, $t_{um,r}$, $\bar{t}_p$, δ_p)
A-4	Number, duration, mean, and deviation of unrecognized commands (n_u, t_u, $\bar{t}_p$, δ_p)
A-5	Number, duration, mean, and deviation of accomplished commands (n_a, t_a, $\bar{t}_p$, δ_p)

Table 4

Classification results using early fusion strategy over 163 features

Model	Class	Precision	Recall	F1 Score	Accuracy
SVM	HC	0.60	0.50	0.55	0.63
	MCI	0.64	0.73	0.68	
DT	HC	0.63	0.67	0.65	0.68
	MCI	0.71	0.68	0.70	
1NN	HC	0.48	0.72	0.58	0.53
	MCI	0.62	0.36	0.46	
2NN	HC	0.67	0.56	0.61	0.68
	MCI	0.68	0.77	0.72	
RF	HC	0.62	0.72	0.67	0.68
	MCI	0.74	0.64	0.68	