



Speech-Based Automated Cognitive Status Assessment

Dilek Hakkani-Tür¹ Dimitra Vergyi² Gokhan Tur²

¹International Computer Science Institute (ICSI), Berkeley, CA

²SRI International, Speech Technology and Research Lab, Menlo Park, CA

dilek@icsi.berkeley.edu, {dverg,gokhan}@speech.sri.com

Abstract

Verbal interviews performed by trained clinicians are a common form of assessments to measure cognitive decline. The aim in this paper is to study the usability of automated methods for evaluating verbal cognitive status assessment tests for the elderly. If reliable, such methods for cognitive assessment can be used for frequent, non-intrusive, low-cost screenings and provide objective and longitudinal cognitive status monitoring data that can complement regular clinical visits and would be useful for early detection of conditions associated with language and communication impairments. This study focuses on two types of tests: a story-recall test, used for memory and language functioning assessment, and a picture description test, used to assess the information content in speech. A data collection was designed for this study involving recordings of about 100 people, mostly over 70 years old, performing these tests. The speech samples were manually transcribed and annotated with semantic units in order to obtain manual evaluation scores. We explore the use of automatic speech recognition and language processing methods to derive objective, automatically extracted metrics of cognitive status that are highly correlated with the manual scores. We use recall and precision based metrics based on semantic content units associated with the tests. Our experiments show high correlation between manually obtained scores and the automatic metrics obtained using either manual or automatic speech transcriptions.

Index Terms: speech recognition, language processing, automated cognitive status assessment, elderly speech

1. Introduction

Cognition defines mental functions such as the ability to think, reason, perceive, judge, and remember. These functions are often evaluated using various non-automated tests to assess the natural history of cognitive decline, to study the need for care and the capacity of independent living, and to evaluate treatment [1]. Declines in cognitive status usually result in declines in recall and memory, concentration or reasoning, and selection of appropriate words. There are several commonly used instruments utilized by general practitioners and specialists to determine cognitive impairment, cognitive decline, Traumatic Brain Injury, driving capability, or aid in the determination of dementia or neurodegenerative diseases such as Alzheimer's disease.

The General Practitioner Assessment of Cognition (GPCOG) [2] consists of cognitive test items, that include a memory test, a test to do with paper and pencil and two verbal questions, in addition to historical questions asked of an informant. The mini mental state examination (MMSE) [3] is the most commonly used instrument for screening cognitive decline. It was developed by psychiatrists and is widely regarded as the 'gold standard' test for dementia. The MMSE

test includes simple questions and problems in a number of areas: the time and place of the test, repeating lists of words, arithmetics, such as the serial sevens, language use and comprehension, and basic motor skills. It provides measures of orientation, registration (immediate memory), short-term memory (but not long-term memory) as well as language functioning. Both GPCOG and MMSE are considered general cognitive assessment tests. Other tests target assessment for more specific impairments. For example, the Western Aphasia Battery (WAB) [4] or the Boston Diagnostic Aphasia Examination (BDAE) [5] are instruments used to assess aphasia, which is a language disorder that may include difficulty in producing or comprehending spoken or written language. These involve a series of subtasks assessing each aspect of language functionality. The WAB comprises of several subtasks that include a fable retelling, a story derived from a picture sequence, single picture descriptions and topic-elicited narratives to evaluate language abilities, and may take more than an hour to administer.

These tests are administered and evaluated by trained clinicians, and, even the simplest ones are expensive and time-consuming, since they require at least one visit to the clinician's office. Thus, they are only administered if the care-taker is alerted of some significant change in the behavior or abilities of the person in concern. Among elderly people who do not present with complaints of memory impairment, or who are not in everyday contact with care-takers, diagnosis of cognitive decline may be delayed, often with critical consequences to the general health of the individual. Therefore, easily administered, reliable and cost effective dementia screening tests are needed for elderly individuals.

Some previous efforts have addressed the need for tests that can be administered easily and at low-cost. The TYM ("test your memory") [6] and the "pencil and paper" Cognitive Assessment Screening Test (CAST) [7] are simple tests that take minimal examiner time, but they still need human scoring. Other efforts propose using computerized instruments with automated scoring [8]. Previous studies have also investigated use of speech input for automatic cognitive assessment. Roark *et al.* [9] investigated spoken language markers of Mild Cognitive Impairment. They found standardized pause rate (the ratio of words uttered to the number of pauses uttered) to be statistically different for two sets of elderly speakers; using a Clinical Dementia Rating (CDR), the cohort was separated into those with MCI and those without. D'Arcy *et al.* [10] investigated temporal language features as indicators for the onset of cognitive decline for the elderly. Roark *et al.* [11] proposed measures for syntactic complexity and investigated the utility of these for discriminating between clinically defined groups. Peintner *et al.* [12] showed that speech and language measures can be used as feature in a classification model for different types of Frontotemporal Lobal Degeneration.

	≤70		> 70		All
Immediate story retelling					
	M	F	M	F	
Number of speakers	15	22	21	38	96
Total No. of words	930	1081	1777	2088	5876
Avg. No. of words	62	49	84	54	61
ASR WER (%)	28.2	23.6	32.5	34.1	30.7
WAB picture description					
	M	F	M	F	
Number of speakers	17	25	21	35	102
Total No. of words	1952	2709	2783	5052	12496
Avg. No. of words	114	108	132	144	122
ASR WER (%)	27.5	21.1	31.9	26.3	26.7

Table 1: Properties of the immediate story retelling and the WAB picture description data set used in experiments.

In this work we focus on automatic evaluation of two types of commonly used verbal tests for assessing cognitive and language impairment: story retelling and picture description. Story retelling is part of the Logical Memory subtest of the Wechsler Memory Scale (WMS) [13] and has been useful to show degradation in memory skills related to very mild dementia of the Alzheimer type [14]. Picture description is included in standard aphasia evaluation instruments, such as the WAB test. In our approach we use automatic transcriptions of patient’s speech produced using automatic speech recognition (ASR). The goal of our work is to show that completely automated metrics derived from the subject’s speech, such as n-gram recall, precision and fmeasure, have high correlation with manual evaluation for the selected subtasks, and that the high correlation is maintained even when the metrics are derived from automatic speech transcriptions.

In the following section we describe the data collected and used in this work. Then in section 3, we describe automatic assessment procedure. In the experiments section, we first analyze the ASR performance for elderly subjects and show that the automatic assessment scores computed from manual and automatic transcriptions correlate well with manual evaluation.

2. Data collection

A variety of discourse elicitation procedures are used by clinicians and researchers. These include conversation interview, story telling, description of common procedures and description of pictures. Since there was no publicly available speech database of adults performing these tasks we designed a data collection specifically targeting the elderly population. Younger speakers were also included, performing the same tasks, to serve as controls.

We interviewed and recorded a total of 123 subjects of ages 20-102. For each session, both the interviewer and subject were wearing head-mounted noise canceling microphones. A high quality digital recorder (Tescam HD-P2) was used to record in 2-channels (one for each speaker) with 44KHz and 16bits. The interviews consisted of 6 parts, each recorded as a separate file: In the first part the subject was given a printout of the “Grandfather Passage” (Figure 1) in big fonts, and was asked to read it aloud, and then return it. In the second part the subject was asked to retell the story he just read (immediate story retell). In the third part the subject was shown the picture in Figure 2 and was asked to describe it in detail. Following that the subject was given a printout of a news story, and was asked to read it aloud. The fifth part was a structured interview on activities of daily

The Grandfather Passage: *You wished to know all about my grandfather. Well, he is nearly ninety-three years old; he dresses himself in an ancient black frock coat, usually minus several buttons; yet he still thinks as swiftly as ever. A long, flowing beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice is just a bit cracked and quivers a trifle. Twice each day he plays skilfully and with zest upon our small organ. Except in the winter when the ooze or snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, Banana oil! Grandfather likes to be modern in his language.*

Figure 1: The passage used in our experiments, from [16].



Figure 2: The picture used for the WAB test [4].

living [15], followed by a more free-form conversation on hobbies and some interesting past experience. Finally, the subject was asked to re-tell whatever he remembered of the story in part 1 (delayed story re-tell).

The overall interview took on average of about 45 minutes. The subjects’ gender, age and education level (in the form of years in school) were recorded. Access to medical/mental health records was not available but the subjects were asked if they had any general health issues.

For this study, only the immediate story retelling (part 2) and the picture description (part 3) are used. The “Grandfather Passage”, used as the story, is one of the standard reading passages used by speech pathologists to assess an individual’s ability to produce connected speech, since it contains almost every sound in the English language. Table 1 presents the details of the data used for this paper for 2 age groups: younger and older than 70 years old. All sessions used were manually transcribed and the audio was segmented in sentences.

3. Automatic Assessment Method

The high level approach proposed here aims to derive automatic scores for the performed tasks based on the spoken words. For a fully automated procedure the words are automatically transcribed using an ASR system. While in this paper we present results only on two tests, story retelling and picture description, the approach can be generalized to other tests as well.

3.1. Automatic Transcription of elderly speech

To obtain automatic transcriptions for the recorded data, we run a recognizer developed for recognition of meetings with close talking microphones [17]. No model adaptation was used for these experiments, even though we expect the performance of the system to improve if we adapt the acoustic models to elderly speakers, since the system performance degrades for older

speakers, and adapt the language models to some example interviews, since both the story retelling and the picture description are lexically constrained tasks. The word error rate (WER) of the speech recognizer for different subsets of the data, is shown in Table 1. We notice a significant degradation in recognition accuracy for speakers over 70, as expected, since the acoustic models used were trained on younger speakers.

3.2. Assessment Approach

In clinical cognitive status assessment, the main evaluation criterion is based on the semantic content units [14, 18].

For the story retelling test, similar to [14], we manually extracted 35 atomic semantic content units for the grandfather passage, such as “My grandfather plays organ” and “He has a long beard.” We manually scored transcriptions of retelling of the grandfather story for each subject for these content units, computing their recall. If the subject includes a sentence from which this content unit can be entailed in his/her retelling of the story, he/she gets a score of one for that content unit.

For the second test, we also rely on the recall aspect. As we do not have a written description of the picture, the information content of their speech was evaluated based in the number of information units produced from a list of 36 units in 4 key categories, subjects (such as “girl” and “couple”), places (such as “on the beach”), objects (such as “kite” and “flag”) and actions (such as “fishing”) as listed in [18].

The goal of automatic assessment is then detecting how many of these semantic content units the speaker has uttered. While this is a non-trivial problem due to linguistic variations coming with natural language, we believe high correlation can be achieved if the score is aggregated over all semantic concepts. The main idea behind automated scoring of input utterances is based on information retrieval and speech summarization metrics. These tests require the speaker to utter key concepts and only those. In this sense, this is very similar to evaluating a natural language summarization system. Hence, we measured recall, precision and F-measure of unigrams and bigrams in the retelling of the story with respect to the original story and recall for description of the picture.

There may be several factors that affect the performance in these tests, such as education level, in addition to the cognitive status, and the experiments can be extended with more information about the health status of the subject. Our aim is testing the usability of ASR technology for performing these tests (for example on the phone) instead, and interpretation of these scores is left to the clinicians.

4. Experiments

4.1. Story Retelling

In this section, we present results of manual recall evaluation tests. The average scores for each group from manual evaluation tests are presented in Table 2.

Table 2 presents the unigram and bigram recall, precision, and F-measures computed using the manual transcription of the subject’s retelling of the story and its ASR transcription, respectively, averaged for each age group. While we observed a small decline in the performance between the younger group and the older group, these experiments aim to compute the correlation between the tests results using manual transcriptions and ASR transcriptions. As can be seen from the table, the performance decline between age groups is also observed when the tests are performed using the ASR transcriptions.

Figure 3 plots the unigram F-measures computed from manual against those computed from ASR transcriptions for

	Man.	1-R	1-P	1-F	2-R	2-P	2-F
Age	Using manual transcriptions						
≤ 70	4.62	42.7	16.5	23.1	8.0	3.3	4.5
> 70	3.36	36.7	15.6	20.5	5.1	2.3	2.9
	Using ASR transcriptions						
≤ 70		37.7	13.9	19.7	6.7	2.6	3.6
> 70		32.9	13.3	17.5	4.0	1.8	2.3

Table 2: Manual (Man.) and automatic average evaluation scores for each age group. The automatic scores for recall (R), precision (P) and F-measure (F) for unigrams (1) and bigrams (2) are obtained both from manual transcriptions and from ASR output for comparison.

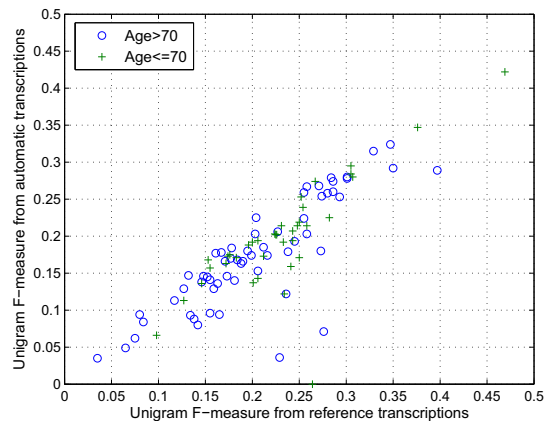


Figure 3: Unigram F-measures per speaker as estimated from reference and speech recognizer transcriptions for the story retelling.

each subject. Pearson’s correlation coefficient between the two F-measure numbers in the figure is measured to be 0.82, which is considered to be high.

Table 3 shows the correlation numbers between manual evaluation scores and automatic metrics derived from manual (Corr-REF) and ASR (Corr-ASR) transcriptions. For both the manual and ASR transcriptions, unigram F-measure has the highest correlation with the manual scores. The correlation for manual transcriptions is very high, showing the usability of these scores for cognitive status assessment, instead of the manual evaluation. The correlation for ASR transcriptions is also high for some measures and can be improved with better ASR.

4.2. Picture Description

In this section, we present results of manual recall evaluation for the WAB picture description test. The average scores for each group from manual evaluation tests are presented in Table 4. This table also presents the unigram recall values computed using the manual transcription of the subject’s description of the picture, averaged for each age group.

While we observed a small decline in the performance between the younger group and the older group, similar to the previous test, these experiments aim to compute the correlation between the tests results using manual and automated scores. Our aim is testing the usability of ASR technology for performing these tests (for example on the phone) instead.

Figure 4 extends these results by plotting the unigram recall computed from ASR transcriptions against that computed from manual transcriptions, for each subject. Pearson’s correlation coefficients for the two recall numbers is measured to be 0.92,

	1-R	1-P	1-F	2-R	2-P	2-F
Corr-REF	0.474	0.737	0.846	0.675	0.815	0.791
Corr-ASR	0.406	0.616	0.703	0.478	0.703	0.668

Table 3: Correlation coefficients between the manual scores and different automatic metrics, computed from manual transcriptions and ASR transcriptions for the story retelling.

	Manual Score		Unigram Recall	
	Age ≤ 70	Age > 70	Age ≤ 70	Age > 70
Overall	16.5	15.1	16.6	14.4
Object	5.8	5.5	5.8	5.3
Place	1.7	1.5	1.8	1.6
Subject	3.9	3.4	3.6	3.2
Action	5.0	4.7	5.4	4.3

Table 4: Manual and automatic evaluation scores computed from manual transcriptions, averaged for each age group for each assessment category.

which is considered to be very high.

Table 5 shows the correlation numbers between manual scores and automatic scores from manual transcriptions and ASR transcriptions. For both the manual and ASR transcriptions, automatic assessment results in very high correlation with the manual scores. This shows the usability of these scores for cognitive status assessment, instead of the manual evaluation. As expected the correlation is much higher for the objects compared to subjects as the physical objects tend to be mentioned with less linguistic variability.

5. Conclusions

We experimented with automatic measures for assessing cognitive status based on automatic speech processing. This study is a first attempt towards building a assessment system which results in some scores highly correlated with the manual evaluation performed by the clinicians. We believe, such systems are important for low-cost, objective, and longitudinal cognitive status monitoring that can complement regular clinical visits. We hope to continue our studies using other parts of the elderly speech corpora, assessing daily living activities speech impairments.

Acknowledgments: We wish to thank Colleen Richey and Shahab Khan for efforts during the elderly speech data collection and transcription which was funded by the SRI.

6. References

- [1] I. McDowell and C. Newell, *Measuring Health: A Guide to Rating Scales and Questionnaires*, pp. XXX–YYY, Oxford University Press, 1996.
- [2] H. Brodaty, N.M. Kemp, and L. Low, “Characteristics of the gpcog, a screening tool for cognitive impairment,” *Int. Journal of Geriatric Psychiatry*, vol. 19, pp. 870–874, 2004.
- [3] M. F. Folstein, S. E. Folstein, and P. R. McHugh, “Mini-mental state,” *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–198, 1975.
- [4] A. Kertesz, *The Western aphasia batter*, Grune and Stratton, New York, NY, 1982.
- [5] H. Goodglass and E. Kaplan, *Boston Diagnostic Aphasia Examination*, Lea and Febiger, Philadelphia, PA, 1993.
- [6] J. Brown, G. Pengas, K. Dawson, L. A Brown, and P. Clatworthy, “Self administered cognitive screening test (tym) for detection of alzheimer’s disease: cross sectional study,” *BMJ*, vol. 338, no. b2030, 2009.
- [7] D. A. Drachman, J. M. Swearer, K. Kane, D. O. Osgood, C. Toole, and M. Moonis, “The cognitive assessment screening test (cast) for dementia,” *Journal of Geriatric Psychiatry and Neurology*, vol. 9, no. 4, pp. 200–208, 1996.

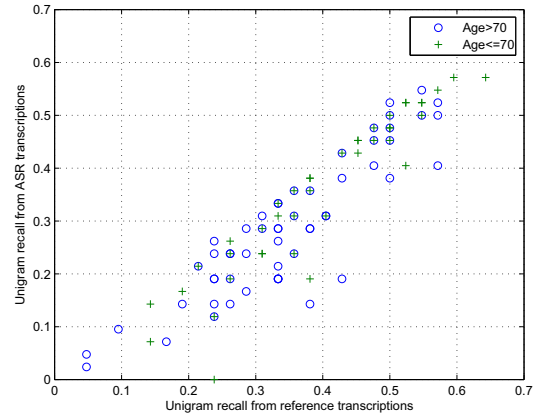


Figure 4: Unigram recall per speaker as estimated from reference and speech recognizer transcriptions.

Automatic Evaluation	Corr-REF	Corr-ASR
Overall	0.93	0.89
Object	0.95	0.91
Place	0.72	0.62
Subject	0.60	0.63
Action	0.85	0.75

Table 5: Correlation between manual scores and automatic metrics (unigram recall) computed from manual (Corr-REF) and ASR (Corr-ASR) transcriptions .

- [8] Cogstate Clinical Trials, “www.cogstate.com/go/clinicaltrials,” .
- [9] B. Roark, J.P. Hosom, M. Mitchell, and J.A. Kaye, “Automatically derived spoken language markers for detecting mild cognitive impairment,” in *2nd International Conference on Technology and Aging (ICTA)*, Toronto, Canada, June 2007.
- [10] S. D’Arcy, V. Rapcan, N. Penard, M. E. Morris, and R. B. Reilly I. H. Robertson, “Speech as a means of monitoring cognitive function of elderly speakers,” in *Interspeech*, Brisbane, Australia, 2008.
- [11] B. Roark, M. Mitchell, and K. Hollingshead, “Syntactic complexity measures for detecting mild cognitive impairment,” in *ACL Workshop on BioNLP: Biological, translational, and clinical language processing*, Prague, Czech Republic, June 2007.
- [12] B. Peintner, W. Jarrold, D. Vergyri, C. Richey, M. Tempini, M. Gorno, and J. Ogar, “Learning diagnostic models using speech and language measures,” in *Proc of the 30th Annual International IEEE EMBS Conference*, Vancouver, British Columbia, Canada, Aug. 2008.
- [13] D. Wechsler, “Standardized memory scale for clinical use,” *Journal of Psychology*, vol. 19, pp. 87–95, 1954.
- [14] D.K. Johnson, M. Storandt, and D. A. Balota, “Discourse analysis of logical memory recall in normal aging and in dementia of the alzheimer type,” *Neuropsychology*, vol. 17, no. 1, pp. 82–92, 2004.
- [15] M.P. Lawton and E.M. Brody, “Assessment of older people: Self maintaining and instrumental activities of daily living,” *The Gerontologist*, vol. 9, no. 3 (Part 1), pp. 179–186, 1969.
- [16] M.S. Cannizzaro, N. Reilly, J. Mundt, and P.J. Snyder, “Remote capture of human voice acoustical data by telephone: A methods study,” *Clinical Linguistics and Phonetics*, vol. 19, no. 8, pp. 649–658, December 2005.
- [17] A. Stolcke, K. Boakye, Ö. Çetin, A. Janin, M. Magimai-Doss, C. Wooters, and J. Zheng, “The SRI-ICSI Spring 2007 meeting and lecture recognition system,” in *Proc. NIST 2007 Rich Transcription Workshop*, 2007.
- [18] A. M. Jensen, H. J. Chenery, and D. A. Copland, “A comparison of picture description abilities in individuals with vascular subcortical lesions and huntington’s disease,” *Journal of Communication Disorders*, vol. 39, no. 1, pp. 62–77, 2006.