Role-specific Language Models for Processing Recorded Neuropsychological Exams

Tuka Alhanai[†]**, Rhoda Au**[‡§]**, and James Glass**[†]

[†]Computer Science and Artificial Intelligence Laboratory,
Massachusetts Institute of Technology, Cambridge MA, USA
[‡]Departments of Anatomy & Neurobiology, Neurology, and Epidemiology,
Boston University School of Medicine and Public Health, Boston MA, USA
[§]The Framingham Heart Study, Framingham MA, USA
tuka@mit.edu, rhodaau@bu.edu, glass@mit.edu

Abstract

Neuropsychological examinations are an important screening tool for the presence of cognitive conditions (e.g. Alzheimer's, Parkinson's Disease), and require a trained tester to conduct the exam through spoken interactions with the subject. While audio is relatively easy to record, it remains a challenge to automatically diarize (who spoke when?), decode (what did they say?), and assess a subject's cognitive health. This paper demonstrates a method to determine the cognitive health (impaired or not) of 92 subjects, from audio that was diarized using an automatic speech recognition system trained on TED talks and on the structured language used by testers and subjects. Using leave-one-out cross validation and logistic regression modeling we show that even with noisily decoded data (81% WER) we can still perform accurate enough diarization (0.02 % confusion rate) to determine the cognitive state of a subject (0.76 AUC).

1 Introduction

Cognitive impairment is a decline in mental abilities that is severe enough to interfere with daily life (Nussbaum and Ellis, 2003). Such conditions are particularly debilitating, with costs of up to $200 billion in the USA alone (Prince et al., 2011; Leifer, 2003; Alzheimers, 2015), and come second only to spinal-cord injuries and terminal cancer in the severity of their effects (Organization, 2003; Ferri et al., 2006).

Several methods exist to screen for cognitive conditions (e.g. Alzheimer's, Parkinson's), ranging from laboratory measures to brain imaging scans (Quadri et al., 2004; Van Himbergen et al., 2012), with the baseline being set by neuropsychological examinations. These exams are composed of multiple components that measure a specific domain of cognition such as: thinking, recall,

speech, and physical movement. Each exam component is assigned a score by the tester according to the established rubric. While this exam can be comprehensive, there is an additional dimension of information that can be passively recorded - the audio of the spoken interactions. Utilizing such data would allow for the identification of spoken language biomarkers of cognitive impairment.

However, with richer information comes additional complexity (Fitch et al., 2016). The application of automatic speech processing technologies to medical domains requires a pipeline with multiple stages. Such a system requires audio pre-processing to locate speech and speaker segments (i.e. diarization) (Anguera et al., 2012), the transcription of spoken utterances (Besacier et al., 2014), and feature representation and modeling of the speaker's latent condition to determine disease biomarkers for classification purposes (Cummins et al., 2015).

Research in this domain can be categorized into two areas. First is the utilization of acoustic and linguistic information to perform speaker diarization and verification using standard corpora (e.g. Switchboard, NIST) (Stolcke et al., 2006; Reynolds et al., 2003). The second category of work seeks to evaluate speech and language biomarkers for the detection of cognitive impairment utilizing measures such as speaking rate, pauses, n-grams, and Word Error Rates (WERs) (Pakhomov et al., 2010; Lehr et al., 2012; Fraser et al., 2014; Pakhomov and Hemmy, 2014; Vincze et al., 2016), as well as Automatic Speech Recognition (ASR) for phonetic alignment and acoustic feature extraction (Tóth et al., 2015). However, systems from the speech community are developed using well-curated data with healthy speakers, while the clinical community develops systems using manually transcribed data, with some exceptions (Tóth et al., 2015; Weiner et al., 2016).

746

Our paper seeks to bridge the two areas by automating data curation for clinical use.

We hypothesize that it is possible to automate data curation for clinical use by conditioning on speaker roles, because speakers (subject/tester) during neuropsychological exams have different word usage and speaking patterns due to the question and answer nature of the evaluation. We also hypothesize that not all segments of the exam will be equally valuable in evaluating for cognitive conditions, due to potential confusion between speakers when automatically annotating speaker segments, polluting the features used for modeling cognitive conditions.

Our study differentiates itself from prior work by combining speaker-specific language modeling and ASR for speaker diarization, with the ultimate goal of assessing the cognitive condition of the subjects using the acoustic information contained in the hypothesized (and less than ideal) segments. This is an extension of work by Alhanai *et al.* that used gold standard speaker segmentations and transcriptions to evaluate cognitive outcomes. Further details on feature selection, modeling, and the relation to previous work in that domain are described in (Alhanai et al., 2017). This approach captures real-world scenarios where automatically diarized and transcribed data may not be at human parity but its usage is necessary for deploying screening technologies at scale. Moreover, audio recordings are often sub-optimal, using digital recorders on a desk, which is the case of the data used in this study. Therefore the ability to detect cognitive conditions must accommodate the presence of noisy data, of which we sought to evaluate.

1.1 Objectives

Our objectives were to (1) automatically extract and identify segments of speech that were most likely to belong to the subject, and (2) to evaluate the type of segments that were most predictive of a subject's cognitive condition.

2 Methods

2.1 Data

The data used in this work was collected from the Framingham Heart Study, an on-going longitudinal population study of 15,447 subjects from 1948 to the present (Mahmood et al., 2014). Since 1999 a subset of subjects have undergone neuropsycho-

logical examinations (Satizabal et al., 2016), and as of 2005, it became standard to record audio of these examinations. The neuropsychological examinations include multiple components to assess memory, attention, executive function, language, reasoning, visuoperceptual skills, and premorbid intelligence. All participants provided written informed consent, with study protocols and consent forms approved by the institutional review board at the Boston University Medical Center.

Our study used 92 mono-channel audio recordings of neuropsychological examinations that had available text transcripts. The exams were composed of several tests measuring memory, recall, logical and thinking. Further details and a full example are found in (Satizabal et al., 2016). The recordings were on average, 65 minutes in duration, contained 2,496 words, with a vocabulary size of 527 words.

Transcripts for each audio file were generated manually. Transcribers were instructed to include timestamps for each speaker turn (subject/tester), indicate who spoke when, transcribe speech orthographically (e.g. nineteen dollars instead of $19), include tags to highlight moments such as filled pauses (<um>), and to insert punctuation.

2.2 Outcome of Interest

Our overarching goal was to determine whether the subject being evaluated was cognitively impaired, but we also needed to determine who spoke when (subject or tester). To this end, we modeled two levels of outcomes. Our first outcome of interest was a binary indicator of the speaker type (subject or tester), with the subject coded as 1.

Our second outcome of interest was a binary indicator of cognitive impairment, with impairment coded as 1. We labeled subjects as cognitively impaired if the date of impairment (as concluded by the dementia diagnostic review panel (Seshadri et al., 2006) was on or before the date of the neuropsychological examination where the audio recording took place. Using this criteria, 21 subjects (22.8%) were cognitively impaired. Ten of these subjects had a severity rating less than mild, six were mild, five were moderate, and none were severe . Fourteen subject were diagnosed as having Alzheimer's disease using the NINCDS-ADRDA criteria (McKhann et al., 2011), and five were diagnosed with Vascular dementia based on the NINCDS-AIRENS criteria (Román et al.,

1993).

2.3 Model Choice and Evaluation Metrics

To evaluate speaker diarization we used the Diarization Error Rate (DER) metric, as well as the percentage of speech classified as non-speech (Miss), the percentage of non-speech classified as speech (False Alarm), and the percentage of speech misclassified as belonging to the other speaker (Confusion Rate) (Tranter and Reynolds, 2006). We used a time-based diarization approach, ignoring segments less than 250ms in duration. To evaluate the performance of the ASR system we used the Word Error Rate (WER) metric. Given the importance of model interpretability for detecting spoken language biomarkers, logistic regression was chosen as our modeling framework. The evaluation metrics we used for detecting cognitive impairment was the Area Under the Receiver Operating Characteristic Curve (AUC) which has the advantage of evaluating model performance across the whole range of probability cutoffs, rather than a single point estimate such as accuracy or F1 score (Huang and Ling, 2005). To assess the generalizability and robustness of our modeling techniques, we performed leave-one-out cross-validation.

3 Experiment 1: Speaker ID from Text

We first investigated the language patterns of speakers to determine whether a subject or tester was speaking (i.e., a 2 class problem). We started with the segmentation from the speaker turns labeled in the transcripts. We trained a trigram language model with Knesser Ney discounting for each speaker type. The language models were then used to generate the language perplexity of the spoken (text) segment. The training and testing was performed with leave-one-out validation (i.e. 92 folds, one fold for each of the 92 subject-tester interactions). Six features were used in the logistic regression model:

- **OOV-rate** (x2): The Out-of-Vocabulary rate of the subjects' and testers' vocabulary (from their respective training sets).

- **Perplexity** (x2): The language model perplexity for the subjects and testers.

- **Perplexity sans** <s> (x2): The language model perplexity for the subjects and testers,

excluding the start and end of sentence tags (<s>,</s>).

This resulted in a classification accuracy of 84% (± 0.06), and an AUC of 0.93 (± 0.07) These results motivated further investigation into classifying speakers from the audio directly.

4 Experiment 2: Speaker ID from ASR

For this experiment, we decoded the audio using an Automatic Speech Recognition (ASR) system with a language model trained on each speaker (subject/tester), and an acoustic model trained on the TEDLIUM corpus. Each component of the ASR system was developed as follows:

- **Acoustic Model:** The TEDLIUM corpus contains over 1,400 audio recordings and text transcription of TED talks, for a total of 120 hours of data and 1.7M words (Rousseau et al., 2012). Using this corpus, we trained the acoustic model as a feedforward Neural Network (6 layers x 2048 hidden units) with the Minimum Bayes Risk (MBR) criterion using 40 mel filterbank features, via the Kaldi speech recognition toolkit using the 's5' TEDLIUM recipe (Povey et al., 2011; Rousseau et al., 2012).

- **Language Model:** A tri-gram language model was trained for each of the speaker and tester using the SRILM toolkit (Stolcke et al., 2002).

- **Lexicon:** We generated the word pronunciations using the LOGIOS lexical tool[1].

We decoded the audio in three ways:

1. **Oracle**: A language model was trained across all 92 transcripts, and utterances were segmented according to manually generated speaker turns.

2. **Leave-one-out**: A language model was trained on all transcripts *excluding* the transcript of the audio being decoded. Utterances were segmented according to manually generated speaker turns.

3. **Leave-one-out + automatic segmentation**: A language model was trained on all transcripts *excluding* the transcript of the audio

---

[1] http://www.speech.cs.cmu.edu/tools/lextool.html

being decoded. Utterances were *not* segmented by speaker turns, the full audio was decoded as a single segment.

The results are displayed in Table 4. Our Oracle system performed with a WER of 66.7%, while decoding without language modeling information (of the audio being decoded) resulted in a WER of 68.6%. This relatively small difference in performance (68.6% vs. 66.7%) indicated that the language usage across the audio recordings was consistent.

We also compared the Diariziation Error Rate (DER) across the different setups (Table 4). This helped us evaluate how well a speaker could be identified given various levels of information about the underlying segments being decoded.

ALL SEGMENTS			
	Oracle	**Loocv**	**Loocv auto seg.**
WER (%)	66.7	68.6	81.3
DER (%)	35.8 ($\pm$ 5.9)	37.2 ($\pm$ 5.5)	40.5 ($\pm$ 05.4)
Miss	00.2 ($\pm$ 0.4)	00.2 ($\pm$ 0.4)	00.2 ($\pm$ 15.3)
False Alarm	03.9 ($\pm$ 1.2)	04.1 ($\pm$ 1.3)	03.7 ($\pm$ 01.3)
Confusion	31.7 ($\pm$ 5.9)	32.9 ($\pm$ 5.5)	36.7 ($\pm$ 05.3)
Cognitive ID			
AUC	0.72	0.70	0.68

OPTIMUM SEGMENTS		
	95% subj. & 10+ words	**Top 9 longest**
DER (%)	-	98.2 ($\pm$ 1.6)	99.9 ($\pm$ 0.2)
Miss	-	97.9 ($\pm$ 2.1)	99.9 ($\pm$ 0.4)
False Alarm	-	00.0 ($\pm$ 0.0)	00.0 ($\pm$ 0.0)
Confusion	-	00.3 ($\pm$ 0.6)	0.02 ($\pm$ 0.2)
Cognitive ID		
AUC	-	0.75	0.76

Table 1: ASR, Speaker ID, and Cognitive ID

5 Experiment 3: Cognitive ID

Using the classified speaker segments, we were interested in determining the subject's cognitive condition (impaired or not). We modeled each segment using logistic regression and 220 acoustic features capturing prosody (pitch, zero-crossing rate, jitter, harmonic-to-noise ration) and energy in the speech (energy, spectral energy, shimmer). Full details on the acoustic feature set, and method for extraction can be found in (Alhanai et al., 2017). To calculate model performance, we took the mean predicted probability across all segments as a single value representing the probability of a subject's cognitive impairment. For this

experiment, we performed leave-one-out cross-validation.

5.0.1 Speaker Turn Segmentations

For the experimental setup that used segmentations by speaker turn, we modeled cognitive impairment within a grid search space along two dimensions: (a) the total number of words that were decoded, and (b) by the percentage of words decoded that were hypothesized to belong to the subject. The results of evaluating cognitive impairment in this search space can be viewed in Figure 1. The highest AUC (of 0.75) was found when modeling with segments that had been decoded with at least 10 words, and 95% of which were hypothesized to belong to the subject.
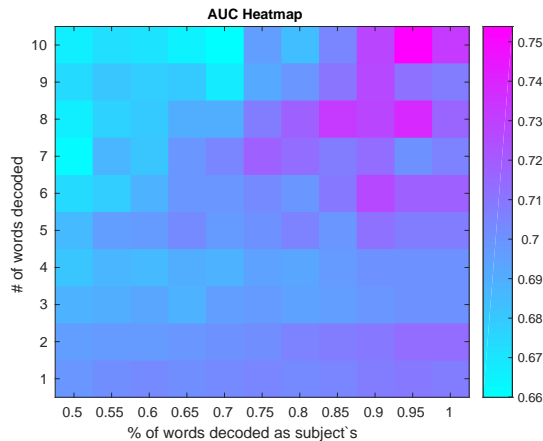

Figure 1: Cognitive ID Heatmap. Heatmap of Subject AUC across two thresholds, (y-axis) minimum number of words decoded, and (x-axis) percentage of words in a segment classified as the subject's.

5.0.2 Discarding Speaker Segmentations

For the experimental setup that was decoded without oracle speaker turn segmentation, we first segmented the decoded hypothesis along silences that were longer than 1.5 seconds, and then segmented according to the hypothesized speaker. For modeling, we selected the top N longest segments that were hypothesized to be the subject's, where N was evaluated from 1 to 15. 99% of segments hypothesized were under 25 seconds in duration, and as a pre-processing step we discarded the longest 1% of hypothesized segments, which were many minutes long and several standard deviations beyond the mean (i.e. spurious decodings). The highest AUC (of 0.76) was found when modeling the 9 longest segments hypothesized as the sub-

ject's. This was an average of 150 seconds (±20 sec) of audio per subject, or 7% of a subject's total audio duration.
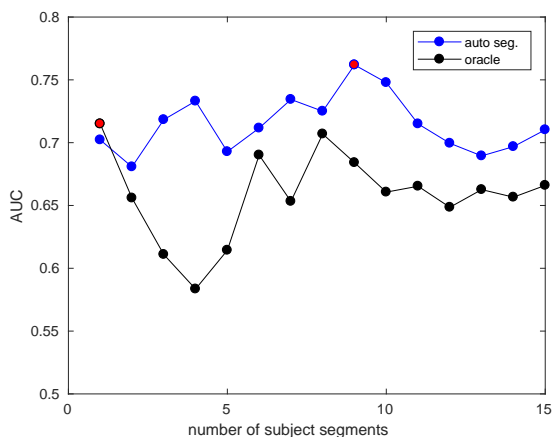

Figure 2: Cognitive ID by Number of Segments. Plot of AUC (y-axis) with respect to the number of segments per-subject (by descending order of length) used for modeling their cognitive impairment (x-axis). Red points indicate best performance (AUC 0.72 and 0.75 for oracle and automatic segmentation systems respectively).

6 Discussion

Utilizing audio recordings of spoken interactions between subjects and tester, the work in this paper sought to: (1) automatically extract and identify segments of speech that were most likely to belong to the subject, and (2) to evaluate the type of segments that were most predictive of a subject's cognitive condition.

6.1 Experiment 1: Speaker ID from Text

Our results from the first experiment showed that language usage between the subject and tester differed significantly, and that each speaker's language style was consistent across recordings (i.e. subjects consistently spoke like other subjects, and testers consistently spoke like other testers). Therefore, with the availability of highly accurate transcriptions of the same structure (neuropsychological exams), a highly accurate text-based speaker diarization can be conducted.

6.2 Experiment 2: Speaker ID from ASR

Our second set of experiments validated the observation from the previous experiment on language usage patterns across speaker roles (i.e. subjects

consistently spoke like other subjects, testers consistently spoke like other testers, and subjects and testers did not speak like each other). Also, seemingly high WERs (between 66.7% and 81.3%) still contained information that was robust enough for further usage in diarization and modeling of cognitive impairment.

6.3 Experiment 3: Cognitive ID

Our last experiment showed that it was possible to perform modeling of cognitive impairment utilizing automatically segmented subject speaker turns that was on par with the oracle speaker segmentation, and that 9 segments was sufficient for evaluation. As shown in Figure 2, we also found that not all diarization was equal, nor were all segment lengths equally powerful at modeling subjects' cognitive state. In the case where no oracle segmentation was available, and automatic segmentation was utilized, longer segments contained information that was more discriminative (AUC 0.68 vs. 0.76). For the oracle system, the longest system was the most and equally predictive of cognitive impairment, as all segments taken together. This highlights that tests that elicited longer responses allowed for more robust diarization, were evaluating cognitive performance that was (via speech) most strongly associated with the outcome, and/or that longer spoken segments provided more opportunity to capture patterns associated with cognitive impairment.

Furthermore, the modeling paradigm we explored was robust enough that neither the underlying neuropsychological test need be explicitly modeled (Lehr et al., 2012), nor do the features utilized require word or phone alignments (alignments which require accurate transcriptions in order to generate) (Tóth et al., 2015).

References

Tuka Alhanai, Rhoda Au, and James Glass. 2017. Spoken language biomarkers for detecting cognitive impairment.

Association Alzheimers. 2015. 2015 alzheimer's disease facts and figures. *Alzheimer's & dementia: the journal of the Alzheimer's Association* 11(3):332.

Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. 2012. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing* 20(2):356–370.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication* 56:85–100.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication* 71:10–49.

Cleusa P Ferri, Martin Prince, Carol Brayne, Henry Brodaty, Laura Fratiglioni, Mary Ganguli, Kathleen Hall, Kazuo Hasegawa, Hugh Hendrie, Yueqin Huang, et al. 2006. Global prevalence of dementia: a delphi consensus study. *The lancet* 366(9503):2112–2117.

W Tecumseh Fitch, Bart de Boer, Neil Mathur, and Asif A Ghazanfar. 2016. Monkey vocal tracts are speech-ready. *Science advances* 2(12):e1600723.

Kathleen C Fraser, Jed A Meltzer, Naida L Graham, Carol Leonard, Graeme Hirst, Sandra E Black, and Elizabeth Rochon. 2014. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *cortex* 55:43–60.

Jin Huang and Charles X Ling. 2005. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering* 17(3):299–310.

Maider Lehr, Emily Prud'hommeaux, Izhak Shafran, and Brian Roark. 2012. Fully automated neuropsychological assessment for detecting mild cognitive impairment. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Bennett P Leifer. 2003. Early diagnosis of alzheimer's disease: clinical and economic benefits. *Journal of the American Geriatrics Society* 51(5s2).

Syed S Mahmood, Daniel Levy, Ramachandran S Vasan, and Thomas J Wang. 2014. The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet* 383(9921):999–1008.

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 alzheimers disease: Recommendations from the national institute on aging-alzheimers association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia* 7(3):263–269.

Robert L Nussbaum and Christopher E Ellis. 2003. Alzheimer's disease and parkinson's disease. *New England Journal of Medicine* 348(14):1356–1364.

World Health Organization. 2003. *The world health report 2003: shaping the future*. World Health Organization.

Serguei VS Pakhomov and Laura S Hemmy. 2014. A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the nun study. *Cortex* 55:97–106.

Serguei VS Pakhomov, Glenn E Smith, Susan Marino, Angela Birnbaum, Neill Graff-Radford, Richard Caselli, Bradley Boeve, and David S Knopman. 2010. A computerized technique to assess language use patterns in patients with frontotemporal dementia. *Journal of neurolinguistics* 23(2):127–144.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, EPFL-CONF-192584.

Martin Prince, Renata Bryce, and Cleusa Ferri. 2011. *World Alzheimer Report 2011: The benefits of early diagnosis and intervention*. Alzheimer's Disease International.

Pierluigi Quadri, Claudia Fragiacomo, Rita Pezzati, Enrica Zanda, Gianluigi Forloni, Mauro Tettamanti, and Ugo Lucca. 2004. Homocysteine, folate, and vitamin b-12 in mild cognitive impairment, alzheimer disease, and vascular dementia. *The American journal of clinical nutrition* 80(1):114–122.

Douglas Reynolds, Walter Andrews, Joseph Campbell, Jiri Navratil, Barbara Peskin, Andre Adami, Qin Jin, David Klusacek, Joy Abramson, Radu Mihaescu, et al. 2003. The supersid project: Exploiting high-level information for high-accuracy speaker recognition. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. IEEE, volume 4, pages IV–784.

Gustavo C Román, Thomas K Tatemichi, T Erkinjuntti, JL Cummings, JC Masdeu, JHet al Garcia, L Amaducci, J-M Orgogozo, A Brun, Aea Hofman, et al.

1993. Vascular dementia diagnostic criteria for research studies: Report of the ninds-airen international workshop. *Neurology* 43(2):250–250.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. Ted-lium: an automatic speech recognition dedicated corpus. In *LREC*. pages 125–129.

Claudia L Satizabal, Alexa S Beiser, Vincent Chouraki, Geneviève Chêne, Carole Dufouil, and Sudha Seshadri. 2016. Incidence of dementia over three decades in the framingham heart study. *New England Journal of Medicine* 374(6):523–532.

Sudha Seshadri, Alexa Beiser, Margaret Kelly-Hayes, Carlos S Kase, Rhoda Au, William B Kannel, and Philip A Wolf. 2006. The lifetime risk of stroke. *Stroke* 37(2):345–350.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2006. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Dialogue* 26(3).

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *Interspeech*. volume 2002, page 2002.

Laszló 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 *Sixteenth Annual Conference of the International Speech Communication Association*.

Sue E Tranter and Douglas A Reynolds. 2006. An overview of automatic speaker diarization systems. *IEEE Transactions on audio, speech, and language processing* 14(5):1557–1565.

Thomas M Van Himbergen, Alexa S Beiser, Masumi Ai, Sudha Seshadri, Seiko Otokozawa, Rhoda Au, Nuntakorn Thongtang, Philip A Wolf, and Ernst J Schaefer. 2012. Biomarkers for insulin resistance and inflammation and the risk for all-cause dementia and alzheimer disease: results from the framingham heart study. *Archives of neurology* 69(5):594–600.

Veronika Vincze, Gabor Gosztolya, Laszlo Toth, Ildiko Hoffmann, Greta Szatloczki, Zoltan Banreti, Magdolna Pakaski, and Janos Kalman. 2016. Detecting mild cognitive impairment by exploiting linguistic information from transcripts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. volume 2, pages 181–187.

Jochen Weiner, Christian Herff, and Tanja Schultz. 2016. Speech-based detection of alzheimer's disease in conversational german. In *INTERSPEECH*. pages 1938–1942.