


Acceptability of collecting speech samples from the elderly via the telephone

Digital Health
Volume 7: 1–10
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076211002103
journals.sagepub.com/home/dhj


Catherine Diaz-Asper¹ , Chelsea Chandler², R Scott Turner³ ,
Brigid Reynolds³ and Brita Elvevåg⁴

Abstract

Objective: There is a critical need to develop rapid, inexpensive and easily accessible screening tools for mild cognitive impairment (MCI) and Alzheimer's disease (AD). We report on the efficacy of collecting speech via the telephone to subsequently develop sensitive metrics that may be used as potential biomarkers by leveraging natural language processing methods.

Methods: Ninety-one older individuals who were cognitively unimpaired or diagnosed with MCI or AD participated from home in an audio-recorded telephone interview, which included a standard cognitive screening tool, and the collection of speech samples. In this paper we address six questions of interest: (1) Will elderly people agree to participate in a recorded telephone interview? (2) Will they complete it? (3) Will they judge it an acceptable approach? (4) Will the speech that is collected over the telephone be of a good quality? (5) Will the speech be intelligible to human raters? (6) Will transcriptions produced by automated speech recognition accurately reflect the speech produced?

Results: Participants readily agreed to participate in the telephone interview, completed it in its entirety, and rated the approach as acceptable. Good quality speech was produced for further analyses to be applied, and almost all recorded words were intelligible for human transcription. Not surprisingly, human transcription outperformed off the shelf automated speech recognition software, but further investigation into automated speech recognition shows promise for its usability in future work.

Conclusion: Our findings demonstrate that collecting speech samples from elderly individuals via the telephone is well tolerated, practical, and inexpensive, and produces good quality data for uses such as natural language processing.

Keywords

Aging, telephone interview, automated speech recognition, cognitive screening, acceptability

Submission date: 15 October 2020; Acceptance date: 17 February 2021

Introduction

With an aging population, the number of elderly people diagnosed with dementias such as Alzheimer's disease (henceforth AD) is projected to escalate rapidly.¹ The diagnosis of probable AD requires a comprehensive clinical examination, which can be both expensive and time consuming. This examination typically necessitates a visit to a clinic or medical center, which raises issues of accessibility and can result in unequal access to services.^{2–4} In fact, estimates suggest that only 5% of community-dwelling elderly in the

¹Department of Psychology, Marymount University, Arlington, VA, USA

²Department of Computer Science, University of Colorado Boulder, CO, USA

³Department of Neurology, Georgetown University, Washington, DC, USA

⁴Department of Clinical Medicine, University of Tromsø, Tromsø- the Arctic University of Norway, Norway

Corresponding author:

Catherine Diaz-Asper, Department of Psychology, Marymount University, 2807 N. Glebe Road, Arlington, VA 22207-4299, USA.
Email: cdiazasp@marymount.edu



United States with memory concerns (without frank dementia) receive a clinical examination,⁴ and only 16% of seniors receive regular cognitive screening assessments as part of their primary care.¹

Hence, there is a pressing need to develop reliable and accessible screeners at the preclinical stage, to maximize treatment success, and increase quality of life for affected individuals and their caregivers by reducing uncertainty. While numerous cognitive screening tools exist, telephone-based screeners provide an attractive option to address barriers to in-person attendance. These tests have the advantage of being easily accessible, inexpensive, well tolerated and sensitive to the diagnosis of frank dementia.⁵⁻⁷ Furthermore, they appear largely accepted by clinicians in the field. For example: our group conducted a pilot study examining the feasibility of telephone-based screening of cognitive decline in elderly Norwegians, and reported that users (both general practitioners and the elderly) were generally positive towards such a service (7). Indeed, an evaluation of 14 studies published between 1988 and 2005 by Martin-Khan et al.⁶ concluded that currently-available telephone-based cognitive assessments can be used as screening tools, or for monitoring over time, and more recently, these assessments have been shown to correlate strongly with face-to-face assessment.⁸

The gold standard in telephone screeners, the Telephone Interview for Cognitive Status (TICS⁹), surveys multiple domains of cognition and is modeled on the widely-used Mini Mental State Examination (MMSE¹⁰). While multi-domain screeners such as the TICS are a useful tool to diagnose dementia, they tend to fall short in the identification of the earliest stages of cognitive decline, with considerably lower sensitivity and specificity for the diagnosis of mild cognitive impairment (MCI) relative to frank dementia.^{11,12}

One aspect of cognition with demonstrated promise in the detection of early decline is language function. Deficits in speech and language, particularly semantic knowledge, are a characteristic of AD¹³ and have been retrospectively detected in language samples years prior to the emergence of overt symptomatology,¹⁴⁻¹⁷ signaling their potential as a screener for early cognitive decline. For example, analyses of speech transcripts from president Ronald Reagan revealed evidence of decline in language function that predated his diagnosis of AD, demonstrated by a decrease in the number of unique words produced over time and an increase in conversational fillers and non-specific nouns.¹⁶ Thus, it follows logically that a telephone-based screening tool that targets speech specifically has the potential to increase both accessibility to screening, and sensitivity to the earliest stages of cognitive decline. A further critical consideration is user satisfaction with such an approach, as up to 35% of older people report an

unwillingness to be screened for memory problems.¹¹ The current study reports on the feasibility of speech-based cognitive screening over the telephone in elderly people, including those both with and without evidence of cognitive decline.

A growing body of research reports on the utility of natural language processing methods to discriminate early cognitive decline from healthy individuals and AD patients using speech data.¹⁸⁻²³ For example, König and colleagues²⁴ recorded the speech of cognitively impaired elderly people using a mobile app, and reported that their automated scoring methods performed as well as humans in discriminating AD from controls Area under the ROC curve (AUC=.94) and mild cognitive impairment from controls (AUC=.76). However, in almost all published reports, these speech data have been audio recorded in clinical or laboratory settings, rather than in the home or other uncontrolled natural environments. Clinical and laboratory settings permit direct control over the quality of the speech data obtained by minimizing noise and distraction, but at the cost of ecological validity.²⁵ If the aim is to assess the speech of older people in the community, many of whom are unable or unwilling to present to a clinician or medical center, then findings from highly controlled settings may not adequately represent this group. Hence, it is unknown the extent to which everyday distractions and ambient background noise might affect the quality of speech recorded over the telephone.

With these important considerations in mind, the current study aims to address the question of whether real-world speech data, recorded over various types of telephones from users in their home environments, can provide the necessary level of accuracy for natural language processing methods to be applied. Specifically, the current study addresses six key questions regarding the practical viability of recording the speech of elderly telephone users, a proportion of whom are cognitively impaired:

1. Will elderly people agree to participate in a recorded telephone interview?
2. Will they complete the interview?
3. Will they find it an acceptable approach?
4. Will the speech that is collected over the telephone be of a good quality?
5. Will the speech data be intelligible to human raters?
6. Will transcriptions produced by automated speech recognition accurately reflect the speech produced?

Method

Participants

Participants (N = 91) were interviewed as part of a study examining the utility of automated language

analysis techniques to diagnose cognitive impairment (AD and MCI versus healthy aging) from speech samples recorded over the telephone (NIA grant number R03AG052416). Participants were community dwelling, native English speakers, aged over 55 years, who were recruited via the Memory Disorders Program at Georgetown University. They were contacted by Memory Disorders Program nurse practitioners from a database of research volunteers who had either completed previous studies through the program or were interested in participating in research. Roughly one third of the sample carried a diagnosis of mild AD, one third carried a diagnosis of amnesic MCI, and one third were cognitively healthy. The current study reports on the acceptability of the approach for older people in general and the quality of the speech data produced, so examination of performance as a function diagnostic group is not discussed.

All participants had adequate hearing, and no self-reported history of neurological disease (e.g., Parkinson's disease, or epilepsy), drug or alcohol abuse, psychiatric hospitalization, current cancer treatment, or stroke or heart attack within the last year, as determined by Memory Disorders Program clinicians. Individuals with minor physical ailments (e.g., diabetes with no serious complications, essential hypertension) were included. Reasons for exclusion from the study were moderate dementia (MMSE <20), poor hearing, significant speech impediment, less than a high-school education, or that they did not speak English, as these would make administration and/or interpretation of the screening tests difficult. Demographic characteristics of the sample are presented in Table 1. Participants received a \$25 gift card as compensation for their time. All participants voluntarily provided written informed consent, and all human subject involvement was approved by the institutional review boards of Marymount University and Georgetown University (MU IRB#260).

Materials

Telephone screener. A standardized cognitive screener, a modified version of the Telephone Interview for Cognitive Status (TICS⁹), was used in the study. In brief, the modified TICS is a short, 13-item test of cognitive functioning administered over the telephone, with scores ranging from 0 to 50. Questions of orientation, repetition, naming, and calculations are some of the items included in the measure. Additionally, a 10-item non-semantically related word list is recalled both immediately and after a delay of about 5 minutes filled with distractor questions, to assess verbal memory.

Table 1. Demographic characteristics of the sample ($N=91$).

Variables	
Age (years)	
Mean ± SD	73.67 ± 6.94
Min-Max	57-93
Gender (%)	
Male	44
Female	56
Education (years)	
Mean ± SD	17.35 ± 2.01
Min-Max	12-20
Ethnicity (%)	
Caucasian	90
African American	9
Asian	1
TICS score	
Mean ± SD	35.26 ± 6.36
Min-Max	18-49
MMSE score	
Mean ± SD	27.26 ± 3.04
Min-Max	19-30

MMSE: Mini Mental State Examination; TICS: Telephone Interview for Cognitive Status.

Speech samples. Participants provided two different speech samples (a short response, based on a verbal fluency prompt, and a longer “free speech” narrative) from which natural language processing metrics will be calculated in ongoing research. The brief samples of speech included standardized verbal fluency tasks assessing semantic (animals and supermarket items) and phonemic (“F, A, S”) word fluency, which were designed to assess recall from semantic memory and from the mental lexicon. These tasks will not be discussed further here. A longer narrative of free speech (describing a favorite memory from childhood) was also included. Finally, participants were asked to rate their satisfaction with the telephone interview procedure, using a 10-point Likert scale in response to

questions concerning ease of use, interest level, and anxiety produced.

Procedures

Participants were called at home on the telephone (mobile phone, landline and/or on speaker) at an agreed upon time by a trained research assistant, who conducted the interview. The telephone interview consisted of two parts (in counterbalanced order), and lasted in total approximately 20-30 minutes. In one part, the modified TICS was administered, and in the second part, participants provided two different speech samples (short responses, based on verbal fluency prompts, and a longer free speech narrative, described above). The speech samples derived from this portion of the interview were digitally recorded for subsequent analysis.

After completion of the above, participants were asked to rate their satisfaction with the telephone interview procedure. Specifically, participants were asked to rate on a scale of 1-10, how enjoyable it was to complete various sections of the interview, with 1 being not at all enjoyable and 10 being very enjoyable. Similarly, they were asked to rate the difficulty of the questions, with 1 being very difficult and 10 being very easy, and to rate how anxious those questions made them feel, with 1 very anxious and 10 not at all anxious. They completed these three ratings at three different points throughout the interview: following the TICS; the verbal fluency prompts; and the free speech portion.

To improve the validity of the telephone testing, the research assistant confirmed adequate hearing of the participants over the telephone before initiating the telephone interview, and participants' spouses/companions were asked to remove any visual memory aids (calendars, newspapers, paper and pens) and to turn off highly audible distractors (e.g., radio, television). However, ambient background noise was present.

Participants were called at home via the Cisco Jabber interface on a laptop computer, and the verbal fluency and free speech portions of the interview were recorded. The system of calling and recording through a laptop computer was adopted to promote equity and inclusivity as a low-cost solution to the typically expensive recording systems used in controlled, laboratory studies. All interviews were transcribed by the first author or a trained research assistant (intraclass correlation coefficient (ICC)=0.988). This allowed for screening of any potentially personally identifying information to ensure participant privacy prior to the speech samples being subjected to automated speech recognition.

Automated speech recognition of the audio files was undertaken to compare with the accuracy of the human transcriptions. Of note, the study was not originally designed with the intention of using automated speech recognition to transcribe the speech samples. However, given the proliferation of such systems and their low cost, we decided to examine how well a widely used automated speech recognition system would perform with speech data recorded in non-ideal conditions. The metric commonly used to compute automated speech recognition accuracy is the word error rate. In this study, the word error rate was defined as the minimum edit distance between the human transcript and the automated speech recognition transcript. Specifically, it is the summation of the number of new words added, the number of words changed, and the number of words deleted from the human transcript divided by the total number of words in the human transcript. We used the Google Speech API,²⁶ which is based on a deep learning model trained on general English language.

Results

1. Will elderly people agree to participate in a recorded telephone interview?

Ninety-six individuals were referred for the study, and 93 agreed to participate (97%). Three withdrew consent prior to the telephone interview being scheduled.

2. Will they complete the interview?

Two participants were excluded from the study subsequent to the interview, for failing to meet all inclusion criteria. Ninety-one out of the 93 participants completed the entire interview (98% completion rate). One individual did not provide satisfaction ratings (apparently due to confusion about the rating scales), although attempted all other sections of the interview. The other participant provided only a very brief narrative speech sample (despite numerous prompts) and then was unwilling to provide satisfaction ratings for that portion of the interview because they "didn't have any memories."

3. Will they find it an acceptable approach?

Participants rated separate portions of the interview, on a 1–10 scale, in terms of enjoyment, ease and lack of anxiety produced. Descriptive statistics are presented in Table 2. As shown in Figure 1, participants rated the free speech portion of the interview as more enjoyable, easier, and less anxiety provoking, than the other sections (which included a standard cognitive screening test and timed measures of verbal fluency).

Table 2. Mean satisfaction ratings, on a scale of 1–10, as a function of interview section. (Higher scores represent more enjoyable, easier, and less anxiety-provoking.)

	<i>N</i>	Mean	Standard deviation	Maximum
TICS <i>enjoyment</i>	90	6.61	2.56	1-10
TICS <i>easy</i>	90	6.17	2.46	1-10
TICS <i>lack of anxiety</i>	90	7.19	2.57	2-10
Fluency <i>enjoyment</i>	90	6.23	2.48	1-10
Fluency <i>easy</i>	90	6.09	2.48	1-10
Fluency <i>lack of anxiety</i>	90	7.06	2.52	1-10
Free speech <i>enjoyment</i>	89	8.63	1.44	5-10
Free speech <i>easy</i>	89	9.18	1.51	4-10
Free speech <i>lack of anxiety</i>	89	9.07	1.76	3-10

TICS: Telephone Interview for Cognitive Status.

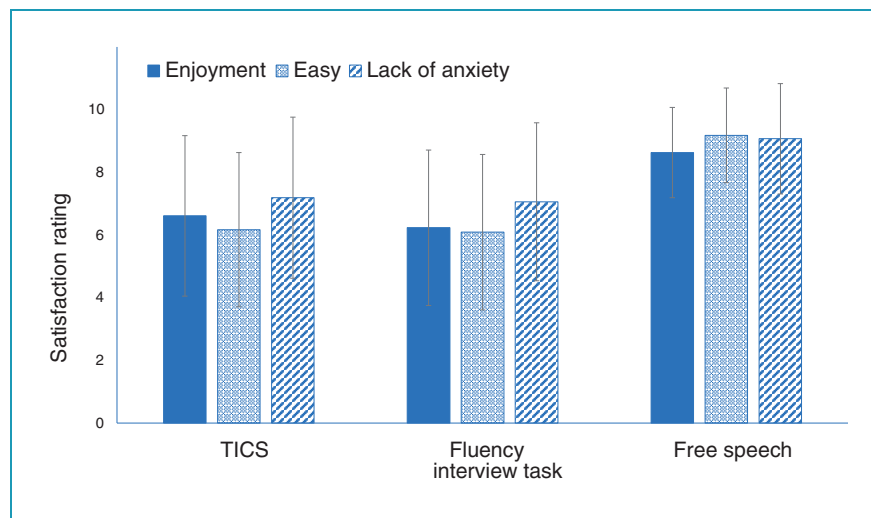


Figure 1. Mean satisfaction ratings on a scale of 1-10 based on enjoyment, ease of questions, and lack of anxiety produced, as a function of interview section.

4. Will the speech that is collected over the telephone be of a good quality?

Table 3 presents the mean number of words produced for the semantic and phonemic verbal fluency tasks and the free speech task. Our operational definition of “good quality” speech data well-suited for the subsequent application of natural language processing methods is somewhat task specific. In our previous work, we have found that on the semantic verbal fluency task where participants are asked to generate as many words as they can that belong to a certain category, that simply one minute of utterances to the

category ‘animals’ is ‘sufficient’ in terms of being informative regarding, for example, thought disorder²⁷ (experiment 2), genetic differences²⁸ and in terms of elucidating how depression affects the speed and flow of thought.²⁹ However, in the case of story recall, as well as tasks that are more open-ended, the amount of speech data required in order to use speech technologies to make incisive comments about underlying neurocognitive processes is naturally different. Indeed, with such tasks we have found that utterances with fewer than 5 words have lacked informational value (e.g., “I don’t know”) and as such, were eliminated in

Table 3. Mean number of words produced, as a function of interview section ($N=91$).

	Mean	Standard deviation	Minimum	Maximum
Fluency ^a - Animals	16.33	7.44	0 ^c	36
Fluency ^a - Supermarket	20.38	8.73	0 ^c	43
Fluency ^a - F	13.05	5.77	2	27
Fluency ^a - A	11.89	5.78	1	30
Fluency ^a - S	13.90	6.24	2	30
Free speech ^b	326.76	166.27	44	1110

^aDoes not include repetitions or intrusions.

^bDoes not include fillers.

^cOne participant spoke, but did not produce any exemplars.

Table 4. Mean word error rate (WER) from automatic speech recognition, as a function of task ($N=91$).

	Mean WER	Standard deviation WER	Minimum WER	Maximum WER
Participant free speech	0.56	0.23	0.12	1.00
Participant free speech, fillers removed	0.54	0.23	0.12	1.00
Interviewer free speech	0.26	0.17	0.00	0.86
Interviewer free speech, fillers removed	0.24	0.23	0.00	0.86

our analyses, so as to allow for the creation of models that could learn based on language produced rather than language that is missing.³⁰ In fact, our previous work showed that the predictive performance of NLP models based on word error rates in the range of 20% lost very few percentage points, presumably because the overall gist, or meaning, of the utterance was retained.

In a series of studies where we have modelled verbal memory recall from speech where participants are read a story (as in the Wechsler Memory Scale – logical memory subtest, or variants of this) and then asked to verbally recall immediately and also after some delay, participants produce on average 57 words (when recalling stories that are between 61 and 82 words). This has been ‘sufficient’ to model memory recall, as well as to model differences in patients with serious mental illness. As shown in Table 2, the average number of words varied considerably by task, but averages all exceeded the minimum number of words found to be useful in our previous work.³⁰ Of note, in our current study the free speech prompt elicited a wide range of response lengths, from 44-1110 words (mean = 326.76 words; standard deviation = 166.27 words), which is a sufficient number of words to be considered

for subsequent natural language processing applications.

5. Will the speech data be intelligible to human raters?

All free speech portions of the audio interview were transcribed by the first author or a trained research assistant. (Inter-rater reliability was assessed using a two-way mixed, intraclass correlation coefficient model to assess the degree of agreement between raters in the transcriptions for the first five cases. The resulting ICC was in the excellent range (ICC = 0.988)). The mean percent of unclear or unintelligible words as a function of total word count was 0.7%, sd = 1.2%, range = 0-6%.

6. Will transcriptions produced by automated speech recognition accurately reflect the speech produced?

Using Google’s acoustic model, the average word error rate of the participant’s speech as compared to the trained human transcriptions was 54%. In contrast, the average word error rate of the interviewer’s speech as compared to the human transcriptions was notably lower, at 24% (see Table 4). While Google’s model is

considered an industry leader, we discuss below alternative models with lower word error rates to use in future research on clinical populations.

Discussion

Results of the current study provide evidence that real-world, unconstrained speech data, recorded over various types of telephones from users in their home environments, can provide the necessary level of quality for natural language processing methods to be applied subsequently. Elderly volunteers, both with and without cognitive impairment, readily agreed to participate in the telephone interview, completed it in its entirety, and rated the overall approach as acceptable. These results can be compared with those of Van Mierlo and colleagues³⁴ who reported an 80% completion rate for their telephone screener, compared with 98% in the current study. Of note, the current study employed an interview style with many opportunities for participants to interact with the interviewer, whereas the Van Mierlo study used a fully automated self-test with all responses coming from touch-tone keys on the telephone. In the current study, participants reported enjoying recalling a favorite memory from childhood, more so than completing formal tests such as the standardized cognitive screening measure (TICS) and timed fluency tasks. This is perhaps unsurprising, given reports from previous telephone-based studies that many older people express an unwillingness to be screened for memory problems, presumably due to their perceived performance on conventional memory tests.¹¹

In terms of the speech data collected, the prompts used for both short response and longer free speech narratives elicited an acceptable quality and quantity of speech for ongoing analyses to be applied.³⁰ Despite varying qualities of recordings (including background noise from pets and environmental sources), almost all recorded words were sufficiently intelligible to be transcribed by humans (less than 1% unintelligible), indicating almost no loss in data due to the recording conditions. In contrast, results from the automated speech recognition analysis revealed that, on average, almost half of the words spoken by the elderly participants were incorrectly transcribed. Considering the human transcript as the base of comparison, the average word error rate of Google's automated speech recognition technology was 54%, compared with a more reasonable 24% word error rate of the interviewer's speech. A number of possible explanations arise when considering this discrepancy. First, the quality of the recordings may be to blame. The interviewer was directly recorded through the laptop interface, whereas the elderly participants' voices were recorded through

the interface after going through a telephone first. This extra step may have degraded the audio signal to the extent that automated speech recognition had difficulty distinguishing the individual words used.

A second possible explanation lies in the quality of the speech itself. Characteristics of elderly speech, such as variable speech rate, hesitations, repetitions, poorer articulation and pronunciation,^{35,36} are all known to inflate the word error rate of automated speech recognition systems.³⁷ A final potential contributor to the discrepancy in word error rate between the participants and the interviewer concerns the content of the speech recorded. The interviewer was either reading a script or asking direct questions, whereas the participant was being asked to recall memories without preparation, potentially leading to hesitation and uncertainty.

We are not the first to report high word error rate in the elderly with automated speech recognition technology. Using a generic automated speech recognition system, Aman and colleagues³⁸ reported an average increase in word error rate between elderly and non-elderly of over 34%. By training on elderly voices specifically, they were able to adapt the automated speech recognition system and reduce the elderly word error rate down to the previous non-elderly word error rate (from 43.5% to 14.5%). Within the elderly group, though, the word error rate performance did not correlate with age, but rather with dependence due to physical degradation. This suggests that generic automated speech recognition systems may not perform well with elderly voices, yet with training on elderly speech, can improve markedly.

Google's automated speech recognition was initially chosen as it is an industry leader in such services, especially in research applications. However, we now realize that our data are not well-suited to their acoustic and language models, so for future research we will explore other automated speech recognition services. On one sample of speech, for example, Google's system generated a transcript with an 83% word error rate. This was a particularly poor audio sample with human transcriptions even noting a couple of "inaudible" words. In an exploration of better-suited automated speech recognition systems, we found word error rates on this particular sample to be still quite high on another industry leader, IBM Watson (WER = 68%³⁹), but much improved for two other services, Sonix (WER = 47%⁴⁰) and Otter.ai (WER = 37%⁴¹). Thus in future work, these improved models will be utilized for automated transcription.

We note that, although high, the word error rate of the elderly participants in the current study is consistent with other studies using automated speech recognition on elderly speech,^{42,43} even in controlled laboratory settings.⁴⁴ There are reports that pre-

processing of elderly speech can decrease the word error rate by up to 12%,⁴⁵ although evidence suggests that natural language processing models are relatively impervious to high word error rate.³⁰ The robust performance of natural language processing models can be attributed to different normalizations of words between a human transcript and an automated speech recognition transcript, trivial word errors that would not change the meaning of a sentence (e.g., “one” vs. “1”), and that natural language processing models are generally trained with a diverse set of language features and are thus able to retain different facets of the language even in the context of a large word error rate. This implies that speech recorded over the telephone can be automatically transcribed with sufficient accuracy for classification purposes.

While errors in transcribing the words themselves may lead to small reductions in performance when applying natural language processing techniques in the future, the speech data holds additional value in automation. Speech samples have been analyzed with acoustic feature extraction software (e.g., openSMILE⁴⁶) and studies have shown that these features can have high predictive potential in patients with serious mental illness.^{47,48} Furthermore, word timing features can be automatically extracted to analyze speed and rhythm in speech. In sum, while fairly high word error rates may exist in automated transcriptions, other modalities of features which are not affected by these errors can still be accurately extracted and used in automated systems.

Results of the current study address the acceptability of using speech recorded over the telephone, in real-world situations, as the basis to develop sensitive metrics that may be used as potential biomarkers by leveraging natural language processing methods. The telephone is ubiquitous and non-threatening to older people, and can likely address a great number of barriers to in-person clinical evaluation. While recent technological breakthroughs show great promise as screeners for cognitive decline and dementia (e.g., eye scanning,⁴⁹ plasma tau⁵⁰), they nonetheless still require costly equipment and/or face-to-face attendance. A semi-automated approach to screening for dementia using speech recorded over the telephone holds the promise of not only allowing more frequent and less costly evaluations, but also for advanced natural language processing models to be applied to the speech in order to find subtle patterns that are not as detectable to a human. For example, in recent work, Orimaye and colleagues²¹ used a combination of deep neural network and deep language models to predict MCI and AD-type dementia from speech data, with impressive accuracy (MCI AUC = 0.80; AD AUC = 0.83). Models such as these are able to efficiently learn patterns from large,

labeled datasets and can be applied to new speech samples to determine relevant classifications or scores on cognitive tasks. While this study made use of state of the art machine learning techniques and had impressive results, it must be noted that the data used to train and test their system is highly controlled and unlikely to match data taken from real-world settings. This means that it would fail to generalize to new data retrieved in a different and less controlled manner or setting, a factor that must be taken into account for all systems that strive to have translational value.

Modern machine learning techniques and analytics of such measures hold the potential to make great advancements in the assessment, diagnosis, and monitoring of mental illness and cognitive decline. With the availability and prevalence of mobile phone applications and wearable devices, data collection and storage is easier and more widespread than it has ever been. Large datasets can easily be retrieved from both target groups and healthy individuals to build powerful models that can assist clinicians in a variety of ways. These powerful machine learning models have the potential to catch subtle deviations in behavior, and when combined with remote data collection, can catch far more critical events from patients than ever before. While the limitations of applying these techniques to the elderly population still apply (e.g., poor speech articulation and limited data availability on the population), there is great potential in these methodologies to advance our understanding of aging and cognitive decline.

Acknowledgements: The authors would like to thank the study participants for their time and Kelly Ha for her support as a research assistant.

Contributorship: CDA and BE researched literature, conceived the study and developed the protocol. RST and BR were involved in gaining ethical approval and patient recruitment. CC completed the data analysis. CDA wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.


Declaration of conflicting interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: The ethics committees of Marymount University & Georgetown University approved this study (MU IRB#260).

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Institutes on Aging [grant number R03AG052416].

Guarantor: CDA.

Peer review: Nicole Lowres, Heart Research Institute Ltd, The University of Sydney has reviewed this manuscript.

ORCID iDs: Catherine Diaz-Asper  <https://orcid.org/0000-0003-2323-7169>

R Scott Turner  <https://orcid.org/0000-0001-7534-2935>

References

1. Alzheimer's Association. Alzheimer's disease facts and figures. *Alzheimer's Dement* 2019; 15: 321–387.
2. Arcury T, Preisser J, Gesler W, et al. Access to transportation and health care utilization in a rural region. *J Rural Health* 2005; 21: 31–38.
3. Averill B. Priorities for action in a rural older adults study. *Fam Community Health* 2012; 35: 358–372.
4. Kotagal V, Langa K, Plassman B, et al. Factors associated with cognitive evaluations in the United States. *Neurology* 2015; 84: 64–71.
5. Gallo J and Breitner J. Alzheimer's disease in the NAS-NRC registry of ageing twin veterans: IV. Performance characteristics of a two-stage telephone screening procedure for Alzheimer's dementia. *Psychol Med* 1995; 25: 1211–1219.
6. Martin-Khan M, Wootton R and Gray L. A systematic review of the reliability of screening for cognitive impairment in older adults by use of standardised assessment tools administered via the telephone. *J Telemed Telecare* 2010; 16: 422–428.
7. Vaskinn A, Wilsgård I, Holm A, et al. A feasibility study of a telephone-based screening service for mild cognitive impairment and its uptake by elderly people. *J Telemed Telecare* 2013; 19: 5–10.
8. Barth J, Nickel F and Kolominsky-Rabas P. Diagnosis of cognitive decline and dementia in rural areas – a scoping review. *Int J Geriatr Psychiatry* 2018; 33: 459–474.
9. Brandt J, Spencer M and Folstein M. The telephone interview for cognitive status. *Cogn Behav Neurol* 1988; 1: 111–117.
10. Folstein M, Folstein S and McHugh P. “Mini-mental state.” a practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* 1975; 12: 189–198.
11. Herr M and Ankri J. A critical review of the use of telephone tests to identify cognitive impairment in epidemiology and clinical research. *J Telemed Telecare* 2013; 19: 45–54.
12. Lin JS, O'Connor E, Rossom RC, Perdue LA, Burda BU, Thompson M, Eckstrom E. Screening for Cognitive Impairment in Older Adults: An Evidence Update for the U.S. Preventive Services Task Force [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2013 Nov. Report No.: 14-05198-EF-1. PMID: 24354019.
13. Verma M and Howard RJ. Semantic memory and language dysfunction in early Alzheimer's disease: a review. *Int J Geriatr Psychiatry* 2012; 27: 1209–1217.
14. Garrard P, Maloney L, Hodges J, et al. The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain* 2005; 128: 250–260.
15. Pakhomov S, Chacon D, Wicklund M, et al. Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing. *Behav Res Methods* 2011; 43: 136–144.
16. Berisha V, Wang S, LaCross A, et al. Tracking discourse complexity preceding Alzheimer's disease diagnosis: a case study comparing the press conferences of presidents Ronald Reagan and George Herbert Walker Bush. *J Alzheimers Dis* 2015; 45: 959–963.
17. Fang C, Janwattanapong P, Martin H, et al. Computerized neuropsychological assessment in mild cognitive impairment based on natural language processing-oriented feature extraction, 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 2017, pp. 543–546, doi: 10.1109/BIBM.2017.8217706.
18. Alhanai R. Au and Glass J. “Spoken language biomarkers for detecting cognitive impairment,” 2017 *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, Japan, 2017, pp. 409–416, doi: 10.1109/ASRU.2017.8268965.
19. Jarrold W, Peintner B, Wilkins D, et al. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. *Workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, 2014, pp.27–37.
20. König A, Satt A, Sorin A, et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement* 2015; 1: 112–124.
21. Orimaye S, Wong J and Wong C. Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia. *PLoS ONE* 2018; 13: e0205636.
22. Themistocleous C, Eckerström M and Kokkinakis D. Identification of mild cognitive impairment from speech in Swedish using deep sequential neural networks. *Front Neurol* 2018; 9: 975.
23. Tóth L, Hoffman I, Gosztolya G, et al. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Curr Alzheimer Res* 2018; 15: 130–138.
24. König A, Satt A, Sorin A, et al. Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. *CAR* 2018; 15: 120–129.
25. Kvavilashvili L and Ellis J. Ecological validity and the real-life/laboratory controversy in memory research: a critical and historical review. *History Philos Psychol* 2004; 6: 59–80.
26. Google API, <https://cloud.google.com/speech-to-text/> (accessed October 2020).
27. Elvevåg B, Foltz PW, Weinberger DR, et al. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophr Res* 2007; 93: 304–316.
28. Nicodemus KK, Elvevåg B, Foltz PW, et al. Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach. *Cortex* 2014; 55: 182–191.

29. Holmlund TB, Cheng J, Foltz PW, et al. Updating verbal fluency analysis for the 21st century: applications for psychiatry. *Psychiatry Res* 2019; 273: 767–769.
30. Chandler C, Foltz PW, Cheng J, Bernstein JC, Rosenfeld EP, Cohen AS, Holmlund TB and Elvevåg B (2019) Overcoming the bottleneck in traditional assessments of verbal memory: Modeling human ratings and classifying clinical group membership. In Niederhoffer, K., Hollingshead, K., Resnik, P., Resnik, R., Loveys, K. (Eds), *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. Minneapolis, Minnesota, USA, June (pp. 137–147). <https://doi.org/10.18653/v1/W19-3016>
31. Rosenstein M, Diaz-Asper C, Foltz PW, et al. A computational language approach to modeling prose recall in schizophrenia. *Cortex* 2014; 55: 148–166.
32. Chandler C, Foltz PW, Cohen AS, et al. Machine learning for longitudinal applications of neuropsychological testing. *Intell Based Med* 2020; 1–2: 100006.
33. Holmlund TB, Chandler C, Foltz PW, et al. Applying speech technologies to assess verbal memory. *NPJ Digit Med* 2020; 3: 33.
34. Van Mierlo L, Wouters H, Sikkes S, et al. Screening for mild cognitive impairment and dementia with automated, anonymous online and telephone cognitive self-tests. *J Alzheimer's Dis: redundant* 2016; 56: 1–11.
35. Horton W, Spieler D and Shriberg E. A corpus analysis of patterns of age-related change in conversational speech. *Psychol Aging* 2010; 25: 708–713.
36. Young V and Mihailidis A. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: a literature review. *Assist Technol* 2010; 22: 99–112.
37. Errattahi R, El Hannani A and Ouahmane H. Automatic speech recognition errors detection and correction: a review. *Procedia Comput Sci* 2018; 128: 32–37.
38. Aman F, Vacher M, Rossato S, et al. Analyzing the performance of automatic speech recognition for ageing voice: does it correlate with dependency level? *SLPAT, 4th workshop on speech and language processing for assistive technologies*, 2013, pp.9–15.
39. IBM Watson, <https://www.ibm.com/watson/speech-to-text> (accessed October 2020).
40. Sonix, <https://sonix.ai> (accessed October 2020).
41. Otter Voice Notes, <https://otter.ai> (accessed October 2020).
42. Vacher M, Aman F, Rossato S, Portet F (2015) Development of Automatic Speech Recognition Techniques for Elderly Home Support: Applications and Challenges. In: Zhou J., Salvendy G. (eds) *Human Aspects of IT for the Aged Population. Design for Everyday Life*. ITAP 2015. Lecture Notes in Computer Science, vol 9194. Springer, Cham. https://doi.org/10.1007/978-3-319-20913-5_32
43. Vipperla R, Frankel J and Renals S. Longitudinal study of ASR performance on ageing voices. In: *Proc. interspeech*, 2008. September 22–26, Brisbane Australia.
44. Lehr M, Prud'hommeaux E, Shafran I, et al. Fully automated neuropsychological assessment for detecting mild cognitive impairment. *Interspeech* 2012; 2012: 1039–1042.
45. Kwon S, Kim S-J and Choeh J. Preprocessing for elderly speech recognition of smart devices. *Comput Speech Lang* 2016; 36: 110–121.
46. Schuller B, Steidl S, Batliner A, et al. The *INTERSPEECH* 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25–29, 2013, In: Bimbot F, Cerisara C, Fougheron C, Gravier G, Lamel L, Pellegrino F and Perrier P (ed.), ISSN 2308-457X; ISCA Archive, http://www.isca-speech.org/archive/interspeech_2013.
47. Cohen AS, Fedechko TL, Schwartz EK, et al. Ambulatory vocal acoustics, temporal dynamics, and serious mental illness. *J Abnorm Psychol* 2019; 128: 97–105.
48. Cohen AS, Cox CR, Le TP, et al. Using machine learning of computerized vocal expression to measure blunted vocal affect and alogia. *NPJ Schizophr* 2020; 6: 26.
49. Yoon SP, Grewal D, Thompson A, et al. Retinal microvascular and neurodegenerative changes in Alzheimer's disease and mild cognitive impairment compared with control participants. *Ophthalmol Retin* 2019; 3: 489–499.
50. Pase MP, Beiser AS, Himali JJ, et al. Assessment of plasma total tau level as a predictive biomarker for dementia and related endophenotypes. *JAMA Neurol* 2019; 76: 598–606.