Telephone-based Dementia Screening I: Automated Semantic Verbal Fluency Assessment

Johannes Tröger  
German Research Center for Artificial Intelligence (DFKI)  
Saarbrücken, Germany  
johannes.troeger@dfki.de

Nicklas Linz  
German Research Center for Artificial Intelligence (DFKI)  
Saarbrücken, Germany  
nicklas.linz@dfki.de

Alexandra König  
Memory Center, CoBTeK - IA CHU  
Université Côte d’Azur  
Nice, France

Philippe Robert  
Memory Center, CoBTeK - IA CHU  
Université Côte d’Azur  
Nice, France

Jan Alexandersson  
German Research Center for Artificial Intelligence (DFKI)  
Saarbrücken, Germany

ABSTRACT

Dementia has a large economic impact on our society as cost-effective population-wide screening for early signs of dementia is still an unsolved medical supply resource problem. A solution should be fast, require a minimum of external material, and automatically indicate potential persons at risk of cognitive decline. Despite encouraging results, leveraging pervasive sensing technologies for automatic dementia screening, there are still two main issues: significant hardware costs or installation efforts and the challenge of effective pattern recognition. Conversely, automatic speech recognition (ASR) and speech analysis have reached sufficient maturity and allow for low-tech remote telephone-based screening scenarios. Therefore, we examine the technologic feasibility of automatically assessing a neuropsychological test—Semantic Verbal Fluency (SVF)—via a telephone-based solution. We investigate its suitability for inclusion into an automated dementia frontline screening and global risk assessment, based on concise telephone-sampled speech, ASR and machine learning classification. Results are encouraging showing an area under the curve (AUC) of 0.85. We observe a relatively low word error rate of 33% despite phone-quality speech samples and a mean age of 77 years of the participants. The automated classification pipeline performs equally well compared to the classifier trained on manual transcriptions of the same speech data. Our results indicate SVF as a prime candidate for inclusion into an automated telephone-screening system.

KEYWORDS

Dementia, Screening, Speech Analysis, Phone-based, Machine Learning

1 INTRODUCTION

Dementia has a large economic impact on our society: according to the World Alzheimer Report 2016, dementia is about to become a trillion dollar disease by 2018 [47]. Since many clinical trials have failed to find a cure, a conceptual shift has occurred considering Alzheimer’s disease (AD) as a continuum for which early intervention may offer the best chance of therapeutic success [11]. This urgent need to identify a treatment that can delay or prevent AD has increased the number of preventative trials targeting disease modifying risk factors for which early screening of subjects at risk to develop cognitive impairment is highly relevant [1]. Recent research has shown that prevention at prodromal stages targeting disease mechanisms show promising results and are more likely to be effective [52]. Many challenges remain detecting these ‘silent’ stages, where clinical signs are not yet very obvious since our understanding of the pathological mechanism is still quite limited [13] and current tools may lack sufficient sensitivity to detect subtle but meaningful changes.

This approach has led to the current discussion on creating and approving more clinically relevant measures for early population based screening with low-cost tests of high sensitivity and lower specificity [11]. For instance, currently, just 50% of cases are diagnosed in Europe and the US [47]. This can be attributed to effective screenings for early signs of dementia (mild neurocognitive disorder) having not reached sufficient coverage. Especially in areas with low population density, clinical facilities and experts are too distributed to screen populations effectively, as this is still done in a face-to-face manner today. Many clinical trials suffer from high drop out rates partly due to visit frequency and study length [21]. This translates into a medical supply resource problem and highlights the opportunities for telemedicine applications.

It has been put forward that new tools may address this need fast, require neither laboratory setup nor external material, and
automatically evaluate and indicate potential clinically relevant persons. Therefore, research should focus on innovative computerized tools that reveal robust psychometric properties for early detection of neurocognitive disorder significantly decreasing the workload of expert clinicians, which represent a very rare resource in most cases. Thus, automatic, inexpensive and remote solutions allowing a broad frontline screening of cognitive abilities in the general population should be developed.

There is growing evidence for the feasibility of automatic speech analysis in addressing exactly this need [25, 33, 56]. Speech-based solutions can be remotely administered via telephone and therefore have minimal technical user interface requirements. This makes them a very attractive solution in the mentioned frontline screening context.

Neuropsychological studies comparing a video and telephone based psychometric dementia screening with a face-to-face assessment, reported good ecological validity for the telemedicine application [39]. However, such studies do not fully exploit the combined opportunities of telemedicine neuropsychological screening empowered by automatic speech analysis and machine learning classification.

Our aim is to develop technology with which raw speech data can be processed via the telephone—facilitated by computational linguistic techniques and machine learning—in order to give a simple risk assessment for dementia. Instead of using free, unconstrained speech, we hope to achieve better performance and shorter assessment times, through analysing performances of cognitive tests. Semantic Verbal Fluency (SVF) tasks are neuropsychological tests in which patients are given limited time (e.g. 60 seconds) to name as many items belonging to a certain semantic category as they can. SVF has been shown to be sensitive to even early forms of dementia [3, 19, 43, 48]. SVF can be considered a multifactorial task, comprising both semantic memory retrieval and executive control processes [24, 49, 58]. Previous studies have concluded the feasibility of automatically analysing SVF performances [31, 44], although no study known to the authors has investigated analysis of telephone quality recordings.

The aim of this study is therefore to benchmark a solution processing raw telephone quality SVF data suitable for inclusion in a fully automated dementia frontline screening for global risk assessment.

2 RELATED WORK

The following section gives an overview of efforts aiming at the automated detection of dementia based on multiple different sensor solutions. For this paper, we would like to differentiate between solutions based on classic pervasive sensing such as home monitoring systems and speech analysis as a special subcategory of pervasive sensing.

2.1 Computerized cognitive screening

Digital tests that seek to assess cognitive functions, briefly and globally, are being developed with the aim to be administered remotely [5]. The exhibited advantages of these tests are standardization of administration and stimulus presentation as well as the measures (e.g. reaction times and latencies) are more accurate: performances can be compared to established norms [59] allowing the clinician to concentrate on a personalized analysis of the patients’ needs.

For instance, the CogState Brief Battery (CogState) is a brief computerized test which assesses reaction and processing speed, episodic memory, attention, working memory, learning, and decision-making. [9] examined the specificity and sensitivity of the CogState test for the diagnosis of mild cognitive deterioration comparing it with classical pen and paper tests with the result that it reaches similar discrimination level as traditional tests.

CANTAB, one of most known cognitive screening tools, offers specialized AD test battery versions for assessing prodromal states, or mild dementia. The batteries measure motor skills, executive function, episodic memory, visual memory information processing and sustained attention. CANTAB has been shown to be highly sensitive to cognitive dysfunction and ties in closely with current neurobiological models for MCI [12, 16].

The TDAS (Touch Panel-type Dementia Assessment Scale) [27] based originally on the pen and paper ADAS-cog test [50], measures word recognition, instruction compliance, temporal orientation, visuospatial skills, recognition of object use, naming, planning of the writing process, money computation, and recognition of the time indicated by an analogue clock. This digital test can be administered in 30 minutes, just two-thirds of the time that ADAS-cog requires.

The CNSVS (CNS Vital Signs) [23] is a digital screening test, assessing working memory, mental flexibility, psychomotor speed, verbal and visual memory, set shifting and inhibition and vigilance and sustained attention. The authors studied test-retest reliability as well as concurrent and discriminant validity concluding that it can be used as a reliable screening tool in medical contexts.

Phone-based screening has been investigated by Castanho et al. (2016) comparing the delayed recall task and a classical neuropsychological assessment with the Telephone Interview of Cognitive Status (TICS) in a population of older adults. The TICS consists of 13 items evaluating spatial, temporal and personal orientation, working memory, attention, and verbal and semantic memory. TICS showed high correlation levels with global scores of classical tests as well as a satisfactory internal consistency. This method could allow faster access to assessment for people living in rural areas...
producing similar results as the usual pencil and paper screening tests.

2.2 Automated Screening Based on Pervasive Sensing

Manifold research has been done into the feasibility of home monitoring systems for modelling domestic circadian activities (activity patterns following a biological 24h rhythm). As such rhythms are typically disturbed by dementia—especially nocturnal activity patterns—these techniques provide a useful basis for automatic dementia detection/screening. Using infrared sensors to monitor nocturnal activities, studies have found significant differences between dementia patients and healthy controls (e.g. [53]). Similarly, the same technical setup has been shown to effectively model daily routines [17]. Following the same rationale and technique [30] leveraged automatic detection of instrumental activities of daily living (IADL) in patients with MCI and healthy participants. Besides promising results, such studies are often carried out with very small sample sizes (N < 50) and focus mainly on the automatic classification of activities rather than the actual neurocognitive disorder. Moreover, the installation of home-monitoring systems require significant resources and a person’s consent to be monitored in their private life; two issues that render such a solution unrealistic in broad population frontline screening.

Also focusing on circadian rhythm monitoring but using less complex wrist-worn technology, [42] found significant correlations between sleep patterns and common dementia staging scales. However, similar to the above-mentioned studies, sample size is relatively small and the main automatic analysis effort was spent on activity monitoring rather than diagnostic classification problems.

Beyond such passive sensing approaches, there is also research on the diagnostic use of pro-active sensing situations: situations that are framed by some task/instruction producing more diagnosis related variance. Leveraging virtual reality technology, [54] used a realistic virtual reality (VR) fire evacuation task to predict amnestic Mild Cognitive Impairment (MCI; often considered as the precursor of dementia), Alzheimer’s disease (AD) and controls from task performance reaching area under the curve (AUC) values of more than 80%. Though very sensitive, the classification setup requires a lot intervention from technicians to analyse the VR task performance. Moreover, the VR screening setup has similar limitations as the classic neurological assessment: it requires the expensive VR application which is phone-based and acts as a passive listener to monitor speech over time. However, as soon as an anomaly is detected the app also uses classic cognitive speech tasks to elicit richer and more controlled variance (i.e. a psychomotor task: continuously repeating pa-ta-ka during a given period of time)

Multiple studies report approaches that are less feasible in phone-based screening scenarios but provide strong evidence for the effectiveness of speech-based screening for dementia patients, including early stages. Overall, reported work either uses speech from conversations, spontaneous speech tasks, reading or repetition tasks, and fluency tasks.

The most liberal setting consists of conversations with clinicians. Audio files of spontaneous speech from conversations [10, 28], or classical autobiographic patient interviews [25] have been used in small setups, yielding significant effects. For such data, considerable effort has to be spent on preprocessing the data (e.g. annotating turns or trimming the audio file) in order to prepare it for further computational learning.

Tasks, eliciting spontaneous speech, are slightly more restricted and therefore easier to process; descriptions of the Cookie Theft Picture or comparable visual material, allows for extracting a wide variety of features and yields very good results [2, 18, 32, 41]. Similarly, some researchers report positive results from speech samples based on an animated film free recall task [20].

Reading or repetition tasks are the most handy to deal with, in the sense of automated processing, as they need little transcription and provide an inherent ground truth. Simple sentence reading has been shown to provide enough variance to effectively discriminate between AD and HC with an accuracy of 84% [38].

Verbal fluency tasks, such as the semantic animal fluency task, have produced rich variance to discriminate between AD patients and HC [32, 34, 61]. The benefits of semantic vs. phonemic fluency tasks have been discussed in multiple publications and there is a large body of neuropsychological evidence reporting dementia patients’ difficulties in semantic fluency tasks, concluding that
dementia patients and MCI patients have significant more difficulties in semantic, e.g., animal, fluency tasks compared to other psychometric standard tests.

In summary, speech analysis provides a powerful opportunity to broaden dementia screening as it has minimal technical requirements and leverages a mature technology—ASR—and can be done remotely in almost all geographic areas. Sensitivity can even be increased through the use of specific psychometric speech tasks, such as the semantic verbal fluency task. Therefore, our aim is to benchmark an entirely automatic pipeline for dementia screening using telephone-quality audio recordings of a classic dementia screening speech task, ASR and machine learning classifiers on top.

3 METHODS

In order to address the above-mentioned challenges, this section will elaborate on the technical pipeline of the proposed system and provide evidence for its feasibility. In the following, the telephone-based speech data processing and the machine learning experiment will be described.

3.1 Participants

Within the framework of a clinical study carried out for the European research project Dem@care, and the EIT Digital project ELEMENT, speech recordings were conducted at the Memory Clinic located at the Institut Claude Pompidou and the University hospital in Nice, France. The Nice Ethics Committee approved the study. Each participant gave informed consent before the assessment. Speech recordings of elderly people were collected using an automated recording app which was installed on a tablet computer.

Participants underwent a clinical assessment including a battery of recorded speech-based tasks.

Each participant went through an assessment including: Mini-Mental State Examination (MMSE) [15], the phonemic and semantic verbal fluency [55], and the Clinical Dementia Rating Scale [40]. Following the clinical assessment, participants were categorised into three groups: control participants that complained about having subjective cognitive impairment (SMC) but were diagnosed as cognitively healthy after the clinical consultation, patients with MCI and patients that were diagnosed with dementia (D), including AD. For the AD group, the diagnosis was determined using the NINCDS-ADRDA criteria [37]. Related mixed/vascular dementia was diagnosed according to the ICD 10 [60]. For the MCI group, diagnosis was conducted according to Petersen criteria [46]. Participants were excluded if they had any major audition or language problems, history of head trauma, loss of consciousness, psychotic or aberrant motor behaviour.

In summary, speech analysis provides a powerful opportunity to broaden dementia screening as it has minimal technical requirements and leverages a mature technology—ASR—and can be done remotely in almost all geographic areas. Sensitivity can even be increased through the use of specific psychometric speech tasks, such as the semantic verbal fluency task. Therefore, our aim is to benchmark an entirely automatic pipeline for dementia screening using telephone-quality audio recordings of a classic dementia screening speech task, ASR and machine learning classifiers on top.

<table>
<thead>
<tr>
<th></th>
<th>SMC</th>
<th>MCI</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>40</td>
<td>47</td>
<td>79</td>
</tr>
<tr>
<td>Age</td>
<td>72.65 (8.3)</td>
<td>76.59 (7.6)</td>
<td>79.0 (6.1)</td>
</tr>
<tr>
<td>Sex</td>
<td>8M/32F</td>
<td>23M/24F</td>
<td>39M/40F</td>
</tr>
<tr>
<td>Education in years</td>
<td>11.35 (3.7)</td>
<td>10.81 (3.6)</td>
<td>9.47 (4.5)</td>
</tr>
<tr>
<td>MMSE</td>
<td>28.27 (1.6)</td>
<td>26.02 (2.5)</td>
<td>18.81 (4.8)</td>
</tr>
<tr>
<td>CDR-SOB</td>
<td>0.47 (0.7)</td>
<td>1.68 (1.11)</td>
<td>7.5 (3.7)</td>
</tr>
</tbody>
</table>

Table 1: Demographic data and clinical scores by diagnostic group; mean (standard deviation); SMC='Subjective Memory Complaints', MCI='Mild Cognitive Impairment', D= 'Dementia', MMSE='Mini Mental State Examination', CDR-SOB='Clinical Dementia Scale - Sum of Boxes'.

3.2 Speech Data Processing

Speech was recorded through a mobile tablet device using the built-in microphone. The recordings were digitised at 22050 Hz sampling rate and at 16 bits per sample. To simulate telephone conditions, the recordings were downsampled to a 8000 Hz sampling rate, using the Audacity software. Since the tablet device’s microphone is used in mobile phones, no further transformations were applied.

Recordings of patients were analysed manually and automatically. For manual analysis, a group of trained speech pathology students transcribed the SVF performances following the CHAT protocol [36] and aligned the transcriptions with the speech signal using PRAAT [4]. For the automatic transcription, the speech signal was separated into sound and silent parts using a PRAAT script based on signal intensity. The sound segments were then analysed using Google’s Automatic Speech Recognition (ASR) service, which returns several possible transcriptions for each segment together with a confidence score. The list of possible transcriptions was searched for the one with the maximum number of words that were in a predefined list of animals in French. In case of a tie, the transcription with the higher confidence score was chosen.

3.3 Features

We extracted a variety of features from the generated transcripts. All hereunder reported features are either clinically accepted (i.e. word count), have been proven to have diagnostic power based on previous medical research (i.e. clusters and switches) or proved to have diagnostic power based on research in the field of computational linguistics (i.e. semantic metrics). Moreover, all features are firmly based on clinical research and therefore explicable and understandable by medical experts.

3.3.1 Word Count: The count of distinct correct responses (animals), excluding repetitions, is the standard clinical measure for evaluation of SVF. Its diagnostic power for even early stages of cognitive impairment has been shown in countless studies.

3.3.2 Clusters and Switches: Many previous researchers [22, 34, 48, 58] have shown that production in SVF is guided by so called clusters—clusters of words that are produced in rapid succession and often shown to be semantically connected. We determine clusters in multiple ways—taxonomy-based [58] and statistical [34]
Figure 2: Technical pipeline: the automatic frontline screening using machine classification and feature selection of clinically relevant features feeding the machine learning classifier for neurocognitive screening.

3.3.3 Semantic Metrics: Many purely semantic metrics have been suggested for the analysis of SVF, that look at the type of words produced. We include frequency norms [35] estimated from large text corpora and computed as the mean frequency of any produced word and semantic distance [35] approximated using neural word embeddings trained on external text resources. We include the mean semantic distance between any produced word, the overall mean of means of semantic distances inside a temporal cluster and the the mean semantic distance between any temporal cluster.

3.4 Classification Experiment
In order to evaluate the feasibility of using SVF in a telephone screening scenario, we performed a machine learning experiment. We built classifiers that discriminate the healthy population from the impaired samples. People were counted into the impaired population, when they belonged to either the MCI or dementia groups. First we established a performance baseline, training models based on features extracted from manual transcripts. After that we used the transcripts from ASR to extract features and constructed models.

In all scenarios we used Support Vector Machines (SVMs)[8] implemented in the scikit-learn framework [45]. Due to our limited amount of data—166 samples—we could not keep a separate hold-out set for testing and instead used leave-one-out cross validation. For each sample, the data is split into a training-set—all samples but the one—and a test-set—the one held-out sample. The classifier is trained on the test set and evaluated on the held-out training set. To find a well-performing set of hyperparameters for the SVM (i.e., kernel, C, γ), we performed parameter selection using cross-validation on the training set of the inner loop of each cross validation iteration.

3.5 Performance Measures
The performance of ASR systems is usually determined using Word Error Rate (WER) as a metric. WER is a combination of the types of mistakes made by ASR systems in the process of recognition. Mistakes are categorized into substitutions, deletions and intrusions. Let S, D and I be the count of these errors and N the number of tokens in the ground truth. Then

\[
\text{WER} = \frac{(S + D + I)}{N}
\]

Since WER considers all utterances, including off-task speech which is not reflected in any of our features, we used a slightly adapted version. Instead of comparing the ground truth annotation of the recording and the ASR results, we transformed both into a list of animals and calculate the WER for these sequences. We refer to the result as the Verbal Fluency Error Rate (VFER) in further discussion.

As performance measures for prediction of each class in the ML classification experiment, we report the receiver operator curve (ROC), as different tradeoffs between sensitivity and specificity are visible. We also report area under curve (AUC) as an overall performance metric.

4 RESULTS
We first evaluate the VFER on the automatic transcript, which is determined to be 33.4%. Of the errors made by the ASR, 69% are deletions, 22% are substitutions and 9% are intrusions. Substitutions are the least problematic error, since they only skew the word count—the single most predictive feature—in rare cases, where a word is substituted with a previously named one.

Figure 3 shows the receiver operator curve (ROC)—a plot of true positive rate vs. false positive rate—for both classification experiments. Models based on features extracted from manual transcripts have an AUC of 0.852 and models built on features extracted from automatic transcripts show an AUC of 0.855. Since a high sensitivity is key for screening applications, a sensible sensitivity-specificity trade-off for the automatic model could be at a sensitivity of around 0.85 and a specificity of 0.65.
The results of our experiments show, that (1) the fully automated SVF for dementia detection. Substitutions affect the word count less, are both problematic for further analysis, as they skew the raw erosion. Considering the types of errors, insertions and deletions to healthy subjects, which can be explained by age-related speech impairments.

In our experiments, the automated ASR-/phone-based screening pipeline and the pipeline based on manually transcribed speech reach comparable classification results. This is very encouraging, as the transcription of speech is the number-one resource-straining factor, showing that an automatic speech-based system has become a powerful alternative to manual analysis of speech-based psychometric tests.

ASR is often considered to be the main weakness in speech based automatic screening approaches [56]. Our results show an overall error rate of 33.4 % for the automated system, compared to the manual transcripts. This result represents an improvement over results of other authors using ASR systems for evaluating the SVF tasks [33, 44]. In line with previous research, more word errors are produced by the ASR for dementia patients, compared to healthy subjects, which can be explained by age-related speech erosion. Considering the types of errors, insertions and deletions are both problematic for further analysis, as they skew the raw word count, the single most predictive performance indicator in SVF for dementia detection. Substitutions affect the word count less, only in rare cases, where a word is substituted with a previously named one, generating a false repetition.

6 CONCLUSION

In this paper we set out to benchmark a telephone-based analysis of SVF for inclusion into a fully automated dementia frontline screening for global risk assessment. Our results show that SVF is a prime candidate for inclusion into an automated pipeline, providing decent sensitivity and specificity scores. Additionally, we show that the phone-based classification is as effective as the gold-standard manual transcription based classifier displaying an acceptable ASR word error rate despite telephone setup and the aged sample for the experiments.

Further research will be directed into finding additional tests, that offer increased sensitivity and specificity in combination with SVF. The idea of this series is to validate and construct a system, that solely based on the telephone as a technological interface and administrable in less than 10 minutes, perfectly fits the need of broad dementia screening tools. It should also serve epidemiological research studies and inclusion for pharmaceutical trials, which aim at including representative shares of the population by cost-effective screening for persons with early onset neurocognitive impairments.

REFERENCES

Telephone-based Dementia Screening I: Automated Semantic Verbal Fluency Assessment


