ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Fully Automated Neuropsychological Assessment for Detecting Mild Cognitive Impairment

Maider Lehr, Emily Prud'hommeaux, Izhak Shafran and Brian Roark

Oregon Health & Science University, Portland, Oregon, USA
{maiderlehr,emilypx,zakshafran,roarkbr}@gmail.com

Abstract

We present an end-to-end system for automatically scoring spoken responses to a narrative recall test administered to seniors when screening for cognitive impairment. In Wechsler Logical Memory (WLM) test, a patient listens to a brief narrative, then retells the story once immediately and again after a brief delay. We transcribe the retellings automatically using an ASR system, align the transcripts to the source narrative, extract features that replicate the standard clinical scoring method, and then use the features for automatic assessment using a classifier. On a test corpus of 72 subjects, we empirically evaluate different ASR adaptation strategies and analyze the errors with respect to clinical assessment. Despite imperfect recognition, the system presented here yields classification accuracy comparable to that of manually assigned scores. Our results show that automatic assessment of neuropsychological tests such as the WLM is practical for screening large cohorts.

Index Terms: clinical diagnostics, classifying mild cognitive impairment

10.21437/Interspeech.2012-306

1. Introduction

One of the earliest identifiable stages of cognitive decline is Mild Cognitive Impairment (MCI). MCI is characterized by impairments in one or more domains of cognitive function that are potentially clinically significant but do not appreciably interfere with daily living activities [1]. When the memory domain is primarily affected (amnestic MCI) there is a high risk of developing Alzheimer's disease in subsequent years. The diagnosis and characterization of MCI typically relies on lengthy interviews with the patient and a family member. Subsequent clinical management requires continual assessment at regular intervals [1, 2]. Thus, an automated method for screening and characterizing MCI is highly desirable.

In this paper we propose a framework that uses speech and language technology to automatically analyze responses to a widely used neuropsychological test used to assess memory function, the Wechsler Logical Memory (WLM) subtest of the Wechsler Memory Scale [3]. In the WLM, a subject listens to the examiner read a brief narrative and then retells the narrative once immediately upon hearing the narrative and a second time after approximately 30 minutes of unrelated activities. The subject's response is graded in real time by the examiner according to how many key *story elements* are recalled in the retelling, in any order, from a list of 25 predetermined story elements. The clinical evaluation guidelines specify what lexical substitutions, if any, are allowed for each element. For some elements, subjects are given credit for recalling a variant of the target word or phrase; for example, *Ann* or *Annie* is an acceptable substitute for *Anna*. Other story elements must be recalled verbatim. Previous work has shown that poor performance on this test is associated with MCI [4], indicating that it is a particularly promising task to use for screening.

In contrast to previous work in automating the evaluation of retellings for diagnostic classification, which has relied on manual transcripts of retellings [5], our end-to-end system takes as input an audio recording of a subject retelling the WLM narrative. The recording is transcribed using an ASR system specifically adapted to this task. From this transcription, story element-based features that parallel published scoring guidelines are extracted. We use the features within a support vector machine classifier to determine whether the subject has MCI.

In the following sections, we present a brief discussion of previous work in the area of automated neuropsychological assessment, followed by an overview of our data, our ASR system, our feature extraction, and our use of those features for diagnostic classification. Although both recognition and alignment error rates in our system are relatively high, the classification results we present here are promising and comparable to results generated with manual expert evaluation. We expect that further improvements in recognition and alignment will yield even more compelling results.

2. Background

The past decade has shown increased interest in applying techniques from language and speech processing to the task of analyzing clinically elicited speech in order to identify signs of neurological disorders. Much of the work in this realm has focused on extracting automatically derived linguistic features, such as measures of syntactic complexity or language model entropy, from transcripts of recorded speech samples. This sort of analysis has been used to assess language development in children [6] and to identify MCI in elderly subjects [7].

An alternative approach is to evaluate a retelling according to how much of the content of the source narrative it contains. Dunn and colleagues [8] found that LSA-

based semantic distance measures between a retelling and the WLM source narrative correlated well with manually assigned summary scores and with independent measures of cognitive function. Hakkani-Tur and colleagues [9] used ASR to transcribe recordings of picture descriptions. The measure of unigram overlap between the transcript and a predefined list of key semantic concepts correlated well with manually assigned counts of semantic concepts.

A third option for analyzing narratives is to attempt to replicate the clinical assessment procedure for evaluating a narrative retelling, in which the story elements from the narrative are identified and tallied to create a summary score for the retelling. In previous work, we outlined our approach for automatically evaluating retellings according to the published guidelines for the WLM [10, 5]. Our techniques, which will be briefly reviewed in Sections 5 and 6, resulted in very high story element identification and diagnostic classification accuracy. The work presented in this paper, however, is the first attempt to combine automatic speech recognition of a narrative retelling in a clinical context with automated analysis and evaluation of that retelling for diagnostic screening.

3. Data

3.1. Experimental subjects

A total of 72 subjects were selected for this experiment from a large group of participants in an existing community cohort study of brain aging at Oregon Health and Science University's Layton Aging and Alzheimer's Disease Center. Of the 72, 35 had a diagnosis of Mild Cognitive Impairment and 37 were typically aging. Table 1 provides demographic information about the two groups. There were no significant between-group differences in age or years of education.

Dx	n	Mean Age	Mean Educ.
MCI	35	87.2	15.0 yr.
Non-MCI	37	87.3	15.5 yr.

Table 1: Subject demographic data.

In the work presented here, we define MCI via the Clinical Dementia Rating (CDR) scale [2]. The CDR is a composite score derived from measures of cognitive function in six domains: Memory; Orientation; Judgment and Problem Solving; Community Affairs; Home and Hobbies; and Personal Care. The CDR ranges from 0, indicating the absence of dementia, to 3, indicating severe dementia. MCI is defined as a CDR of 0.5 [1]. The CDR has high inter-rater reliability when conducted by experts [2], and it is independent of the WLM.

3.2. Clinical evaluation of the WLM

An excerpt from the WLM narrative used in this study is presented in Figure 1, with slashes indicating the boundaries between the story elements. Figure 2 shows an example retelling from our data. This retelling received a score of 12, with one point for each of the 12 of the 25 total story elements recalled: *Anna, Boston, employed as a cook, and robbed of, she had four, small children, reported, station, touched by he woman's story, took up a collection* and *for her*.

> Anna / Thompson / of South / Boston / employed / as a cook / in a school / cafeteria / reported / at the police / station / that she had been held up / on State Street / the night before / and robbed of / fifty-six dollars. . . .

Figure 1: An excerpt of WLM and its story elements.

> Ann Taylor worked in Boston as a cook. And she was robbed of sixty-seven dollars. Is that right? And she had four children and reported at the some kind of station. The fellow was sympathetic and made a collection for her so that she can feed the children.

Figure 2: Sample retelling of the WLM narrative.

For some story elements, the guidelines allow lexical substitutions; for example, *employed* is considered correctly recalled as long as the subject provides an indication that she had a job. Other story elements, such as *Thompson* and *cafeteria* must be recalled verbatim.

3.3. Speech Corpus

The WLM was administered to each of the experimental subjects as part of an interview and set of structured activities designed to elicit responses that can be used to assess cognitive function. The audio of the session for each subject was recorded using a stationary microphone attached to either a laptop or a digital recorder. From these recordings, the segments corresponding to the two full WLM retellings, typically ranging between 30 and 60 seconds, were extracted for each subject. The recordings were sometimes made in an informal setting, such as the subject's home or a senior center. For this reason, there are often extraneous noises in the recordings such as music and footsteps. Although this presents a challenge for ASR, part of the goal of our work is to demonstrate the robustness of our methods to noisy audio.

Our system was evaluated on the spoken retellings from 72 subjects. In addition, the recordings of two retellings from 91 other participants in the existing community cohort study, 15 with MCI and 76 typically aging, about 2 hours of speech in all, were held out for adapting acoustic models on this domain.

3.4. Manual Transcripts

Each experimental subject's two retellings were transcribed according to commonly used utterance segmentation rules for conversational speech [11]. These transcriptions enable us to evaluate the ASR output, as discussed in Section 4. In addition to the transcriptions of the retellings from the 72 experimental subjects, we have manual transcriptions from 91 other participants in the

same existing community cohort study. These retellings are used to adapt the acoustic model for the ASR system, as described in Section 4, below.

4. Automatic Transcription

Since we have limited number of transcripts to be able to train a robust in-domain speech recognizer, we need to adapt a baseline recognizer trained from a publicly available corpus. Based on performance in initial experiments with Switchboard and Broadcast News systems, we picked Broadcast News system as a baseline recognizer, which is modeled after [12]. Briefly, the acoustics of speech are modeled by 4000 clustered allophone states defined over a pentaphone context, where states are represented by a Gaussian mixture models with a total of 150K mixture components. The observation vectors consists of PLP features, stacked from 10 neighboring frames and projected to a 50-dimension space using a single semi-tied covariance. The acoustic models were trained on 430 hours of transcribed speech from Broadcast News corpus (LDC97S44, LDC98S71). The language model is defined over 84K vocabulary and consists of about 1.8M, 1M and and 331K bigrams, trigrams and 4-grams, estimated from standard Broadcast news corpus. The decoding is performed in three stages using three successively refined acoustic models – a context-dependent model, a vocal-tract normalized model and a speaker-adapted MLLR model. The system gives a word error rate of 21.6% on the 2004 Rich Transcription benchmark by NIST [13], which is comparable to state-of-the-art for equivalent amounts of acoustic training data.

The spoken retellings were decoded in three different modes to gauge the impact of adaptation and WER on the automated scoring. In the baseline mode, we decoded the utterances using the three stages of the baseline recognizer. In the unsupervised mode, we decoded the utterances after adapting the baseline acoustic models on this clinical domain using MLLR [14] transforms estimated on the 2-hour held-out adaptation data. The transforms are estimated using the automatic transcription from the final stage of the baseline system. In the supervised mode, we adapt the baseline acoustic models, but now estimating the transforms using the accompanying manual transcriptions. Empirically, we found 600 and 500 transforms optimal for the unsupervised and the supervised modes. The word error rate in the automatically generated transcriptions from the three models on the test set are reported in Table 2. Systematic errors could confound subsequent analysis and to tease them apart, we report the error rate incurred on MCI and control subjects separately. The accuracy on words related to story elements have a greater impact on subsequent automated analysis and they are reported in parentheses.

Surprisingly, the ASR systems commit significantly more errors on recognizing speech from the MCI subjects than from the control, in all three modes. At a first glance, the supervised mode appears to have consistently higher error rate than the unsupervised mode, belying

Systems	Total	Control	MCI
Baseline	47.5 (46.8)	45.0 (43.4)	50.6 (53.3)
Unsupervised	39.8 (34.0)	36.1 (30.0)	44.3 (41.4)
Supervised	41.7 (27.3)	37.5 (23.4)	47.0 (34.5)

Table 2: Comparison of the average WER using three acoustic models. Story elements specific WER are shown in parentheses.

prior expectations. Closer examination revealed that in the supervised mode more words spoken by the examiner were recognized which were often absent in the manual transcripts and hence the reference, resulting in a higher insertion error rate. This is clear from the error rate computed on words related to the story elements, where supervised mode clearly gives lower error rate than the unsupervised mode. Probing further, we observed that domain adaptation of acoustic model helped almost all subjects in the control group and relatively fewer subjects in the MCI group.

5. Story Element Feature Extraction

In our earlier work [5, 10], we outlined an alignment-based method for extracting the recalled story elements from a retelling, which we applied to manual transcripts. Here, we apply this same method to ASR-derived transcripts. We use the Berkeley aligner [15], trained on a source-to-retelling and retelling-to-retelling parallel corpus from a larger group of held-out study participants, to derive word-level alignments between each of the experimental subjects' retellings and the WLM source narrative. Using these pairwise alignments and the boundaries between elements defined in the WLM administration guidelines, we can determine which retelling words are matches for the story elements. We can then compare the story elements extracted in this way to the elements manually identified by the examiner in order to evaluate the accuracy of our extraction technique. We refer the reader to our earlier papers for further details on our alignment and element extraction methods.

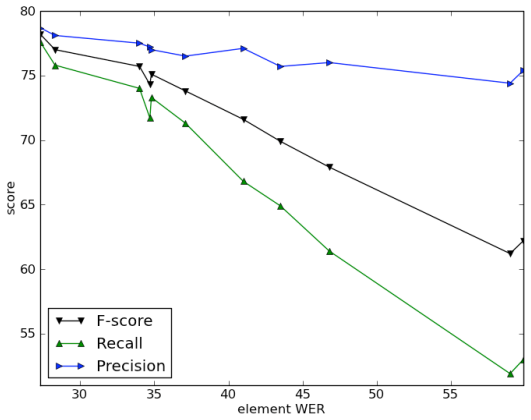

Figure 3: Story element extraction accuracy vs. element WER.

As shown in figure 3, we find that as WER improves, the accuracy of element extraction in terms of precision,

recall, and f-measure correspondingly improves, underscoring the importance of accurate ASR output. The fully automated method of extracting story elements is highly accurate, which bodes well for diagnostic classification.

The story elements are used as features for diagnostic classification, as follows. From the list of story elements recalled in each retelling, we generate a set of 25 binary features, with one feature for each of the 25 WLM story elements having a value of 1 if the story element was recalled and 0 otherwise. Each subject is therefore associated with a feature vector of length 50, containing 25 story element features for the immediate retelling and 25 story element features for the delayed retelling.

6. Automatic Classification

In order to compare the diagnostic sensitivity of the ASR-derived element features to that of the manually assigned story element features, we build a support vector machine (SVM) using the LibSVM [16] extension to the WEKA data mining Java API [17]. The SVM was trained on manually extracted story element feature vectors from the held-out subjects previous described. We test the SVM on the story element feature vectors extracted from the ASR output from the three acoustic models described in Section 4. Table 3 shows the classification accuracy for

Systems	AUC	Element WER
Baseline	75.4	43.4
Unsupervised	77.7	30.0
Supervised	80.9	23.4
Manual Transcripts	81.5	n/a

Table 3: Comparison of the performance of different ASR acoustic models on MCI classification and element WER.

MCI as measured by the area under the receiver operating characteristic curve (AUC) for each of the three acoustic models, along with the WER on words related to the story elements. As ASR quality improves, classification accuracy also improves. In summary, we find that the ASR-derived features yield classification accuracy comparable to that of manually-derived features.

7. Conclusions

The work presented here demonstrates the efficacy of our end-to-end system for automatic diagnostic screening for MCI. Our results show that the acoustic models trained on the publicly available Broadcast News corpus can be adapted to recover a majority of the semantic concepts that are essential for automatically scoring the retellings reliably. Surprisingly, unsupervised adaptation of acoustic models bridges a significant portion of the gap between the out-of-domain baseline model and models adapted with supervision. In future studies on related clinical tasks, effort may be better spend on collecting retellings from more subjects than manually transcribing fewer retellings for the sake of supervised adaptation.

8. Acknowledgements

This research was supported in part by NIH awards 5K25AG033723-02 and P30 AG024978-05 and NSF awards 1027834, 0958585, 0905095 and 0826654. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not reflect the views of the NIH or NSF. We thank Brian Kingsbury and IBM for making their ASR software tools available to us. We are grateful to Jeffrey Kaye and Diane Howeison for their valuable input. We thank the clinical team at Oregon Center for Aging and Technology for their meticulous care and effort in collecting the data.

9. References

[1] K. Ritchie and J. Touchon, "Mild cognitive impairment: Conceptual basis and current nosological status," *Lancet*, vol. 355, pp. 225–228, 2000.

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

[3] D. Wechsler, *Wechsler Memory Scale - Third Edition Manual*. The Psychological Corporation, 1997.

[4] A. Nordlund, S. Rolstad, P. Hellstrom, M. Sjogren, S. Hansen, and A. Wallin, "The goteborg mci study: mild cognitive impairment is a heterogeneous condition," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 76, pp. 1485–1490, 2005.

[5] E. T. Prud'hommeaux and B. Roark, "Alignment of spoken narratives for automated neuropsychological assessment," in *Proceedings of ASRU*, 2011.

[6] K. Sagae, A. Lavie, and B. MacWhinney, "Automatic measurement of syntactic development in child language," in *Proceedings of ACL*, 2005, pp. 197–204.

[7] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2081–2090, 2011.

[8] J. C. Dunn, O. P. Almeida, L. Barclay, A. Waterreus, and L. Flicker, "Latent semantic analysis: A new method to measure prose recall," *Journal of Clinical and Experimental Neuropsychology*, vol. 24, no. 1, pp. 26–35, 2002.

[9] D. Hakkani-Tur, D. Vergyri, and G. Tur, "Speech-based automated cognitive status assessment," in *Proceedings of Interspeech*, 2010, pp. 258–261.

[10] E. T. Prud'hommeaux and B. Roark, "Extraction of narrative recall patterns for neuropsychological assessment," in *Proceedings of Interspeech*, 2011.

[11] S. Strassel, *Simple metadata annotation specification v6.2*. Linguistic Data Consortium, 2004.

[12] B. Kingsbury, H. Soltau, G. Saon, S. M. Chu, H.-K. Kuo, L. Mangu, S. V. Ravuri, N. Morgan, and A. Janin, "The ibm 2009 gale arabic speech transcription system," in *ICASSP*, 2011, pp. 4672–4675.

[13] J. Fiscus, J. Garofolo, A. Le, A. Martin, greg Sanders, M. Przybocki, and D. Pallett, "2004 spring nist rich transcription (rt-04s) evaluation data," 2007.

[14] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hmm," *Computer Speech and Language*, pp. 9:171–185, 1995.

[15] P. Liang, B. Taskar, and D. Klein, "Alignment by agreement," in *Proceedings of HLT NAACL*, 2006.

[16] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 27, pp. 1–27, 2011.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.