AUTOMATICALLY DERIVED SPOKEN LANGUAGE MARKERS FOR DETECTING MILD COGNITIVE IMPAIRMENT

Brian Roark, John-Paul Hosom, Margaret Mitchell
Center for Spoken Language Understanding
OGI School of Science & Engineering
Oregon Health & Science University
{roark,hosom,meg.mitchell}@cslu.ogi.edu

Jeffrey A. Kaye
Layton Aging & Alzheimer's Disease Center and
Oregon Center for Aging & Technology (ORCATECH)
Oregon Health & Science University
kaye@ohsu.edu

ABSTRACT

Speech produced by subjects during neuropsychological exams can provide markers other than test performance, via spoken language characteristics that discriminate between subject groups. We present preliminary results on the utility of such markers, automatically derived from spoken responses to narrative recall tests, in discriminating between healthy elderly and subjects with Mild Cognitive Impairment (MCI). Given the audio and transcript of the retellings, a range of markers were automatically derived, including (among others) pause frequency and grammatical complexity. Certain spoken language derived markers, which do not measure the fidelity of the retelling to the original narrative, show statistically significant differences between the group means, when calculated either manually or automatically.

1. INTRODUCTION

Mild Cognitive Impairment (MCI), and in particular amnestic MCI, the earliest clinically defined stage of Alzheimer's related dementia, often goes undiagnosed due to the inadequacy of common screening tests such as the MMSE for reliably detecting relatively subtle impairments. Linguistic memory tests, such as word list and narrative recall, are more effective than the MMSE in detecting MCI, yet are still individually insufficient for adequate discrimination between healthy and impaired subjects. Because of this, a battery of examinations is typically used to improve psychometric classification. Yet the summary recall scores derived from these linguistic memory tests (total correctly recalled) ignore potentially useful information in the characteristics of the spoken language itself. Narrative retellings provide a natural, conversational speech sample that can be analyzed for many characteristics of the speech and language that have been shown to discriminate between healthy and impaired subjects, including syntactic complexity [10, 11] or mean pause duration [16]. These measures go beyond simply measuring fidelity to the narrative, thus providing key additional dimensions for improved diagnosis of impairment.

This study focuses on spoken language markers derived from transcribed audio of narratives elicited as part of the Wechsler Logical Memory (LM) tests, which include both an immediate (LM I) and delayed (LM II) recall. Working with neuropsychological tests has several key benefits. First, methods will have direct clinical applicability, because they will apply to standard tests that are already in use in clinical settings. Second, to the extent that additional discriminative utility can be derived from the output of any particular neuropsychological test, the number of tests required for reliable screening will be reduced, leading to a better chance of widespread testing due to reduced demands on both clinicians and patients. Finally, the relatively constrained elicitation, focused on a fixed narrative, makes automation of marker extraction particularly feasible, because of the narrow topic-focused language use. As we shall demonstrate below, this ease of automation does not come at the expense of the utility of the spoken language markers: there remain enough differences in the spoken language to provide markers of good discriminative utility.

2. METHODS

2.1. Data

We collected audio recordings of 44 neuropsychological examinations administered at the Layton Aging & Alzheimer's Disease Center, an NIA-funded Alzheimer's center for research at OHSU. For this study, we partitioned subjects into two groups: those who were assigned a Clinical Dementia Rating (CDR) of 0 (healthy) and those who were assigned a CDR of 0.5 (MCI). The CDR [13] is assigned with access to clinical and cognitive test information, independent of performance on the battery of neuropsychological tests used for research study.[1] Studies at the Layton Center, from which our subjects were drawn, define MCI in two ways: via the CDR scale and via a psychometrically driven concept of degraded performance on neuropsychological tests. Given that we are studying some of the very neuropsychological tests that play a role in the latter definition of MCI, we must rely on those which do not depend on these test scores – in particular the CDR scale. The global CDR score has been shown to have high expert inter-annotator reliability [14], and, critically, provides subject assessments that are independent of the neuropsychological exams being used in this study.

Of the collected recordings, three subjects were recorded twice; for the current study only one recording was used for each subject. Two subjects were assigned a CDR of 1.0 and were excluded from the study; two further subjects were excluded for errors in the recording that resulted in missing audio. Of the remaining 37 subjects, 22 were in the healthy elderly group (CDR = 0) and 15 were in the MCI elderly group (CDR = 0.5).

[1] See [13] for specific details about the CDR.

	CDR = 0 (n=22)		CDR = 0.5 (n=15)		
Measure	M	SD	M	SD	$t(35)$
Age	87.0	9.6	91.3	4.5	-1.59
Education (Y)	14.6	2.2	14.1	2.8	0.68
MMSE	28.6	1.3	25.8	2.8	4.10^{***}
Word List (A)	20.5	3.9	15.7	3.1	4.00^{***}
Word List (R)	7.0	1.7	3.9	1.4	5.97^{***}
Wechsler LM I	17.5	4.3	10.3	4.0	5.17^{***}
Wechsler LM II	16.0	4.2	9.0	4.8	4.69^{***}
Cat.Fluency (A)	17.0	3.5	13.3	4.1	3.00^{**}
Cat.Fluency (V)	13.0	4.5	9.1	3.4	2.86^{**}
Digits (F)	6.3	1.5	5.9	1.2	0.84
Digits (B)	4.6	1.0	4.6	1.2	0.10

Table 1. Neuropsychological test results for subjects. $^{***}\ p < 0.001$; $^{**}\ p < 0.01$

2.2. Neuropsychological Tests

Table 1 presents means and standard deviations for age, years of education and the scores of a number of standard neuropsychological tests that were administered during the recorded session. These tests include: the Mini Mental State Examination (MMSE); the CERAD Word List Acquisition (A) and Recall (R) tests; the Wechsler Logical Memory (LM) I (immediate) and II (delayed) narrative recall tests; Category Fluency, Animals (A) and Vegetables (V); and Digit Span (WAIS-R) forward (F) and backward (B).

2.3. Spoken language markers

To evaluate the utility of automation of spoken language marker extraction, we first must extract markers manually. We manually annotated the Wechsler Logical Memory I/II retellings to allow for manual marker extraction, and established markers for which there was a statistically significant difference between group means. Data annotation included producing (i) a time-aligned transcript of each retelling, from which speech duration-based markers could be derived; and (ii) a full syntactic parse tree, from which syntactic markers could be derived. In addition to allowing for manual marker extraction, this annotation provides training and evaluation data for automated annotation.

2.3.1. Syntactic complexity markers

For this study, we followed the syntactic annotation style of the well-known University of Pennsylvania Wall St. Journal Treebank [12], an example of which is shown in Figure 1. This tree segments the string "she was a cook in a cafeteria" into hierarchically arranged labeled constituents. For example, in this particular tree, there is a prepositional phrase (PP) consisting of the words "in a cafeteria", and three noun phrases (NP) for "she", "a cook", and "a cafeteria". The spoken language of the narrative retellings contain disfluencies as well as sentence fragments, ungrammaticalities, and filled pauses. We assume that the transcription includes disfluencies in the utterance, but that disfluent regions ("EDITED" in the Penn Treebank) are indicated in the transcription.

There are many approaches to scoring syntactic complexity, such as the scoring methods of Yngve [17] and Frazier [7]. Yngve [17] is a very simple scoring approach based on tree shapes which we will discuss below. Frazier [7] differs from the Yngve score by explicitly penalizing embedding. D-Level [15] is a composite score associated with certain kinds of constructions. In addition to these composite metrics, the counts of certain kinds of syntactic phenomena, e.g., embedded clauses, can be used as indicators of syntactic complexity. Cheung and Kemper [6] found that all of these syntactic complexity measures were very highly correlated, hence for this study we will focus on the relatively simple Yngve metric, as well as measures such as words per clause.

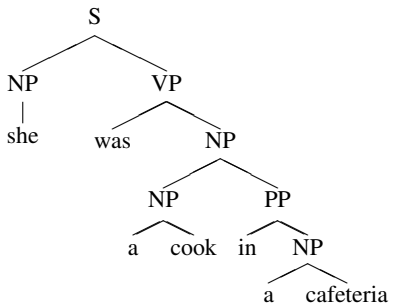

Fig. 1. Penn Treebank style syntactic annotation

The Yngve approach gives scores to branches as follows: the rightmost branch receives a score of 0, and each branch moving towards the left gets one more than the branch to its right. So, for example, the ternary branching tree in Figure 2, with three branches at the root of the tree (labeled with D), gives a score of 0 to the rightmost branch, a score of 1 to the middle branch, and a score of 2 to the leftmost branch. The algorithm gives a score to each word, calculated by summing the weights on all branches from the root of the tree at the top, down to the word. So, for example, the letter "a" in the left-branching tree of Figure 2(C) is reached by following the left-branch three times. Each left-branch is given a score of 1, hence the letter "a" in the left-branching tree gets a score of 3. In that tree, the letter "b" gets a score of 2, the letter "c" a score of 1 and the letter "d" a score of 0. For the string "a b c d" in that tree, the mean Yngve syntactic complexity score is 1.5 (6 total points over 4 words). In contrast, the right-branching structure gives a score of 1 to a, b and c, and a score of 0 to d, for a total score of 3 and a mean of 0.75, half of the left-branching score. The other two tree shapes in Figure 2 give a mean score of 1.25.

We also calculated the mean words per clause in a retelling. A clause is defined as a constituent in the parse tree with one of the Penn Treebank clause labels: S, SBAR, SQ, SBARQ or SINV. The number of words is provided by the transcript, and the number of clauses by the syntactic parse tree, either manually or automatically annotated.

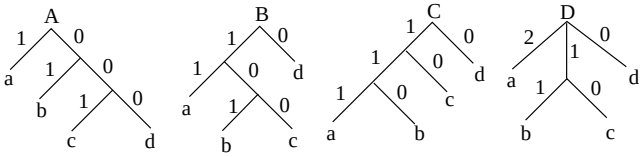

Fig. 2. Four tree shapes, and branch scores for calculating Yngve syntactic complexity metric: A) right-branching; B) center-embedded; C) left-branching; and D) ternary branching.

2.3.2. Speech duration markers

In addition to markers derived from the syntactic structure of the utterances in the retelling, we derived markers from temporal aspects of the speech sample, such as number of pauses and total pause duration. We follow Singh et al. [16] in setting a 1 second minimum for counting silence as a pause. Given the number of words in the sample (W), the number of pauses (N), the total pause time (P) and the total time of the sample (T), we can calculate a number of markers, again following [16], including: Verbal Rate (W/T), Phonation Rate ($(T-P)/T$), Mean Duration of Pauses (P/N) and Standardized Pause Rate (W/N). These markers can be deterministically calculated given a time-aligned transcription, whether the time-alignment is manually or automatically produced.

2.4. Automated Marker Extraction

2.4.1. Parsing for Syntactic Complexity Markers

Automation of syntactic complexity marker extraction was done through the use of a high-accuracy statistical parser. For this study, we chose to use the Charniak parser, which has the highest reported accuracy on more than one standard parsing task [3, 4, 5]. This parser is available for research purposes, and is trainable. When parsing spontaneous speech, the best practice is to remove disfluencies in advance of parsing [4], which was done for these trials.

To evaluate parsing accuracy, constituents are associated with spans of words. For example, in the tree in Figure 1, there is a VP (verb phrase) constituent that spans the words *"was a cook in a cafeteria"*, as well as an NP constituent spanning the words *"a cook"* and a PP constituent spanning *"in a cafeteria"*. We can compare the manually annotated parse tree with the tree that an automatic parser produces by counting how many of the constituents in the tree have the same label (e.g., NP, PP, VP) and the same span of words. If a labeled constituent with a particular span exists in both the true (manually annotated) tree and in the tree produced by the automatic parser, we say that the constituents match. *Labeled precision* (LP) is the number of matching constituents divided by the number of constituents in the automatic parse. *Labeled recall* (LR) is the number of matching constituents divided by the number of constituents in the true parse. F-measure accuracy is the harmonic mean of LP and LR. These are the most widely used parser evaluation measures in the research literature, known as the PARSEVAL metrics [2].

The baseline system was trained on the Switchboard Treebank, part of the Penn Treebank-3, released through the Linguistics Data Consortium[2]. This consists of syntactically annotated telephone conversations on a variety of topics, which is closer to the conversational style of the narrative retellings than other syntactically annotated corpora, such as the WSJ Treebank. The Switchboard Treebank contains approximately 1 million words, and uses the same annotation style that we used when manually annotating the collected narrative retellings. The results using the model trained on this out-of-domain data are the Baseline row in Table 2.

[2]http://www.ldc.upenn.edu catalog number LDC99T42

System	LR	LP	F-measure
Baseline	84.4	86.8	85.6
Domain adapted	87.0	88.4	87.7

Table 2. Parser accuracy using a baseline system (trained on out-of-domain data only) versus using a domain adapted system.

To perform domain adaptation, we used a cross-validation technique, so that we could evaluate parsing accuracy over all retellings. For each of the 37 subjects, we created a small in-domain training corpus consisting of the retellings of the other 36 subjects. We then performed MAP adaptation [8, 1] of the baseline model, using count merging with an in-domain scale of 100, to produce a domain adapted model. In such a way, for each subject, the subject's own utterances (and the parses of those utterances) were not seen in the training data, thus simulating a real test-time application of the parser. The results are shown in Table 2. Adaptation improved the parsing accuracy by over two percent absolute (15% relative error reduction) versus the baseline. We used the adapted models to produce trees for automated syntactic complexity marker extraction.

2.4.2. Forced alignment for Pause Durations

Given a word-level transcript, a process called "forced alignment" can be used [9]. Forced alignment uses an existing ASR system, and constrains it so that it can only recognize the (known) word sequence. The output contains the location in the speech signal of each word and pause event. The forced alignment system developed at OHSU is state-of-the-art, placing 92.6% of phonemes boundaries within 20 milliseconds of manual boundaries on the TIMIT corpus.[3]

3. RESULTS

Table 3 shows group means and standard deviations of a selection of spoken language markers derived from the Logical Memory I/II retellings, both manually and automatically extracted. Of the two reported syntactic complexity markers, Words per clause and the Yngve score per word, both were statistically significantly different between the groups for the delayed test (Logical Memory II), though the Yngve score for Logical Memory I was not statistically significantly different between the two groups. These differences held for both manual and automatic marker extraction. In contrast, of the speech duration based markers – Verbal rate, Phonation rate, Mean pause duration and Standardized pause rate – only Standardized pause rate for Logical Memory I showed a significant difference when manually extracted. This statistical significance was not maintained when automatically extracted, despite the maintenance of relatively large differences between the groups.

4. DISCUSSION

There are several points that can be made from the presented results. First, we have strong evidence that automated extraction of these markers can be effective, given the preser-

[3]Boundary agreement between two humans on the TIMIT corpus is 93.5% within 20 milliseconds.

	Logical Memory I					Logical Memory II				
	CDR = 0		CDR = 0.5			CDR = 0		CDR = 0.5		
Measure	M	SD	M	SD	$t(35)$	M	SD	M	SD	$t(35)$
Total words in retelling	74.9	34.0	60.7	35.7	1.22	80.3	33.3	59.7	35.9	1.79
Manually extracted: Words per clause	6.53	1.23	5.31	1.21	2.98^{**}	6.58	1.02	4.86	1.79	3.72^{***}
Yngve score per word	1.42	0.22	1.41	0.22	0.15	1.54	0.28	1.25	0.46	2.35^{*}
Verbal rate	1.39	0.47	1.34	0.41	0.37	1.75	0.62	1.75	0.70	0.04
Phonation rate	0.48	0.13	0.46	0.08	0.64	0.57	0.12	0.56	0.21	0.25
Mean pause duration	4.40	1.39	3.53	1.41	1.85	3.44	1.56	2.67	1.37	1.55
Standardized pause rate	12.29	6.09	8.32	3.44	2.28^{*}	15.75	18.04	11.35	6.91	0.90
Auto extracted: Words per clause	6.54	1.44	5.13	1.07	3.22^{**}	6.54	1.25	4.98	1.97	2.94^{**}
Yngve score per word	1.35	0.21	1.30	0.22	0.66	1.48	0.25	1.19	0.45	2.44^{*}
Verbal rate	1.43	0.49	1.38	0.38	0.34	1.79	0.64	1.83	0.81	-0.19
Phonation rate	0.36	0.12	0.36	0.06	0.23	0.43	0.11	0.45	0.20	-0.44
Mean pause duration	4.17	0.85	3.59	1.12	1.79	3.80	1.29	3.32	2.78	0.71
Standardized pause rate	10.16	6.07	7.34	1.74	1.74	13.20	10.36	11.43	13.17	0.46

Table 3. Manual and automatic spoken language marker results for subjects. $^{***}\ p < 0.001$; $^{**}\ p < 0.01$; $^{*}\ p < 0.05$

vation of significant group differences for all but one marker, and the overall small changes in the automated group means compared to manual extraction. Second, several spoken language derived markers do appear to have discriminative utility when extracted from these narrative recall tests, though only one of the speech duration based markers showed significant group differences on this small data set.

Interestingly, syntactic complexity differences were larger in the delayed retelling, versus larger pause based differences in the immediate retelling. The combination of such markers should ultimately result in composite markers of even higher utility, though more data will be required to investigate this. Also, improved forced alignment models, obtained via, e.g., speaker-dependent adaptation, should improve the utility of automated pause based markers.

In summary, we have presented preliminary results of the utility of spoken language markers derived automatically from transcribed audio of neuropsychological exams. Though the size of the data set was small, we found statistically significant differences in a number of spoken language derived markers. We were able to extract these markers automatically sufficiently accurately to retain group differences.

Acknowledgments This work was supported in part by grants PHS P30-AG008017, P30-AG024978 and PHS 5 M01-RR000334; and by a grant from the Oregon Partnership for Alzheimer's Research. Thanks to Tracy Zitzelberger, Jessica Payne-Murphy and Robin Guariglia for help with the data.

5. REFERENCES

[1] M. Bacchiani, M. Riley, B. Roark, and R. Sproat. MAP adaptation of stochastic grammars. *Computer Speech and Language*, 20(1):41–68, 2006.

[2] E. Black, S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. P. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. A procedure for quantitatively comparing the syntactic coverage of english grammars. In *DARPA Speech and Natural Language Workshop*, pages 306–311, 1991.

[3] E. Charniak. A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the ACL*, pages 132–139, 2000.

[4] E. Charniak and M. Johnson. Edit detection and parsing for transcribed speech. In *Proceedings of the 2nd Conference of the North American Chapter of the ACL*, 2001.

[5] E. Charniak and M. Johnson. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the ACL*, 2005.

[6] H. Cheung and S. Kemper. Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, 13:53–76, 1992.

[7] L. Frazier. Syntactic complexity. In D. R. Dowty, L. Karttunen, and A. M. Zwicky, editors, *Natural Language Parsing*. Cambridge University Press, Cambridge, UK, 1985.

[8] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994.

[9] J. Hosom. Automatic phoneme alignment based on acoustic-phonetic modeling. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2002.

[10] S. Kemper, E. LaBarge, F. Ferraro, H. Cheung, H. Cheung, and M. Storandt. On the preservation of syntax in Alzheimer's disease. *Archives of Neurology*, 50:81–86, 1993.

[11] K. Lyons, S. Kemper, E. LaBarge, F. Ferraro, D. Balota, and M. Storandt. Oral language and Alzheimer's disease: A reduction in syntactic complexity. *Aging and Cognition*, 1(4):271–281, 1994.

[12] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[13] J. Morris. The clinical dementia rating (CDR): Current version and scoring rules. *Neurology*, 43:2412–2414, 1993.

[14] J. Morris, C. Ernesto, K. Schafer, M. Coats, S. Leon, M. Sano, L. Thal, and P. Woodbury. Clinical dementia rating training and reliability in multicenter studies: The Alzheimer's disease cooperative study experience. *Neurology*, 48(6):1508–1510, 1997.

[15] S. Rosenberg and L. Abbeduto. Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics*, 8:19–32, 1987.

[16] S. Singh, R. Bucks, and J. Cuerden. Evaluation of an objective technique for analysing temporal variables in DAT spontaneous speech. *Aphasiology*, 15(6):571–584, 2001.

[17] V. Yngve. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104:444–466, 1960.