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

Johannes Tröger, Nicklas Linz, Alexandra König, Philippe Robert, Jan Alexandersson

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

Pervasive Health, 2018,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Johannes Tröger, Nicklas Linz, Alexandra König, Philippe Robert, and Jan Alexandersson. 2018. Telephone-based Dementia Screening I: Automated Semantic Verbal Fluency Assessment. In *Proceedings of 12th EAI International Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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 preventional 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

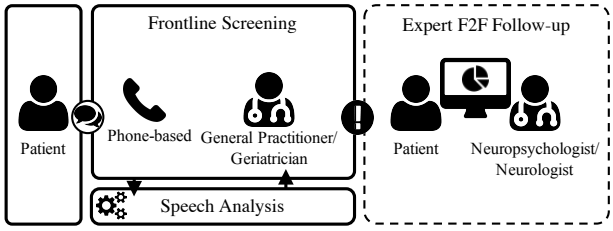

Figure 1: Telephone-based frontline screening scenario: speech gets sent to the analysis server which automatically indicates the general practitioner (GP) or the geriatrician (G) in charge the risk for neurocognitive disorder, the GP/G also checks via phone for excluding/confounding conditions (e.g. substance abuse) and forwards the patient to the specialist who would efficiently continue with the in-depth assessment.

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 prognostic 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 laboratory and test persons have to leave their home.

Other studies combine gait and balance analysis through a hip-/foot-worn accelerometer and specific walking tasks [7, 26]. Such approaches take advantage of classic geriatric assessments showing age-/dementia-related gait irregularities when confronted with a simple straight-line walking task or dual task paradigms (e.g. walking and mental arithmetic task).

These pervasive sensing approaches reveal several *shortcomings* for our use case. They are either very technology-heavy, which implies significant investments, and rely heavily on activity recognition which represents an ongoing classification research challenge in itself. Alternatively, they have to be done in laboratories far away from peoples' homes. Conversely, automatic speech analysis recently has reached a technical readiness level that renders it very attractive for speech based pervasive solutions. Moreover, the only technical requirement is a working telephone which can be considered as ubiquitous in most countries even for an aged population such as the dementia screening target group.

2.3 Automated Screening Based on Speech

Authors have reported studies on automated dementia screening with possible applications in phone-based telemedicine scenarios. [57] extracted paralinguistic features from speech based cognitive tests and trained classifiers to discriminate between healthy controls and patients with AD. Furthermore, [33] used ASR to extract features from a story retelling task and was able to discriminate between MCI and healthy controls with an Area Under the Curve (AUC) score of 80.9%. [51] used four spoken cognitive tests (Countdown, Picture description, Repetition and SVF), extracted paralinguistic features to discriminate individuals with MCI, early AD and healthy controls (HC). Trained models achieve an accuracy of 87% for early AD vs. HC and 81% for MCI vs. HC. Not focusing on dementia detection but on Parkinson's Disease, [29] report an 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

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

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'.

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 broad 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.

Each participant performed the SVF task during a regular consultation with one of the Memory Center's clinician who operated the mobile application which was installed on an iPad tablet. Instructions for the vocal tasks were pre-recorded by one of the psychologist of the center ensuring a standardised instruction over the experiment. Administration and recording were controlled by the application and facilitated the assessment procedure.

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[1] 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[2], 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]

[1] http://www.audacityteam.org/
[2] https://cloud.google.com/speech/

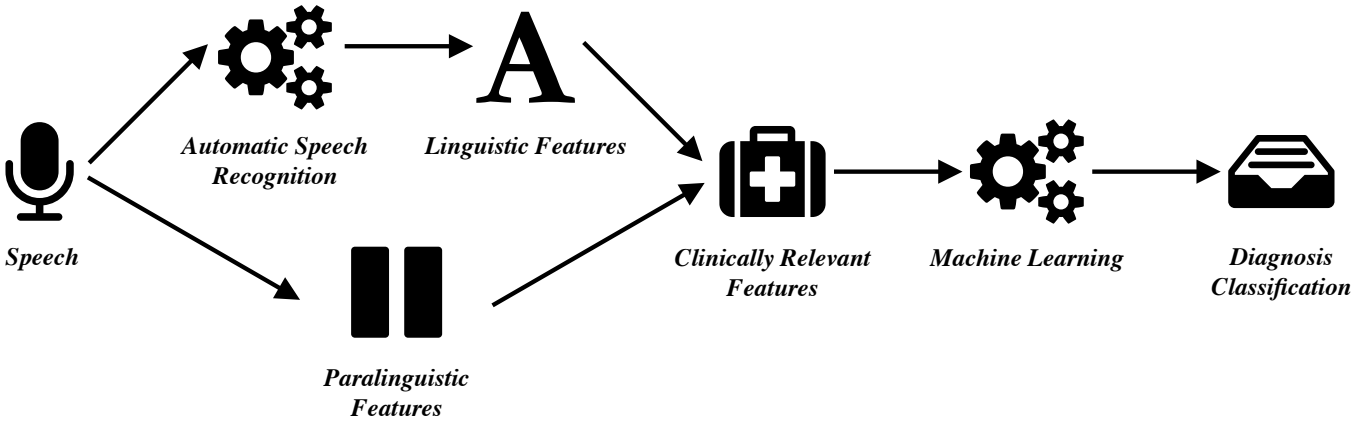

Figure 2: Technical pipeline: the automatic frontline screening using machine classification and feature selection of clinically relevant features feeding the machine learning classifier for neurocognitive screening.

semantic, as well as temporal analysis [14]—and compute mean cluster size and number of switches between clusters as features.

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

$$WER = \frac{(S + D + I)}{N}$$

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.

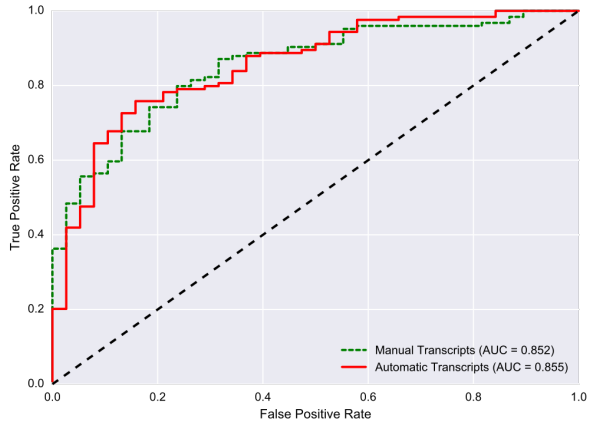

Figure 3: Receiver Operator Curve (ROC) for features based on manual transcripts (green) and on automatic transcripts (red). Area under curve (AUC) is reported in the legend.

5 DISCUSSION

The results of our experiments show, that (1) the fully automated analysis of phone-based SVF is feasible for dementia screening, (2) the phone-based pipeline produces classification results comparable to the gold-standard manual transcription based classifiers and (3) the word error rate for the ASR approach is acceptable despite the reduced telephone bandwidth and the aged population.

In general, regarding screening scenarios, high sensitivity scores are important. Our classification experiment based on the fully automated pipeline shows a good AUC and for screening scenario a good sensitivity of 0.85 and decent specificity of 0.65. For achieving better specificity results, it may be necessary to include additional tasks, especially focusing on the differentiation of MCI and healthy controls. Nevertheless, this is not the main goal for broad screening, as false positives are less expensive for a health-care system than false negatives.

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

[1] P. Aisen, J. Touchon, R. Amariglio, S. Andrieu, R. Bateman, J. Breitner, M. Donohue, B. Dunn, R. Doody, N. Fox, S. Gauthier, M. Grundman, S. Hendrix, C. Ho, M. Isaac, R. Raman, P. Rosenberg, R. Schindler, L. Schneider, R. Sperling, P. Tariot, K. Welsh-Bohmer, M. Weiner, and B. Vellas. 2017. EU/US/CTAD Task Force: Lessons Learned from Recent and Current Alzheimer's Prevention Trials. *J Prev Alzheimers Dis* 4, 2 (2017), 116–124.

[2] Sabah Al-hameed, Mohammed Benaissa, and Heidi Christensen. 2016. Simple and robust audio - based detection of biomarkers for Alzheimer' s disease. In *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*. 32–36.

[3] Sophie Auriacombe, Nathalie Lechevallier, Hélène Amieva, Sandrine Harston, Nadine Raoux, and J-F Dartigues. 2006. A Longitudinal Study of Quantitative and Qualitative Features of Category Verbal Fluency in Incident Alzheimer's Disease Subjects: Results from the PAQUID Study. *Dementia and geriatric cognitive disorders* 21, 4 (2006), 260–266.

[4] Paul Boersma and David Weenink. 2001. PRAAT, a system for doing phonetics by computer. *Glot international* 5 (2001), 341–345.

[5] Estefania Brando, Raquel Olmedo, and Carmen Solares. 2017. The application of technologies in dementia diagnosis and intervention: A literature review. 16 (05 2017), 1–11.

[6] T. C. Castanho, C. Portugal-Nunes, P. S. Moreira, L. Amorim, J. A. Palha, N. Sousa, and N. Correia Santos. 2016. Applicability of the Telephone Interview for Cognitive Status (Modified) in a community sample with low education level: association with an extensive neuropsychological battery. *Int J Geriatr Psychiatry* 31, 2 (Feb 2016), 128–136.

[7] Pau-Choo Chung, Yu-Liang Hsu, Chun-Yao Wang, Chien-Wen Lin, Jeen-Shing Wang, and Ming-Chyi Pai. 2012. Gait analysis for patients with Alzheimer's disease using a triaxial accelerometer. In *2012 IEEE International Symposium on Circuits and Systems*. IEEE, 1323–1326.

[8] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Learning* 20, 3 (1995), 273–297. https://doi.org/10.1023/A:1022627411411

[9] C. A. de Jager, A. C. Schrijnemaekers, T. E. Honey, and M. M. Budge. 2009. Detection of MCI in the clinic: evaluation of the sensitivity and specificity of a computerised test battery, the Hopkins Verbal Learning Test and the MMSE. *Age Ageing* 38, 4 (Jul 2009), 455–460.

[10] Hiroko H Dodge, Nora Mattek, Mattie Gregor, Molly Bowman, Adriana Seelye, Oscar Ybarra, Meysam Asgari, and Jeffrey A Kaye. 2015. Social Markers of Mild Cognitive Impairment: Proportion of Word Counts in Free Conversational Speech. *Current Alzheimer research* 12, 6 (2015), 513–519.

[11] B. Dubois, H. Hampel, H. H. Feldman, P. Scheltens, P. Aisen, S. Andrieu, H. Bakardjian, H. Benali, L. Bertram, K. Blennow, K. Broich, E. Cavedo, S. Crutch, J. F. Dartigues, C. Duyckaerts, S. Epelbaum, G. B. Frisoni, S. Gauthier, R. Genthon,

A. A. Gouw, M. O. Habert, D. M. Holtzman, M. Kivipelto, S. Lista, J. L. Molinuevo, S. E. O'Bryant, G. D. Rabinovici, C. Rowe, S. Salloway, L. S. Schneider, R. Sperling, M. Teichmann, M. C. Carrillo, J. Cummings, and C. R. Jack. 2016. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. *Alzheimers Dement* 12, 3 (Mar 2016), 292–323.

[12] A. Egerhazi, R. Berecz, E. Bartok, and I. Degrell. 2007. Automated Neuropsychological Test Battery (CANTAB) in mild cognitive impairment and in Alzheimer's disease. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 31, 3 (Apr 2007), 746–751.

[13] S. Epelbaum, R. Genthon, E. Cavedo, M. O. Habert, F. Lamari, G. Gagliardi, S. Lista, M. Teichmann, H. Bakardjian, H. Hampel, and B. Dubois. 2017. Preclinical Alzheimer's disease: A systematic review of the cohorts underlying the concept. *Alzheimers Dement* 13, 4 (Apr 2017), 454–467.

[14] Sven-Erik Fernaeus, Per Östberg, Åke Hellström, and Lars-Olof Wahlund. 2008. Cut the coda: Early fluency intervals predict diagnoses. *Cortex* 44, 2 (2008), 161 – 169. https://doi.org/10.1016/j.cortex.2006.04.002

[15] M. F. Folstein, S. E. Folstein, and P. R. McHugh. 1975. "Mini-Mental State". A Practical Method for Grading the Cognitive State of Patients for the Clinician. *J Psychiatr Res* 12, 3 (1975), 189–198.

[16] K. S. Fowler, M. M. Saling, E. L. Conway, J. M. Semple, and W. J. Louis. 1997. Computerized neuropsychological tests in the early detection of dementia: prospective findings. *J Int Neuropsychol Soc* 3, 2 (Mar 1997), 139–146.

[17] Céline Franco, Jacques Demongeot, Christophe Villemazet, and Nicolas Vuillerme. 2010. Behavioral Telemonitoring of the Elderly at Home: Detection of Nycthemeral Rhythms Drifts from Location Data. In *2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops*. IEEE, 759–766.

[18] Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz. 2016. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease* 49, 2 (2016), 407–422.

[19] Rowena G. Gomez and Desirée A. White. 2006. Using verbal fluency to detect very mild dementia of the Alzheimer type. *Archives of Clinical Neuropsychology* 21, 8 (2006), 771 – 775. https://doi.org/10.1016/j.acn.2006.06.012

[20] Gábor Gosztolya, László Tóth, Tamás Grósz, Veronika Vincze, Ildikó Hoffmann, Greta Szatloczki, Magdolna Pókáski, and János Kálmán. 2016. Detecting mild cognitive impairment from spontaneous speech by correlation-based phonetic feature selection. In *INTERSPEECH 2016–17th Annual Conference of the International Speech Communication Association*. 107–111. https://doi.org/10.21437/Interspeech.2016-384

[21] J. D. Grill and J. Karlawish. 2010. Addressing the challenges to successful recruitment and retention in Alzheimer's disease clinical trials. *Alzheimers Res Ther* 2, 6 (Dec 2010), 34.

[22] Paul J Gruenewald and Gregory R Lockhead. 1980. The Free Recall of Category Examples. *Journal of Experimental Psychology: Human Learning and Memory* 6 (1980), 225–240.

[23] C. T. Gualtieri and L. G. Johnson. 2006. Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs. *Arch Clin Neuropsychol* 21, 7 (Oct 2006), 623–643.

[24] Julie Henry and John R Crawford. 2005. A meta-analytic review of verbal fluency deficits in depression. *Journal of clinical and experimental neuropsychology* 27, 1 (jan 2005), 78–101. https://doi.org/10.1080/138033990513654

[25] Ildikó Hoffmann, Dezso Nemeth, Cristina D Dye, Magdolna Pákáski, Tamás Irinyi, and János Kálmán. 2010. Temporal Parameters of Spontaneous Speech in Alzheimer's Disease. *International Journal of Speech-Language Pathology* 12, 1 (2010), 29–34. https://doi.org/10.3109/17549500903137256

[26] Yu-Liang Hsu, Pau-Choo Chung, Wei-Hsin Wang, Ming-Chyi Pai, Chun-Yao Wang, Chien-Wen Lin, Hao-Li Wu, and Jeen-Shing Wang. 2014. Gait and Balance Analysis for Patients With Alzheimer?s Disease Using an Inertial-Sensor-Based Wearable Instrument. *IEEE Journal Of Biomedical And Health Informatics* 18, 6 (2014), 1822–1830.

[27] M. Inoue, D. Jimbo, M. Taniguchi, and K. Urakami. 2011. Touch Panel-type Dementia Assessment Scale: a new computer-based rating scale for Alzheimer's disease. *Psychogeriatrics* 11, 1 (Mar 2011), 28–33.

[28] Ali Khodabakhsh, Fatih Yesil, Ekrem Guner, and Cenk Demiroglu. 2015. Evaluation of Linguistic and Prosodic Features for Detection of Alzheimer's Disease in Turkish Conversational Speech. *EURASIP Journal on Audio, Speech, and Music Processing* 9 (2015), 1–15. https://doi.org/10.1186/s13636-015-0052-y

[29] Philipp Klumpp, Thomas Janu, Tomás Arias-Vergara, and Juan Camilo Vásquez Correa. 2017. Apkinson–A Mobile Monitoring Solution for Parkinson's Disease. *INTERSPEECH 2017–18th Annual Conference of the International Speech Communication Association* (2017), 1839–1843.

[30] Alexandra König, Carlos Fernando Crispim Junior, Alexandre Derreumaux, Gregory Bensadoun, Pierre-David Petit, François Bremond, Renaud David, Frans Verhey, Pauline Aalten, and Philippe Robert. 2015. Validation of an automatic video monitoring system for the detection of instrumental activities of daily living in dementia patients. *Journal of Alzheimer's Disease : JAD* 44, 2 (2015), 675âĂŤ685. https://doi.org/10.3233/jad-141767

[31] A. König, N. Linz, J. Töger, M. Wolters, J. Alexandersson, and P. Robert. 2018. Fully automatic analysis of semantic verbal fluency performance for the assessment of cognitive decline. *Dementia and Geriatric Cognitive Disorders* (2018). Accepted.

[32] Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H. Robert, and Renaud David. 2015. Automatic Speech Analysis for the Assessment of Patients with Predementia and Alzheimer's Disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 1, 1 (2015), 112–124.

[33] Maider Lehr, Emily Prud'hommeaux, Izhak Shafran, and Brian Roark. 2012. Fully automated neuropsychological assessment for detecting Mild Cognitive Impairment. In *INTERSPEECH 2012–13th Annual Conference of the International Speech Communication Association*. 1039–1042.

[34] Nicklas Linz, Johannes Tröger, Jan Alexandersson, and Alexandra König. 2017. Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*.

[35] Nicklas Linz, Johannes Tröger, Jan Alexandersson, Maria Wolters, Alexandra KÃűnig, and Philippe Robert. 2017. Predicting Dementia Screening and Staging Scores from Semantic Verbal Fluency Performance. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. 719–728. https://doi.org/10.1109/ICDMW.2017.100

[36] Brian MacWhinney. 1991. *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Inc.

[37] Guy M McKhann, David S Knopman, Howard Chertkow, Bradley T Hyman, Clifford R Jack, Claudia H Kawas, William E Klunk, Walter J Koroshetz, Jennifer J Manly, Richard Mayeux, et al. 2011. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia* 7, 3 (2011), 263–269.

[38] Juan José G Meilán, Francisco Martínez-Sánchez, Juan Carro, Dolores E. López, Lymarie Millian-Morell, and José M. Arana. 2014. Speech in Alzheimer's Disease: Can Temporal and Acoustic Parameters Discriminate Dementia? *Dementia and Geriatric Cognitive Disorders* 37, 5–6 (2014), 327–334.

[39] C. Munro Cullum, L.S. Hynan, M. Grosch, M. Parikh, and M.F. Weiner. 2014. Teleneuropsychology: Evidence for Video Teleconference-Based Neuropsychological Assessment. *Journal of the International Neuropsychological Society* 20, 10 (2014), 1028–1033. https://doi.org/10.1017/S1355617714000873

[40] Sid E O'Bryant, Stephen C Waring, C Munro Cullum, James Hall, Laura Lacritz, Paul J Massman, Philip J Lupo, Joan S Reisch, Rachelle Doody, and Texas Alzheimer's Research Texas Alzheimer's Research Consortium. 2008. Staging dementia using Clinical Dementia Rating Scale Sum of Boxes scores: a Texas Alzheimer's Research Consortium Study. *Arch. Neurol.* 65, 8 (2008), 1091–1095.

[41] Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen Jennifer Golden. 2014. Learning Predictive Linguistic Features for Alzheimer's Disease and related Dementias using Verbal Utterances. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 78–87.

[42] Paula Paavilainen, Ilkka Korhonen, Luc Cluitmans, Jyrki Lötjönen, Antti Särelä, and Markku Partinen. 2005. Circadian activity rhythm in demented and non-demented nursing-home residents measured by telemetric actigraphy. *Journal of Sleep Research* 14, 1 (2005), 61–68. https://doi.org/10.1111/j.1365-2869.2004.00433.x

[43] Serguei V.S. Pakhomov, Lynn Eberly, and David Knopman. 2016. Characterizing Cognitive Performance in a Large Longitudinal study of Aging with Computerized Semantic Indices of Verbal Fluency. *Neuropsychologia* 89 (2016), 42–56.

[44] Serguei V.S. Pakhomov, Susan E. Marino, Sarah Banks, and Charles Bernick. 2015. Using Automatic Speech Recognition to Assess Spoken Responses to Cognitive Tests of Semantic Verbal Fluency. *Speech Communication* 75 (2015), 14–26. https://doi.org/10.1016/j.specom.2015.09.010

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[46] Ronald C Petersen, Glenn E Smith, Stephen C Waring, Robert J Ivnik, Eric G Tangalos, and Emre Kokmen. 1999. Mild Cognitive Impairment: Clinical Characterization and Outcome. *Arch. Neurol.* 56, 3 (1999), 303–308.

[47] Martin Prince, Adelina Comas-Herrera, Martin Knapp, Maëlenn Guerchet, and Maria Karagiannidou. 2016. *World Alzheimer Report 2016 Improving Healthcare for People living with Dementia. Coverage, Quality and Costs now and in the Future.* Technical Report. 1–140 pages.

[48] Nadine Raoux, Hélène Amieva, Mélanie Le Goff, Sophie Auriacombe, Laure Carcaillon, Luc Letenneur, and Jean-François Dartigues. 2008. Clustering and switching processes in semantic verbal fluency in the course of Alzheimer's disease subjects: Results from the PAQUID longitudinal study. *Cortex* 44, 9 (2008), 1188–1196. https://doi.org/10.1016/j.cortex.2007.08.019

[49] Philippe H. Robert, Valérie Lafont, Isabelle Medecin, Laurence Berthet, Sandrine Thauby, Claude Baudu, and Guy Darcourt. 1998. Clustering and switching strategies in verbal fluency tasks: Comparison between schizophrenics and healthy adults. *Journal of the International Neuropsychological Society* 4, 6 (1998), 539–546.

[50] W. G. Rosen, R. C. Mohs, and K. L. Davis. 1984. A new rating scale for Alzheimer's disease. *Am J Psychiatry* 141, 11 (Nov 1984), 1356–1364.

[51] Aharon Satt, Ron Hoory, Alexandra König, Pauline Aalten, and Philippe H Robert. 2014. Speech-based Automatic and Robust Detection of very Early Dementia. In *INTERSPEECH 2014–15th Annual Conference of the International Speech Communication Association*. 2538–2542.

[52] S. Sindi, F. Mangialasche, and M. Kivipelto. 2015. Advances in the prevention of Alzheimer's Disease. *F1000Prime Rep* 7 (2015), 50.

[53] Toshiro Suzuki, Sumio Murase, Tomoyuki Tanaka, and Takako Okazawa. 2007. New Approach for The Early Detection of Dementia by Recording In-House Activities. *Telemedicine and e-Health* 13, 1 (2007), 41–44.

[54] Ioannis Tarnanas, Winfried Schlee, Magda Tsolaki, René Müri, Urs Mosimann, and Tobias Nef. 2013. Ecological Validity of Virtual Reality Daily Living Activities Screening for Early Dementia: Longitudinal Study. *JMIR Serious Games* 1, 1 (2013).

[55] Tom N Tombaugh, Jean Kozak, and Laura Rees. 1999. Normative Data Stratified by Age and Education for Two Measures of Verbal Fluency: FAS and Animal Naming. *Archives of Clinical Neuropsychology* 14, 2 (1999), 167–177. https://doi.org/10.1016/S0887-6177(97)00095-4

[56] László Tóth, Gábor Gosztolya, Veronika Vincze, Ildikó Hoffmann, Gréta Szatlóczki, Edit Biró, Fruzsina Zsura, Magdolna Pákáski, and János Kálmán. 2015. Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech using ASR. In *INTERSPEECH 2015–16th Annual Conference of the International Speech Communication Association*. 1–5.

[57] Johannes Tröger, Nicklas Linz, Jan Alexandersson, Alexandra König, and Philippe Robert. 2017. Automated Speech-based Screening for Alzheimer's Disease in a Care Service Scenario. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*.

[58] Angela K Troyer, Morris Moscovitch, and Gordon Winocur. 1997. Clustering and Switching as Two Components of Verbal Fluency: Evidence From Younger and Older Healthy Adults. *Neuropsychology* 11, 1 (1997), 138–146.

[59] Katherine Wild, Diane Howieson, Frank Webbe, Adriana Seelye, and Jeffrey Kaye. 2008. Status of computerized cognitive testing in aging: A systematic review. *Alzheimer's & Dementia* 4, 6 (2008), 428 – 437. https://doi.org/10.1016/j.jalz.2008.07.003

[60] World Health Organization. 1992. *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. World Health Organization.

[61] Bea Yu, Thomas F. Quatieri, James R. Williamson, and James C. Mundt. 2015. Cognitive impairment prediction in the elderly based on vocal biomarkers. In *INTERSPEECH 2015–16th Annual Conference of the International Speech Communication Association*. 3734–3738.