Identifying and Avoiding Confusion in Dialogue with People with Alzheimer's Disease

Hamidreza Chinaei*
University of Toronto
Toronto Rehabilitation Institute

Leila Chan Currie**
University of Toronto

Andrew Danks**
University of Toronto

Hubert Lin**
University of Toronto

Tejas Mehta**
University of Toronto

Frank Rudzicz†
Toronto Rehabilitation Institute
University of Toronto

Alzheimer's disease (AD) is an increasingly prevalent cognitive disorder in which memory, language, and executive function deteriorate, usually in that order. There is a growing need to support individuals with AD and other forms of dementia in their daily lives, and our goal is to do so through speech-based interaction. Given that 33% of conversations with people with middle-stage AD involve a breakdown in communication, it is vital that automated dialogue systems be able to identify those breakdowns and, if possible, avoid them.

In this article, we discuss several linguistic features that are verbal indicators of confusion in AD (including vocabulary richness, parse tree structures, and acoustic cues) and apply several machine learning algorithms to identify dialogue-relevant confusion from speech with up to

* Department of Computer Science, University of Toronto, and Toronto Rehabilitation Institute-University Health Network.
** Department of Computer Science, University of Toronto.
† Toronto Rehabilitation Institute-University Health Network, and University of Toronto, Department of Computer Science. E-mail: frank@cs.toronto.edu.

Submission received: 17 November 2015; revised version received: 29 June 2016; accepted for publication: 7 November 2016.

doi:10.1162/COLI_a_00290

82% accuracy. We also learn dialogue strategies to avoid confusion in the first place, which is accomplished using a partially observable Markov decision process and which obtains accuracies (up to 96.1%) that are significantly higher than several baselines. This work represents a major step towards automated dialogue systems for individuals with dementia.

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that deteriorates cognitive, social, and executive functions. Typical effects of AD include declines in memory, executive capacity, and (crucially) linguistic ability (Cummings 2004). These declines complicate or prohibit the completion of activities of daily living, and more severe declines often require caregiver assistance. Caregivers who assist individuals with AD at home are common, but their involvement is often the precursor to placement in a long-term care facility (Gaugler et al. 2009). Approximately $100 billion are spent annually in North America on the direct and indirect care for patients with AD and other dementias, the majority of which is attributed to long-term institutional care (Ernst et al. 1997). As the population ages, the incidence of AD will double or triple, with Medicare costs alone reaching $189 billion in the United States presently (Bharucha et al. 2009).

Given the growing need to support this population, we are designing software that engages in two-way speech communication with individuals with dementia for two purposes: 1) to help guide that individual towards the completion of daily household tasks (e.g., brushing one's teeth), and 2) to fulfill a social function often left empty in relative isolation. Our goal is to encode in software verbal techniques that caregivers use when interacting with their patients, including automatically identifying and recovering from breakdowns in communication. In this article, we show that these breakdowns can reliably be identified by surface-level linguistic features, and we propose a statistical dialogue system for avoiding these breakdowns in the first place.

2. Previous Work

Formal caregivers are often trained in responsive verbal **communication strategies** that are meant to adapt to an individual's state-of-mind (Hopper 2001; Goldfarb and Pietro 2004). These include (Small et al. 2003) but are not limited to:

1. Relatively slow rate of speech rate.
2. Verbatim repetition of misunderstood prompts.
3. Closed-ended questions (i.e., that elicit yes/no responses).
4. Simple sentences with reduced syntactic complexity.
5. Giving one question or one direction at a time.
6. Minimal use of pronouns.

These and similar strategies are often based on observational studies but are rarely grounded in *quantitative* empirical analysis. Tomoeda et al. (1990) showed that rates of speech that are too slow may interfere with comprehension if they introduce problems of short-term retention of working memory. Small, Andersen, and Kempler (1997) showed that paraphrased repetition is just as effective as verbatim repetition (indeed, syntactic variation of common semantics may assist comprehension). Furthermore,

Rochon, Waters, and Caplan (2000) showed that the syntactic complexity of received utterances is not necessarily the only predictor of comprehension in individuals with AD; rather, comprehension of the semantics of sentences is inversely related to the increasing *number of propositions* used, and therefore to the number of clauses, as well. However, there is some evidence that, at least in automated dialogue systems, neither simple confirmations nor reducing the number of options have a measurable effect on user performance, at least within specific appointment-scheduling applications (Wolters et al. 2009).

2.1 Communication Breakdown

Several theoretical models apply to communication breakdown. The trouble source-repair (TSR) model describes difficulties in communication in any dyads, including how repairs are initiated and carried out (Schegloff, Jefferson, and Sacks 1977). Difficulties can be phonological (e.g., mispronunciation), morphological/syntactic (e.g., incorrect agreement among constituents), semantic (e.g., disturbances related to lexical access), and discourse-based (i.e., misunderstanding of shared knowledge, or cohesion) (Orange, Lubinsky, and Higginbotham 1996). The majority of TSR sequences involve self-correction (e.g., by repetition, elaboration, or reduction of a troublesome utterance).

Orange, Lubinsky, and Higginbotham (1996) showed that whereas 18% of non-AD dyad utterances involve TSR, 23.6% of early-stage AD dyads and 33% of middle-stage AD dyads involved TSR. Of these, individuals with middle-stage AD exhibited more discourse-related difficulties including inattention, failure to track propositions and thematic information, and deficits in working memory. Contrary to TSR, the most common repair initiators and repairs given communication breakdown involved frequent *wh*-questions and hypotheses (e.g., "*Do you mean...?*").

An alternative, closely related, paradigm for measuring communication breakdown is trouble-indicating behavior (TIB) involving implicit or explicit requests for aid. In a study of 10 seniors with varying degrees of dementia, Watson (1999) showed that there was a significant difference in TIB use ($p < 0.005$) between individuals with AD and the general population. This article adopts this model, which incorporates 12 distinct types of TIB (examples are derived from our own annotations of the DementiaBank corpus, described in Section 3):

1. **Neutral or non-specific requests for repetition (local)**. Minimal queries indicating non-understanding, which did not identify the problem specifically. *E.g., What? Huh?*
2. **Request for confirmation – repetition with reduction**. Partial repair of a trouble source, often in the form of a question. *Speaker 1: just tell me everything that you see happening there.... Speaker 2: uh these are things that are happening ?*
3. **Request for confirmation – complete repetition**. Recapitulatory "echo" questions, often with pronoun alternation. *Speaker 1: You tell me everything you see. Speaker 2: I'm to tell everything I see ?*
4. **Request for confirmation – repetition with elaboration**. Same as TIB 3, but with the inclusion of additional semantic content. *Speaker 1: Tell me everything that you see happening in this picture. Speaker 2: Uh do do you want like the window's open that sort of thing?*

5. **Request for specific information**. Contains a specific semantic concept, content word, or referent to the previous or recent turn. *Speaker 1: Did I say about some water?*

6. **Request for more information**. A non-specific request (i.e., without direct mention of semantic concepts in a recent utterance). *I don't understand. Tell me more. What do you mean?*

7. **Corrections**. Are the result of a violation in the quality of message or message inaccuracies. Here, semantic confusion often originates from the individual not indicating the TIB. *Speaker 1: She dropped a dish no, no she didn't drop a dish.*

8. **Lack of uptake / lack of continuation**. Verbal behaviors including 1) minimal feedback where back channel responses indicate non-understanding or lack of contribution to or elaboration on topic extension; 2) overriding where a participant does not allow the floor; and 3) topic switch where one participant abruptly changes topic. *Speaker 1: Do you know what indexed means? Speaker 2: Yes. Speaker 1: What? Speaker 2: Oh, it's a bit too hard, bit late too late to...*

9. **Hypothesis formation**. Guessing behaviors involving supplying words or speaking for or on behalf of the other participant. This does not include hypotheses in the form of rhetorical questions (which are instead categorized as TIB 5). *Speaker 1: We went to the farm. Speaker 2: You went to Riverdale Farm.; Speaker 1: we forgot to turn off the spigot.*

10. **Metalinguistic comment**. This includes "talk about talk" which explicitly refer to non-understanding of message, the interpersonal manner in which the message was conveyed, or the production of the message. *I can't remember. I don't understand. I guess there's more things I'm supposed to see.*

11. **Reprise / minimal dysfluency**. Reprises with partial or whole repetition (or revision) of the message. Minimal dysfluencies indicate difficulties producing a message that involve sound, syllable and word repetition, pauses, and fillers. These are deemed more excessive than the typical dysfluencies that occur in typical speech. *the children are wearing um completely outfitted.*

12. **Request for repetition–global**. Minimal queries indicating non-understanding. *Wait – go back to the part about... You just lost me.*

In previous work, we studied ten older adults with Alzheimer's disease as they were guided, through speech, by a robot in the task of making a cup of tea. In that work, older adults were especially likely to exhibit TIB 8 (lack of uptake), but exhibited fewer TIBs, proportionally, when interacting with the robot than with a human conversant (Rudzicz et al. 2015).

2.2 Linguistic Factors in Dementia

Although memory impairment is the main symptom of AD, language impairment can be an important marker. Faber-Langendoen et al. (1988) found that 36% of mild AD

patients and 100% of severe AD patients had aphasia, according to standard aphasia testing protocols. Ahmed et al. (2013) found that two-thirds of their participants showed subtle, but significant, changes in connected speech production up to a year before their diagnosis of probable AD. Weiner et al. (2008), in a study of 486 AD patients, reported a significant correlation between dementia severity and a number of different linguistic measures, including confrontation naming, articulation, word-finding ability, and semantic fluency.

Bucks et al. (2000) described eight linguistic measures for detecting dementia. The first three are related to vocabulary richness, which is estimated through type-token ratio (TTR), Brunét's index (BI), and Honoré's statistic (HS). These methods compute relationships between the number of unique words in an utterance versus the total number of words spoken (Bucks et al. 2000). TTR $= \frac{U}{N}$ is simply the ratio of the total number of unique words in some dialogue, U, to the total word count, N, and Brunét's index is:

$$\text{BI} = N^{U^{-0.165}} \tag{1}$$

Evidently, BI favors unique words relative to the total number of words. Bucks et al. suggest that BI values between 10 and 20 are indicative of individuals with AD during spontaneous conversation, and that smaller BI indicates greater lexical richness.

Honoré's statistic (HS) accounts for words that are only used once, namely, *hapax legomena*, denoted by N_1, so as to give the number of these words a stronger factor in measuring lexical richness. It assumes that the greater number of words used just once, the richer the lexicon.

$$\text{HS} = \frac{100 \log N}{1 - \frac{N_1}{U}} \tag{2}$$

Other indicative measures include noun rate, pronoun rate, verb rate, and adjective rate, per 100 words, which have been used in studies of aphasia and speech in affective disorders (Bucks et al. 2000; Fraser, Rudzicz, and Rochon 2013). The noun rate is associated with word-finding difficulties, and the pronoun rate measures the frequency of indirect referencing, which is higher for individuals who commonly forget names. Also, Bucks et al. (2000) suggested that the adjective rate measures the quality and expressiveness of speech and that the verb rate is related to the fluency of speech. Additional measures are described in Section 4.

3. Data

We use the DementiaBank (Goodglass and Kaplan 1983; Becker et al. 1994) database for this study. DementiaBank consists of narrative speech during the standard Cookie Theft picture description task from the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan 1983). In this task, an examiner shows the picture in Figure 1 to the patient, requests a description of its contents, and is permitted to periodically encourage or prompt the patient. Each speech sample was recorded and manually transcribed at the word level following the TalkBank CHAT protocol (MacWhinney 2014). Narratives were segmented into utterances and annotated with filled pauses, paraphasias, and unintelligible words. The transcriptions were tokenized to remove whitespace, punctuation, and CHAT code tokens (e.g., "[//]"). Tokens that were CHAT-coded to indicate repetition were expanded; e.g., *that's a dog [x 3]* becomes *that's a dog dog dog*

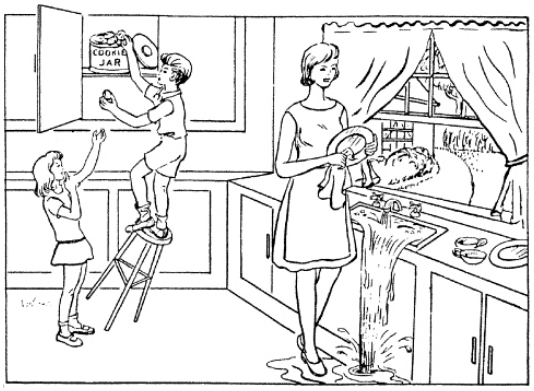


Figure 1
The standardized Cookie Theft picture (Goodglass and Kaplan 1983).

(MacWhinney 2014). The remaining tokens were lemmatized, which is effective in reducing lexical sparseness, but which also removes some information, such as verb tense, that can differentiate between control subjects and those with AD (Cortese et al. 2006).

Every participant is associated with a professionally administered Mini-Mental State Examination (MMSE) score, which provides a unified and validated scale for measuring the severity of dementia on a scale of 0 (greatest cognitive decline) to 30 (no cognitive decline), based on a series of questions in five areas: orientation, registration, attention, memory, and language (Folstein, Folstein, and McHugh 1975). English was the first language of all speakers.

Participants were assigned to the Dementia group primarily based on a history of cognitive and functional decline, and the results of the MMSE. In 1992, several years after the study had ended, the final diagnosis of each patient was reviewed on the basis of their clinical record and any additional relevant information (in 66.7% of the deceased, autopsy). From this group, we consider participants with diagnoses of 'possible' or 'probable' AD, resulting in 240 samples from 167 participants, because a few performed the description task annually (Boller and Becker 2005). We also consider control participants, resulting in 233 additional files from 97 speakers. Demographics are given in Table 1. The participants with dementia produced an average of 104.3 (s.d. 59.0) words per narrative, and the control participants produced an average of 114.4 (s.d. 59.5) words per narrative, although the distribution in both cases is somewhat right-skewed.

Table 1
Demographics of DementiaBank data.

	AD ($n = 167$)	Control ($n = 97$)
Mean age in years (s.d.)	71.8 (8.5)	65.2 (7.8)
Mean education in years (s.d.)	12.5 (2.9)	14.1 (2.4)
Sex (n male/female)	52/115	37/60
Mean MMSE score (s.d.)	18.5 (5.1)	29.1 (1.1)

Some analysis (Section 4.1.4) uses the Carolina Conversations Collection (CCC), in which transcriptions and audio of *undirected* conversations between interviewees (older adults) and interviewers (younger adults) are recorded longitudinally (Pope and Davis 2011). In our analysis, we use the data of 31 interviewees with AD (7 men) and 41 without (9 men).

Across both databases, each transcript has an associated audio file, which we convert to 16-bit mono PCM format with a sampling rate of 16 kHz. All data were annotated at the utterance-level with specific TIBs by two speech-language pathologists, with an inter-annotator Fleiss' kappa agreement of $\kappa = 0.84$. Utterances with more than one TIB were allowed, but no more than two per utterance were ever detected.

4. Features

The following subsections outline the lexical and acoustic features used across our analyses, with special attention paid to features that are meant to be task-relevant, namely, cue words of TIBs, semantic similarity across utterances, word specificity, vocabulary richness, repetition, incomplete words, and fillers.

4.1 Lexical Features

We extract 81 lexical features from the textual transcriptions of each utterance. We use Lu's L2 Syntactic Complexity Analyzer (Lu 2010), which computes 23 features that measure the syntactic complexity of text. Lu's Syntactic Complexity Analyzer was initially designed to measure the syntactic complexity of second-language writing and, in turn, uses the Stanford parser (Klein and Manning 2003). It includes measuring lengths of production units (i.e., clauses, sentences, T-units), the ratio of clauses to sentences, subordination (e.g., clauses per T-unit, dependent clauses per clause or T-unit), coordination (e.g., coordinate phrases per clause or T-unit), and particular structures (e.g., complex nominals per clause or T-unit, verb phrases per T-unit). We further compute the average sentence length within an utterance, in terms of the number of words, and the automated readability index, which measures word and sentence difficulty (Smith and Senter 1967), specifically,

$$4.71\frac{c}{w} + 0.5\frac{w}{s} - 21.43 \tag{3}$$

where c is the number of characters, w is the number of words, and s is the number of sentences. Because these measures are derived from textual transcripts of spoken utterances, the number of characters is merely an approximation for the complexity of words.

We include various part-of-speech (POS) features, such as the proportion of nouns, verbs, light verbs, adjectives, adverbs, prepositional words, demonstratives, functional words generally, and the noun-to-verb ratio, obtained using the Stanford POS tagger (Toutanova et al. 2003). We measure the mean word frequency using the SUBTL norms (Brysbaert and New 2009), and each of the Bristol norms: imageability, age-of-acquisition, and familiarity. Stadthagen-Gonzalez and Davis (2006) describe imageability as "a semantic variable that measures how easy it is for a word to elicit mental images... used to evaluate the effects of meaning on memory and word recognition" and as closely related to "concreteness," with some exceptions (Bird, Franklin, and Howard 2001); raters of imageability were asked to indicate how easily each word elicited mental images, on a 7-point scale. Familiarity is "often ... interpreted as a measure of the

frequency of exposure to a word"; ratings of familiarity were also on a 7-point scale, from words that they never had seen to those that they had seen very often (nearly every day). The Bristol norms were obtained from the ratings by 100 native English speakers across nouns, verbs, and some additional words (Stadthagen-Gonzalez and Davis 2006).

We estimate vocabulary "richness" through type-token ratio, average word length, BI, and HS, as described in Section 2.2. We also compute objectivity measures using the MPQA Subjectivity Lexicon (Wilson, Wiebe, and Hoffmann 2005), which categorizes words by their subjectivity (strong or weak) and their prior, out-of-context, polarity (positive, negative, both, or neutral). Using the subjectivity lexicon, we compute the proportion of all possible permutations (e.g., strong negative, weak negative, weak positive).

To estimate statistical deviations from typical language, we also compute the average log probability of each word in each utterance given two trigram models with Good-Turing smoothing, from the Brown and Switchboard (Godfrey, Holliman, and McDaniel 1992) corpora, respectively. Additional features are described in the following subsections.

4.1.1 Word Specificity. Word specificity is closely related to the word's position in a hypernym tree. For example, *sedan* is more specific than *car*. Because degraded cognitive function is correlated with a smaller active vocabulary, individuals with AD may be more likely to resort to more common, general words and therefore a lower degree of word specificity (Hirst and Feng 2012).

Word specificity, for nouns, can be approximated using WordNet (Miller 1995). Specifically, hypernym relationships between synonym sets form a hierarchal tree where the root (*Entity*) provides the most abstract sense. Therefore, the "word specificity" feature is the average depth of nouns in the WordNet (noun) hypernym tree, after recent work (Le 2010).

Adjectives and adverbs, unlike nouns, are not organized in hierarchies, but in bipolar scales between pairs of extremes; verbs are structured into several smaller hierarchies with different roots, which may limit the utility of computing depths of verbs in this way.

4.1.2 Semantic Similarity. Watson (1999) showed that almost half of TIBs exhibited by speakers without AD are either TIB 5 (request for specific information) or TIB 9 (hypothesis formation). These TIBs, and others, involve expressions that are semantically related to recent or preceding utterances. This is in stark contrast to how individuals with AD express trouble—over 30% of TIBs exhibited by these individuals are TIB 8 (lack of uptake) and over 65% are TIB 11 (reprise / minimal dysfluency), neither of which involves particular relations with preceding utterances. Our hypothesis is therefore that measures of semantic similarity with previous utterances will be indicative of AD.

To estimate semantic similarity between two utterances, we first use singular value decomposition (Yu 2009) given co-occurrence counts among terms and utterances in latent semantic space. We then calculate a matrix B of correlation coefficients between the semantic-space vectors for each utterance in this representation. The score of utterance t is therefore:

$$S(t) = B[t, t-1] + \frac{1}{2}B[t, t-2] + \frac{1}{4}B[t, t-3] + \ldots + \frac{1}{2^{t-1}}B[t, 1] \quad (4)$$

Our intuition is that, additionally, TIB 1 (neutral or non-specific requests for repetition (local)) and TIB 10 (metalinguistic comment) would give low similarity scores.

4.1.3 Repetition, Incomplete Words, and Fillers. In slight contrast to Section 4.1.2, repetition of a sequence of words is a common coping mechanism for brief cognitive lapses (Guinn and Habash 2012) and is increased with degradation in cognitive ability. Watson (1999) and Guinn and Habash (2012) showed that those with AD exhibited repetition twice as often as their interviewers, on average.

Incomplete words occur when uttered words are abruptly terminated before completion. This can be caused by brief lapses in cognition (Guinn and Habash 2012) or might imply that the speaker is unsure of what they are trying to communicate. These are about twice as common in people with AD relative to the general population (Guinn and Habash 2012).

Fillers are non-words or short phrases that signal cognitive lapse (e.g., *"umm," "uh"*). Single-word fillers are extracted using the *UH* tag in the Penn tagset. Near the beginning of utterances, fillers often indicate that the speaker has difficulty either comprehending what was said or forming a sentence. In the middle, they often indicate the need for a trouble source repair (Bortfeld et al. 2001). About 2.6% of words spoken by people with AD are filler phrases, compared with 1.2% of words spoken by healthy older adults (Guinn and Habash 2012).

4.1.4 Cue Words. Cue words are expressions that either link spans of discourse or signal semantic relationships in text. They are frequently used as identifying markers of the relative importance of sentences in document summarization (Teufel and Moens 1997) and in dialogue (Gravano, Hirschberg, and Beňuš 2012). We do not remove stop words here from consideration because AD patients are known to use a high proportion of function words such as pronouns, conjunctions, and articles that convey little meaning (Kempler 1995; Almor et al. 1999). Therefore, even non-content words that would typically belong to a stop list can be cues for AD.

To determine which tokens are cue words, we considered term frequency-inverse document frequency (*tf-idf*), but this was discarded empirically. Instead, we use two-tailed Welch's *t*-test to compare the mean frequency of each token in utterances with and without each TIB. The tokens were then ranked by p-value, as shown in Table 2, which includes data from the CCC. Despite differences in cue words across data sets, there is also overlap. Across both sets, the word *pardon* is a cue word for TIB 1 (neutral or non-specific request for repetition), *no* is a cue word for TIB 7 (corrections), and *know* is a cue word for TIB 10 (metalinguistic comment). This appears to be consistent with previous analysis of AD language use (Nicholas et al. 1985; Almor et al. 1999).

4.2 Acoustic Features

As in previous work (Fraser, Rudzicz, and Rochon 2013), we measure pause-to-word ratio (i.e., the ratio of speech segments to silent segments longer than 150 msec), mean and variance of fundamental frequency (F0), total duration of speech, long and short pause counts (> 0.4 msec and < 0.4 msec, respectively) (Pakhomov et al. 2010), mean pause duration, and phonation rate (the amount of the recording spent in voiced speech) (Roark et al. 2011). These acoustic measures are, generally, heuristics that estimate the difficulty with which words are recalled and produced. We also include the mean and variance for the first three formants ($F1, F2, F3$), mean instantaneous power, the mean and maximum of the first autocorrelation function, sample-level skewness and

Table 2
Up to the five most significant cue words identified for TIBs 1–6 in the CCC and in DementiaBank. All selected features have $p < 0.005$ after Bonferonni correction.

TIB	CCC	DementiaBank
1. Neutral or non-specific request for repetition (local)	*huh, pardon, what*	*hm, mention, pardon, really, anything*
2. Request for confirmation – repetition with reduction	N/A	*did, mhm, mention, hear, say*
3. Request for confirmation – complete repetition	N/A	*really, good*
4. Request for confirmation – repetition with elaboration	*said*	*missed, again, bad, why, huh*
5. Request for specific information	*ummm, what*	*no, talk, did, wanted, nope*
6. Request for more information	*why, what*	*describe, am, matter, detail, more*
7. Corrections	*ohhh, no*	*no*
8. Lack of uptake/ lack of continuation	*umm*	*please, yes, uhhuh, what*
9. Hypothesis formation	*uhoh, uhm, remember, yes*	*use, watch, my, suppose, find*
10. Metalinguistic comment	*remember, know, forget, unbelievable, disappointing*	*mentioned, know, what, must, can't*
11. Reprise/minimal dysfluency	*uh, the, um, but, because*	*k, hmm, uhhuh*
12. Request for repetition – global	N/A	N/A

kurtosis, zero-crossing rate, and mean recurrence period density entropy (which measures the periodicity of a signal, and has been applied to pathological speech generally [Little et al. 2006]). With the exception of power, which may generally be a heuristic for energy in the voice, these measures are all related to repetitiveness in the vocal signal, which have been associated with pathological speech in older voices, following our recent work (Zhao et al. 2014). Additionally, jitter and shimmer (measures of the cycle-to-cycle variations of $F0$ and amplitude, respectively, also largely applied to pathological speech) are computed (Silva, Oliveira, and Andrea 2009) as:

$$\begin{aligned} \text{jitter}(x) &= \tfrac{1}{N-1} \sum_{k=1}^{N-1} |P_0[k+1] - P_0[k]| \\ \text{shimmer}(x) &= \tfrac{1}{N-1} \sum_{k=1}^{N-1} |x[k+1] - x[k]| \end{aligned} \tag{5}$$

where $P_0(k)$ is the pitch period length ($1/F0$) at time k in a sequence x with N observations, and $x[k]$ is third-order median filtered. To this we also add the kurtosis and skewness (again, to measure repetitiveness and regularity) of each of the 12^{th}-order autocorrelation linear predictive coding coefficients of the signal, as well as the energy in the residual of this analysis (sometimes called the "gain" of the filter). We also compute the variance, mean, kurtosis, and skewness for each of the first 14 Mel-frequency cepstral coefficients (including the 0^{th} and the log energy), their velocities (δ)

and accelerations (δδ), as well as the kurtosis and skewness of their means taken over the entire utterance.

Aperiodicity and bradykinesia (slowed articulation) of speech, which are symptoms of neurodegenerative disorders such as Parkinsons disease (Bhatnagar 2002), are typically not associated with AD. Nevertheless, we perform recurrence quantification analysis of the cross recurrence (Marwan and Kurths 2002) of each utterance. We add these features because our recent work has found an acoustic component to AD (Fraser, Meltzer, and Rudzicz 2016). Specifically, for 3^{rd}-order recurrence quantification analysis with delay 2, neighborhood 0.5, we compute the mean recurrence rate, determinism, {mean, maximal, and entropy of} diagonal line length, laminarity, trapping time, maximal vertical line length, recurrence time of 1^{st} and 2^{nd} types, clustering coefficient, and transitivity over windows of length 1,000 samples (with a 500-sample window shift). In total, there are 178 acoustic features.

5. Experiment 1: Identifying TIBs

Our first goal is to automatically identify TIBs using only lexical and acoustic features described in Section 4, and machine learning. Here, we construct three processed data sets: control, dementia, and both, as follows. The nature of the DementiaBank conversations is such that the interviewer is never in a position to exhibit a TIB. Therefore, all interviewer data points are removed. Initially, we are only interested in identifying the presence of a TIB without distinguishing between the different types of TIBs, so we simply conflate all TIB-annotated utterances into one class, and all others into a second class.

A large number of features include not-a-number (NaN) values (typically when the feature represents a ratio, e.g., number of light verbs to number of verbs in the utterance "Cookie" is 0/0). To account for these values, we use k-nearest neighbor (kNN) imputation (Batista and Monard 2002). This imputation substitutes each NaN with a weighted average of the k data points closest to the datum in question. Empirically, we set $k = 3$.

Furthermore, the density of TIB-class data points is very sparse, especially within the control data set. Models built on this data have the danger of performing majority class classification. Because the TIB-class is the minority class, a majority class classification model would do extremely poorly in identifying TIBs. The imbalance in the data set can be alleviated by several means. In this experiment, we generate synthetic data points using the synthetic minority oversampling technique (SMOTE), which under-samples the majority class *and* over-samples the minority class by creating synthetic minority class examples (Chawla et al. 2002). Here, each minority class sample introduces synthetic examples randomly along line segments that join the $k = 5$ minority class nearest neighbors.

5.1 Methodology

We use multilinear logistic regression as a baseline and a mixture density network (MDN) to estimate the presence of TIB. An MDN is a multilayer perceptron with one hidden layer with a variable number of hidden units and a Gaussian mixture probability density as the output (Bishop 1994). For this experiment, the output is a single Gaussian, given that increasing this number reduced performance. We train the network for a maximum of 6,000 cycles optimized by scaled conjugate gradient and negative log likelihood as the error function. The mean of the output Gaussian is rounded to estimate

the TIB class. A rounded estimate of 0 signifies non-TIB and a rounded estimate of 1 signifies TIB. As with the logistic regression, if some rounded estimate does not equal 0 or 1, then the estimate is considered to indicate non-TIB.

Evaluation of the models are based on accuracy (% of correct estimates), sensitivity (true positives / all positives), and specificity (true negatives / all negatives) where TIBs are considered positive and non-TIBs are negative. We use 10-fold cross validation on the chosen processed data set (control, dementia, or combined) in which models are iteratively trained on data in 9 of the 10 random partitions and are then evaluated using the remaining partition.

The large dimensionality of the feature space is reduced by selecting only $N = 10, 20, \ldots, 250$ features to build the model. To select features, we use the Pearson product-moment correlation coefficient:

$$\rho(\text{TIB, feature}) = \frac{\text{cov}(\text{TIB, feature})}{\sigma_{\text{TIB}}\sigma_{\text{feature}}} \tag{6}$$

where cov(X,Y) is the covariance between X and Y, and σ_X is standard deviation of X. The correlation coefficients between the features and TIB class is calculated using only the data in the training set. N features are selected in descending order of absolute correlation with TIB class. Cross-validation is repeated for different values of N for each model.

Three models are evaluated: logistic regression, MDN with 100 hidden units, and MDN with a variably scaled number of hidden units where the number of hidden units is five times the number of features selected. The models are tested on each of control, dementia, and combined data.

5.2 Results of Identifying TIBs

First, we compare data from controls and those with dementia separately, before combining the data in Section 5.2.2.

5.2.1 Control vs. Dementia. Table 3 presents the top five features by absolute correlation with TIB class over all of the Control and Dementia data. Note that no features in common are shared between groups, indicating that TIBs may manifest differently in each population. On Dementia data, the MDN with 100 hidden units and 20 to 25 features

Table 3
Top five features, by absolute correlation with TIB class, in data from people without and with dementia. For example, "S1 → SQ" is a count of sentential structures that are also interrogative, and DT, NN, and PRP are determiners, nouns, and personal pronouns, respectively. Imageability and familiarity are scores obtained from the Bristol norms (Stadthagen-Gonzalez and Davis 2006).

Control Features	Dementia Features
S1 → SQ (interrogative)	Num. of PRPs :: Num. NNs + Num. PRPs
Num. DTs	Num. NNs
Num. DTs :: Num. words	Num. NNs :: (Num. NNs + Num. VBs)
Mean imageability	NP → PRP (personal pronoun)
Mean familiarity	Mean NN imageability

shows the most promise (78.9% accuracy); on Control data, the logistic regression with 50 features performs the best (93.6% accuracy). Note that although Control models produce higher accuracy and specificity across both logistic regression and MDN methods (two-tailed heteroscedastic *t*-tests, $p < 0.001$), sensitivity is far lower than in Dementia data, suggesting that more TIBs are proportionally "missed" in Control data.

The logistic regression model on Dementia data achieves peak accuracy of 76.48% at 60 features, peak sensitivity of 74.04% at 60 features, and peak specificity of 80.06% at 35 features. On Control data, its peak accuracy is 93.57% at 70 features selected, its peak sensitivity is 31.96% at 50 features, and its peak specificity is 100.00% for $\geq$95 features. These results are shown in Figure 2.

The MDN with 100 hidden units on Dementia data achieves peak accuracy of 78.87%, peak sensitivity of 75.32%, and peak specificity of 84.05%, each at 20 features. On Control data, it has peak accuracy of 92.70% at 20 features, peak sensitivity of 13.51% at 30 features, and peak specificity of 99.70% at 10 features. This is shown in Figure 3.

The MDNs with five times as many hidden units as selected features on Dementia data achieve peak accuracy of 77.86% at 20 features, peak sensitivity of 75.376% at 20 features, and peak specificity of 82.42% at 25 features. On Control data, they have peak accuracy of 92.70% at 5 features, peak sensitivity of 61.91% at 180 features, and peak specificity of 100.00% at 5 features. This is shown in Figure 4.

5.2.2 Combined Data Set (Control and Dementia). The top 20 features selected by absolute correlation with TIB class are shown in Table 4, along with p-values for these features after fitting the data with a multilinear logistic regression, given all data together. These features are composed entirely of lexical features, in contrast with related work on the same data in which we found that acoustic measures are highly indicative of AD (Fraser, Meltzer, and Rudzicz 2016). Indeed, pausing and phonation rate, along with word repetition, incomplete words, and fillers (see Section 4.1.3) have high diagnostic value, and we expected these to be related to TIB 11 (reprise / minimal dysfluency), but this was not the case. Interestingly, the average WordNet depth and mean noun familiarity are among the top five features, but make no appearance among the top five features selected from either the Control or Dementia sets separately. In particular, the ratio of pronouns to all nouns correlates positively with the presence of TIBs (as one might expect, given the relatively generic nature of the former) with a correlation of 0.43.

The logistic regression model achieves stable results with the exception of a large variance in sensitivity at 90 features selected. The logistic regression model has the highest rate of identifying TIBs (79.43% accuracy) with 75 selected features with a sensitivity of 60.93% while retaining good specificity (88.49%). The MDN with 100 hidden units obtains greater variance than the logistic regression model, but has several advantages. The peak accuracy and sensitivity at 81.55% and 63.26% of this MDN are strictly higher than the peak accuracy and sensitivity of the logistic regression. Furthermore, the MDN achieves these results at only 20 features selected compared with 75–80 features selected for the logistic regression. In environments where TIB estimation needs to be performed dynamically, it can be computationally intensive to extract a large number of features in real-time. The MDN with five times as many hidden units as input features selected (Figure 5) performs adequately well for up to 80 features selected although it exhibits a high variance as the number of features increases beyond 100, perhaps because of overfitting. There is a large variance in the results for 30 features selected as well, which could be due to an unusual partitioning of the data during cross-validation.

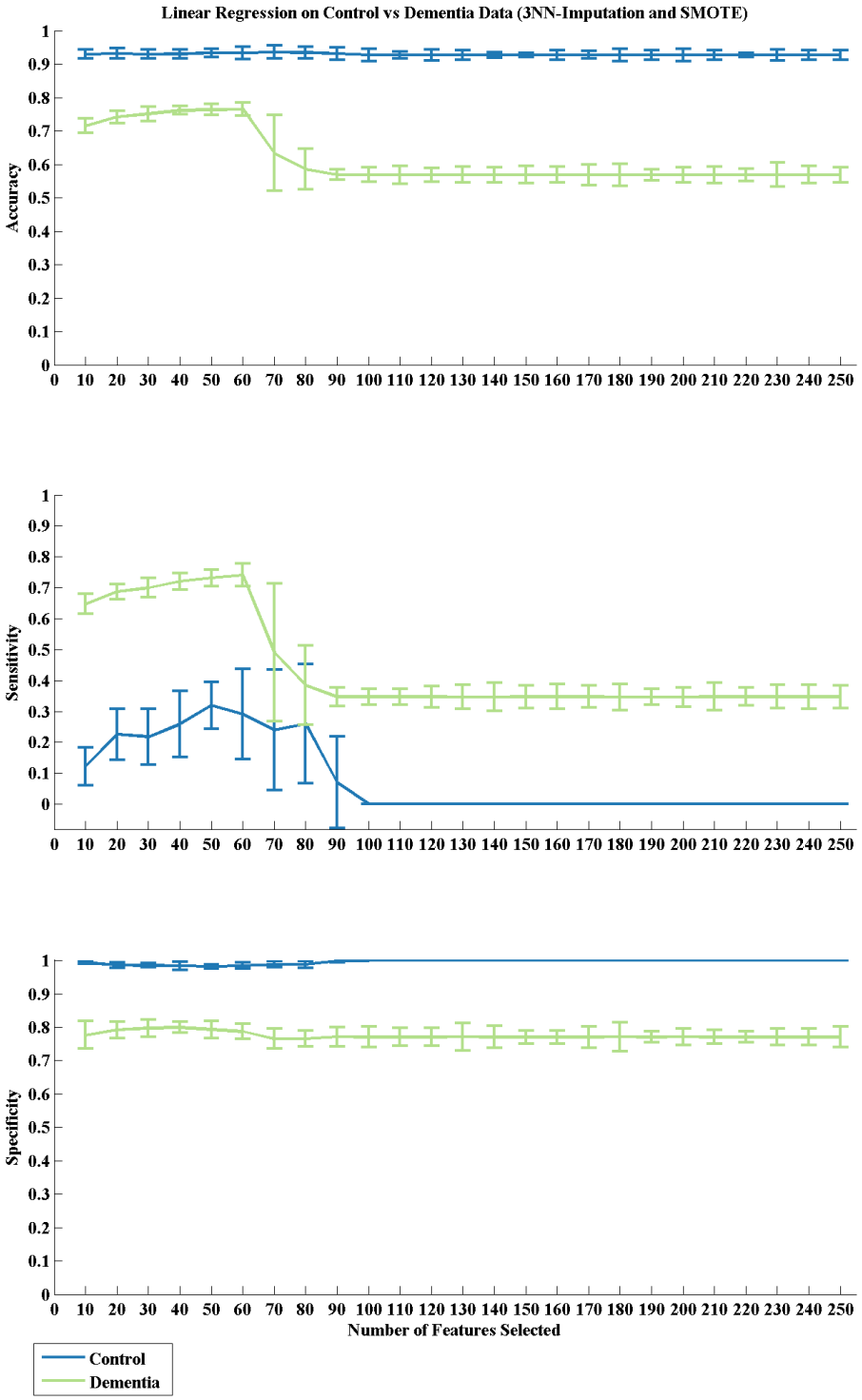


Figure 2
Accuracy, sensitivity, and specificity of TIB detection with logistic regression on Control and Dementia data (processed with 3NN-Imputation and SMOTE).

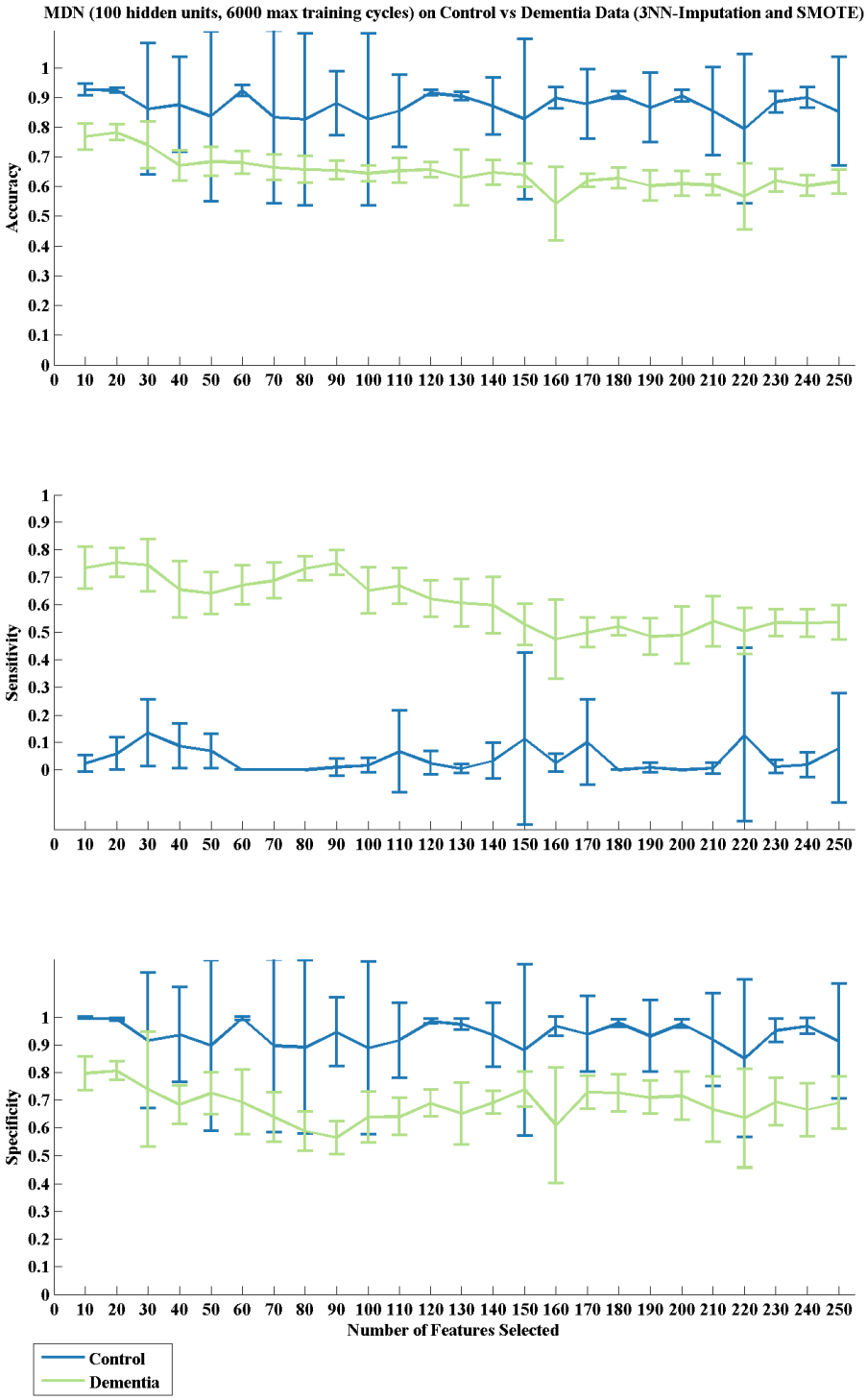


Figure 3
Accuracy, sensitivity, and specificity of TIB detection with MDN (100 hidden units, 6,000 max training cycles) on Control and Dementia data (processed with 3NN-Imputation and SMOTE).

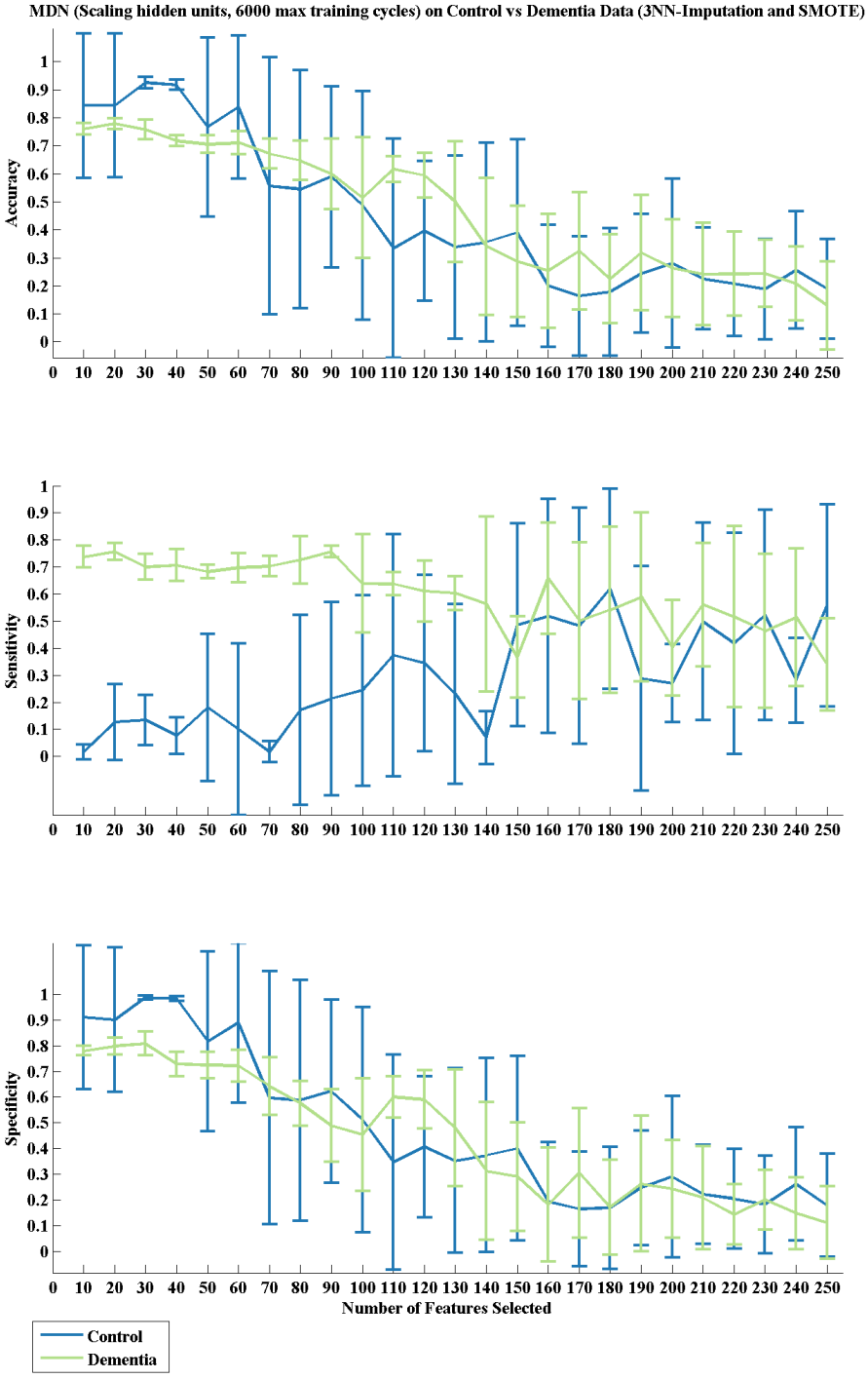


Figure 4
Accuracy, sensitivity, and specificity of TIB detection with MDN (scaled number of hidden units, 6,000 max training cycles) on Control and Dementia data (processed with 3NN-Imputation and SMOTE).

Table 4
Top features using all data, ordered by absolute correlation r with TIB class, with LR p-values.

Feature Name	r	LR p-value
Number of Pronouns :: Number of Pronouns and Nouns	0.42812	< 0.001
Mean Noun Imageability	−0.33533	0.395
Mean Noun Familiarity	−0.32535	0.0127
Average WordNet Depth (i.e., word specificity §4.1.1)	−0.31758	< 0.001
NP → PRP	0.31205	0.560
Number of Nouns	−0.31170	< 0.001
Number of Nouns :: Number of Nouns and Verbs	−0.30733	0.353
Mean Imageability	−0.30579	0.187
Mean Age of Acquisition	−0.27172	0.00721
Number of Demonstratives :: Number of Words	0.26869	< 0.001
WHNP → WP (*wh*-NP with an NP GAP → *wh*-pronoun)	0.26614	< 0.001
Cue words (§4.1.4)	0.25988	0.641
Number of Nouns :: Number of Words	−0.25929	0.0411
NP → DT NN	−0.25321	< 0.001
Number of Complex T-units :: Number of T-units	0.24475	0.00231
Number of Dependent Clauses	0.24265	0.00131
Noun Imageability	−0.24141	0.456
Semantic Similarity (§4.1.2)	−0.23932	0.633
Number of Dependent Clauses :: Number of T-units	0.23818	0.0483
Number of Complex T-units	0.23598	0.0694

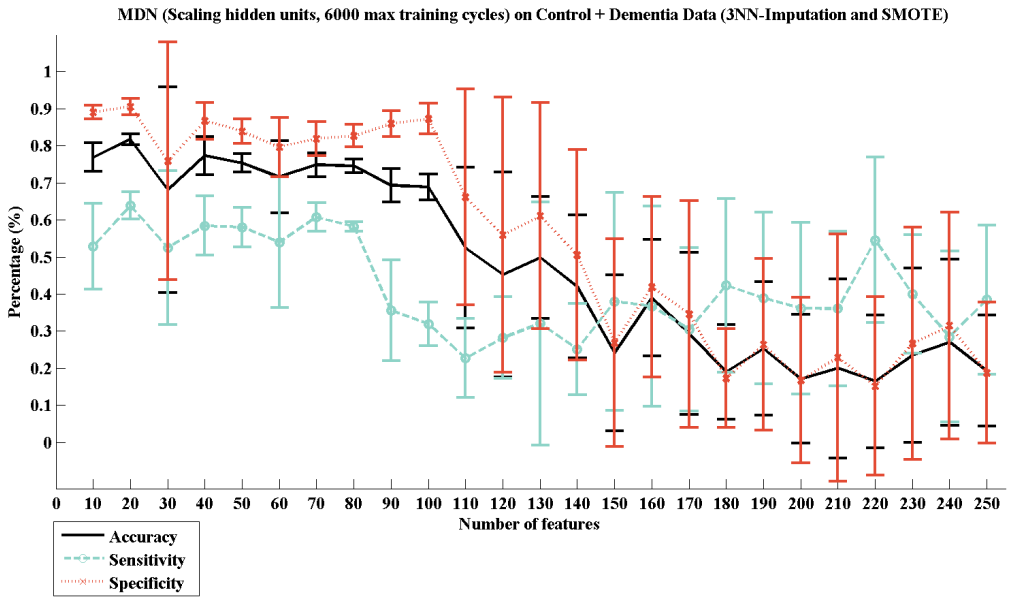
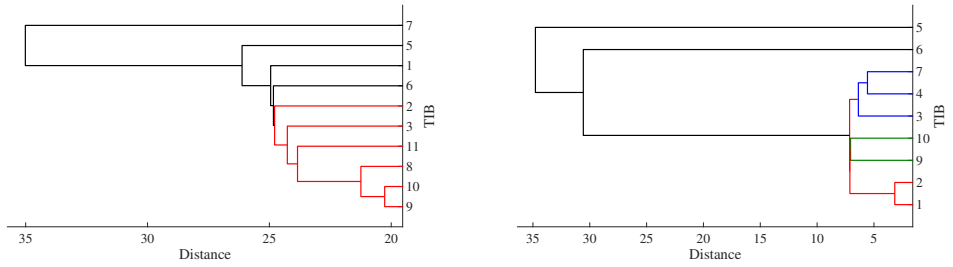


Figure 5
MDN (scaling hidden units, 6,000 max training cycles) on *combined* data (processed with 3NN-Imputation and SMOTE). Values corresponding to 10, 20,... features are shown. Peak accuracy of 81.70% at 20 features, peak sensitivity of 63.88% at 20 features, and peak specificity of 90.51% at 20 features.



(a) Dementia (b) Controls

Figure 6
Dendrograms showing clusters of TIBs for Dementia (a) and Control data (b). Distance is Euclidean in z-scored feature space. TIBs that are individually separable are on black branches; colored subtrees represent smallest clusters of the remaining TIBs.

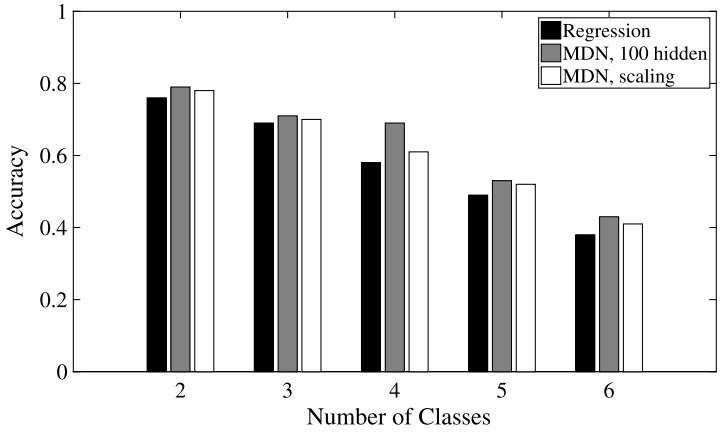
5.3 Clustering TIBs

Although this work is focused on the detection of *any* TIB, an automated system should react differently to different *types* of TIB. In this section, we perform a brief tangential but related experiment in n-way classification for groups of TIBs, rather than the binary case of Section 5.2. To cluster TIBs into those groups by similarity, we first z-score each feature to eliminate the influence of those dimensions with excessively large variances. We then compute the mean of all feature vectors representing utterances for each TIB, and finally perform agglomerative hierarchical clustering using the shortest Euclidean distance between those vectors. The resulting dendrograms are shown in Figure 6.[1]

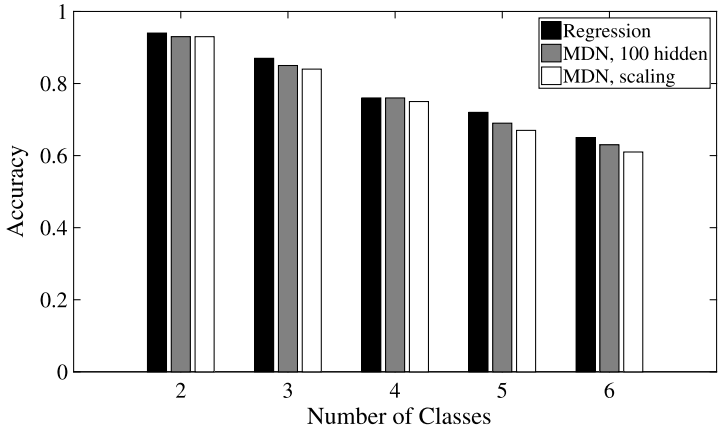
Qualitatively, certain patterns are evident; for example, there is high similarity between TIBs 8, 9, and 10 (metalinguistic or unrelated speech) and between TIBs 2 and 3 (different requests for confirmation) in Dementia. However, the latter pair are not as similar among Controls, perhaps because of a different degree of reduction in TIB 2. Also, Control subjects do not exhibit any lack of uptake (TIB 8) at all in this view, which is in sharp contrast to its relatively frequent occurrence among people with dementia, especially in our related work (Rudzicz et al. 2015). Control subjects also exhibit neither reprise (TIB 11) nor global request for repetition (TIB 12) in these data, and subjects with Dementia exhibit neither the latter nor request for confirmation (repetition with elaboration, TIB 4).

For these experiments (and those in Section 6.3, where these clusters are used), we sweep the number of TIB clusters (excluding no-TIB) from 1 to 5. Specifically, one class is "no-TIB" and another is all remaining TIBs, combined, after isolating, iteratively, TIBs 7, 5, 1, and 6 in Dementia; TIBs 5, 6, {3, 4, and 7}, and {9 and 10} in Controls, and TIBs 4, 7, 1, and 6 in the combined set. Figure 7 shows accuracy in TIB cluster classification for each of the Dementia, Controls, and combined data sets, using the top 30 features in each, for each value of n (including no-TIB). As expected, there is a significant trend towards lower accuracy as the number of classes increases.

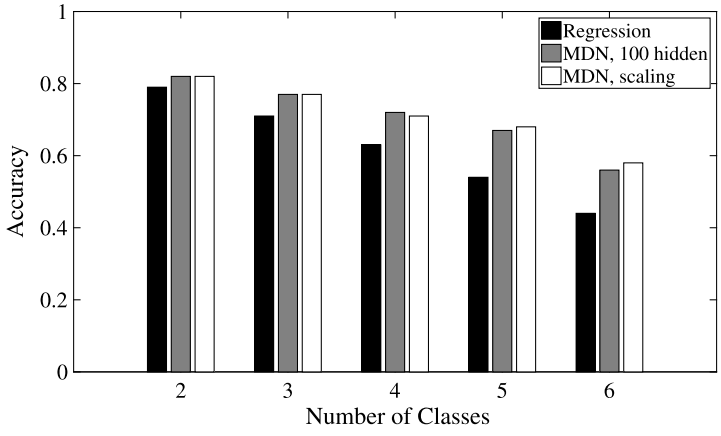
1 A third dendrogram, for the combined set, is not shown.



(a) Dementia



(b) Controls



(c) Combined

Figure 7
Accuracy of *n*-way classification, using TIB-clusters and the no-TIB class, for each of Dementia, Control, and combined data sets using each of the regression and MDN models.

6. Experiment 2: Avoiding TIBs (While Completing Tasks) with POMDPs

Software that can engage older adults with AD in dialogue can benefit from tracking the speech and language characteristics of those individuals and from automatically identifying when the dialogue breaks down. Such software must *also* optimize its output to the user, especially if the assistance offered by caregivers in the completion of daily tasks is to be mimicked.

Here, we build a dialogue system to avoid TIBs using a partially observable Markov decision process (POMDP; see Section 6.1) using DementiaBank data. The specification of the data that we used are shown in Table 5. We removed the dialogues that do not include at least one interviewer and one participant turn. Individuals with dementia were much more likely (two-tailed $t(1) = 318.36, p < 0.001$) to exhibit TIBs in an utterance (10.17%) than Controls (2.25%). Note that in order to design a POMDP for reducing confusion, there must be some other function (e.g., an extrinsic task) to optimize. Our pilot work involved focusing only on TIBs in the CCC database, which involves free-form conversation. However, without a task to complete, the POMDP consistently decided that the best way to avoid confusion is to simply say nothing at all; therefore, we only consider the DementiaBank data here, as described subsequently.

Table 6 shows a sample dialogue with an individual with dementia. Here, the Speaker column indicates interviewer (INV) or participant (PAR) turns. We recruited speech-language pathologists (SLPs) to annotate the dialogues with *actions*. For interviewer turns, actions are ASK (explicit requests for information), ENCOURAGEMENT (e.g., "*you're doing great*"), and REPEAT (verbatim or paraphrased repetitions of previous requests, including continuations). In the interviewer's CHAT transcripts, there are occasionally turns consisting of the "mhm" token that corresponds to "Yes" or ACKNOWLEDGMENT. Note that it is possible for an interviewer to utter *none* of these actions. As the interviewers in this database were explicitly instructed against other forms of interaction, no other actions (e.g., responses to requests for information, TIB 6) were recorded.

For participant turns, the SLPs instead annotated both the presence of specific TIBs and whether picture elements were mentioned. Those picture descriptions relate to the state of *task completion* and are, if extant, either SINK (i.e., information about the right part of the picture in Figure 1) or COOKIE (i.e., information about the left part of the picture). Fleiss' kappa, $\kappa = 0.84$, indicates satisfactory agreement, so we consider the annotations of the first SLP here.

6.1 Partially Observable Markov Decision Processes

POMDPs are generalized Markov decision processes that inherently model uncertainty with regards to latent factors in a system, namely, the state space S which can be further

Table 5
Specifics of Control, Dementia, and combined data.

	No. of dialogues	No. of turns	TIB%	Mean turns/dialogue (s.d.)	Max turns/dialogue
Dementia data	283	5128	10.17%	3.69 (3.38)	39
Control data	159	2080	02.25 %	2.25 (1.37)	7
Combined data	442	7208	07.89%	3.16 (2.20)	39

Table 6
A sample dialogue from DementiaBank.

Speaker	Actions/States transcripts	Transcript, in reduced CHAT format
INV	ASK	just tell me whats happening in the picture . .
PAR	SINK	the pearl [: poor] [* p:w] &mo moms gettin(g) her wet [/ /] feet wet (be)cause she thinking of days gone by and then the water run . [+ gram] .
PAR	COOKIE	(.) and &uh that boy whether he knows or not hes gonna [: going to] crack his head on the back of that counter trying to get too many cookies out . .
PAR	COOKIE-TIB	(..) he [/ /] (.) &uh I dont know the significance of that little girl putting her hand to her mouth reaching . .
PAR	TIB	unless thats the way all silly girls act at that age . [+ gram].
INV	REPEAT	(...) anything else ? .
PAR	TIB	pardon ? [+ exc].
INV	REPEAT	anything else ? .
	...	

factorized into substates with conditional dependencies. It is a model of an active agent in the world whose objective is to map a set of observations O from the environment to the optimal action A to perform to maximize the reward over time. This results in a policy of actions to take to optimize the reward. To learn a full POMDP model, we solve for the transition probabilities between states, $P(s_{t+1}|s_t, A)$ where $s_t \in S$, the observation function $P(o_t|s_t)$ where o_t is the observation in O at time t, and an optimal mapping of actions to observations $\pi = f(o_t) \in A$.

In some conditions, the parameters of a POMDP can be learned together through online reinforcement. This is intractable in many real-world scenarios, so Hoey et al. (2010) simplified the problem by explicitly designing a reward function $R(s_t, a_t), s \in S, a \in A$, learning a prior model of user behavior (via maximum likelihood estimation given appropriately annotated data) and a prior model of transitions and observations. Similarly, we use these models together to develop the optimal policy. Because the states of the system are latent, we maintain a belief state $b_t(s_t)$, which defines the probability of s_t. In order to optimize the policy π, all possible future rewards must be considered:

$$V^{\pi}(b) = \sum_{t} \gamma^t E_{b_t|\pi}[R] \tag{7}$$

where γ is the discount factor (< 1) that reduces the value of future rewards. Poupart (2005) used point-based value iteration to optimize this function by only planning for the most probable belief states. This algorithm first picks a finite set of belief states $B = \{b_0, b_1, ..., b_q\}$ then finds the optimal value function for each of these states (Zhang, Lee, and W 1999). The value function is then updated for each of the belief states. Next, we stochastically expand the set of belief states by greedily choosing the furthest reachable

states. These two steps are then repeated for a fixed number of iterations. Although this algorithm is not guaranteed to converge, it has error bounds that depend on the density of B (Pineau, Gordon, and Thrun 2003).

6.2 POMDP Design

We define the POMDP **states** associated with each turn in the dialogue with the triple *[TIB, SINK, COOKIE]*. Because each of the component variables is binary, there are $2^3 = 8$ POMDP states in total. Similarly, the POMDP **actions** are defined using zero or more actions from [ASK, ENCOURAGEMENT, REPEAT]. Thus, we also have eight possible action values. The POMDP **observations** are based on all features described in Section 4. Each feature is discretized to four different classes. This is done by fitting values of *each feature* to a normal distribution. The middle class (class 2) is centered on mean μ and covers the range $[\mu - \sigma, \mu + \sigma]$; class 1 covers feature values less than $\mu - \sigma$, and class 3 covers feature values greater than $\mu + \sigma$. Finally, class 0 is reserved for missing values.

Discretizing features in this way allows for discrete-space modeling, but results in $4^{(81+178)}$ possible observation vectors, which is clearly intractable. We therefore empirically select the top three features from Table 4 with maximal correlation to TIBs. These features include (1) the proportion of number of pronouns to the number of pronouns and nouns, (2) the number of nouns, and (3) the proportion of the number of demonstratives to the number of nouns. We tried more and fewer features; accuracy increases slightly up to eight features, but it was not feasible to add more than 2,048 observations in the POMDP. Furthermore, we use one feature for the SINK state and one for the COOKIE state. For each participant utterance, the SINK feature encodes the number of keywords mentioned about the SINK part of the picture (the right part) and the COOKIE feature encodes the number of keywords concerning the COOKIE part of the picture (the left part).

POMDP reward functions are often defined to encourage task completion (Young et al. 2013). Similarly, we set a reward of $+1$ when either SINK or COOKIE is mentioned, $+10$ if both are mentioned, and 0 otherwise. In order to avoid confusion in the participant, TIBs are penalized by -5 if any TIB occurs, otherwise the reward is increased by $+5$. This reward function naturally favors states in which both SINK and COOKIE occur, without TIB.

Given all annotations of state s_t, action matrix A and observation vector o_t, we calculate the following likelihoods for DementiaBank data using maximum likelihood estimation:

1. $P(s_{t+1}|s_t, A)$
2. $P(o_t|s_t, A)$

In our experiments we learn three POMDPs: Dementia POMDP, Control POMDP, and Combined POMDP, where those transition and observation functions are learned separately from dementia dialogues, control dialogues, or both, respectively. The discretization step for observations are also done separately for each type of model given the representative data.

6.3 POMDP Results

After learning the parameters of each POMDP model, the policy is approximated using point-based value iteration, which uses a finite number of beliefs for solving a POMDP model approximately. The number of beliefs is a parameter that is generally hand-tuned empirically. For our experiments, the solver used 10,000 random samples to estimate the optimal policy.

POMDP policies can be evaluated based on aspects such as dialogue lengths and accumulated discounted rewards in simulation runs (Young et al. 2013). We evaluate the learned POMDP policies in 1,000 simulation runs using the default simulation in Perseus (Spaan and Vlassis 2005) in which the true state of the user is estimated from the provided transition function, and the observations are sampled from the provided observation function. The discounted rewards are calculated based on the reward function and the "true" user state.

We compare the mean dialogue turns (until the maximum reward is reached) and the mean of accumulated discounted rewards for the POMDP policy against baseline policies of only performing one action, that is, either ASK, ENCOURAGEMENT, or REPEAT. A random policy (RAND) is also attempted in which one of the eight possible POMDP actions is performed at each turn. We also estimate POMDP *accuracy* using existing dialogue annotations. This is done by representing each dialogue in the form of an action observation trajectory. After learning the POMDP policy, we start by the initial belief, b_0. The action suggested by the POMDP for the initial belief, a_0, is compared with the interviewer's action, h_0. The belief is updated to b_1 based on the received observation. Subsequent POMDP actions are then compared to caregiver actions until the end of that dialogue (trajectory). Accuracy is the proportion of actions taken by the POMDP that match at least one of the annotated actions, in each turn. Because this is a disjunctive operation on binary vectors, we expect a random policy to do well.

The results in Table 7 show that the learned POMDPs perform well based on simulation runs (shorter and more task-oriented conversations are preferred) and accuracy. In particular, the results show that the Dementia POMDP performs better than all the baseline policies, in all cases at $p < 0.05$ given two-tailed *t*-tests with Bonferroni correction. The Dementia POMDP produces shorter dialogues to achieve its maximum reward, on average around four turns. The Dementia POMDP also accumulates more rewards on average. The standard deviations for both dialogue turns and accumulated rewards are smaller than those values for the baseline policies. We also calculate the observed agreement of the PODMP actions with those of human caregivers based on Cohen's kappa, achieving $\kappa = 0.66$ (Control), $\kappa = 0.42$ (Dementia), and $\kappa = 0.55$ (Combined), each of which constitutes moderate agreement between the POMDP and human annotators.

As mentioned in Section 5.3, we split the TIB class into $K = 1 \ldots 5$ clusters. Then, we construct new POMDPs given each K, that is, we designed POMDPs with 1, 2, 3, 4, and 5 TIB states, plus one no-TIB state; there are also four task states, namely, two values on the left and two values on the right of the cookie-theft picture. We also update the transition function of each POMDP based on their subsequent states using a simple maximum-likelihood function with an $\epsilon = 0.1$ smoothing parameter. The new POMDPs are created using Dementia, Control, and Combined data.

We compare the mean number of dialogue turns (until the maximum reward is reached) and the mean of accumulated discounted rewards for the POMDP policy against baseline policies of only performing one action (i.e., either ASK, ENCOURAGEMENT, or REPEAT). A random policy (RAND) is also attempted in which one

Table 7
Evaluation of learned POMDP policies based on mean number of dialogue turns, mean accumulated discounted rewards in simulation, and accuracy relative to human annotations, as compared with baseline policies of choosing either the indicated action only, or uniformly random actions.

Dementia	POMDP	ASK	ENCG	REPEAT	RAND
Dialogue turns	4.23 (2.22)	5.12 (4.04)	11.11 (8.61)	7.29 (6.89)	8.69 (7.20)
Accumulated rewards	51.75 (8.59)	45.27 (13.24)	30.23 (11.99)	42.63 (9.23)	31.59 (11.53)
Accuracy	91.4%	32.5%	36.3%	27.8%	86.2%
Control	POMDP	ASK	ENCG	REPEAT	RAND
Dialogue turns	2.94 (1.53)	2.95 (1.82)	6.53 (4.55)	3.20 (1.95)	4.16 (2.87)
Accumulated rewards	57.64 (8.30)	49.70 (10.47)	45.15 (7.14)	36.16 (11.39)	38.47 (11.12)
Accuracy	96.1%	25.6%	19.5%	14.8%	84.1%
Combined	POMDP	ASK	ENCG	REPEAT	RAND
Dialogue turns	4.03 (2.47)	3.73 (2.53)	9.52 (7.05)	6.28 (6.24)	6.17 (4.76)
Accumulated rewards	50.86 (8.99)	49.34 (10.11)	33.30 (10.79)	42.09 (10.92)	34.42 (11.92)
Accuracy	94.1%	30.8%	32.0%	24.4%	86.7%

of the eight possible POMDP actions is performed at each turn. We also defined two hand-crafted policies. The first (HC policy) starts with the ASK action (explicit request for information), then performs ENCOURAGEMENT (e.g., "*you're doing great*"), and then the REPEAT action (verbatim or paraphrased repetitions of previous requests, including continuations). This HC policy continues these last two actions until the end of conversation. The second hand-crafted policy takes a frequency-based approach. For each state, the fHC policy (frequentist HC) performs the action that has been performed by caregivers most frequently, given the state. At the time of interaction, the fHC policy uses the POMDP belief; that is, it takes the maximum likelihood state from the belief and performs the caregivers' most frequent action for that state. Note that one separate fHC policy is learned for each population (i.e., Dementia, Control, and Combined) separately. The POMDP performance is shown, over 1,000 simulation runs, in Figures 8 and 9, as the number of TIB states increase from 2 to 6 (note that there is no-TIB state besides $K = 1 \ldots 5$ TIB clusters).

Figures 8 (top) and 9 (top) show the performance of the policies on the Dementia model. These show that the optimized policies of the learned Dementia POMDP have consistently better performance than the baseline policies, including the heuristic ones. Figure 9 (top) shows that by increasing the number of classes from 2 to 5 (increasing the number of POMDP states from 8 to 24), the mean of rewards for the Dementia POMDP slightly decreases from 49.23 (s.d. 10.12) to 42.42 (s.d 16.24), over the simulation runs. The second best performance is the ASK policy that, on average, collects 39.21 (s.d. 10.53) and 35.83 (s.d. 13.38) reward, when two and six classes are used, respectively. The third best policy is the Dementia fHC policy that accumulates 30.05 (s.d. 13.77) and 31.48 (s.d. 12.83) rewards on average, respectively, when two and six classes are used.

In addition, the POMDP policies, on average, have shorter dialogue turns to reach the maximum reward than any other policies, based on Figure 8 (top). When six classes are used (24 POMDP states), it takes 4.97 (s.d. 2.99) dialogue turns, on

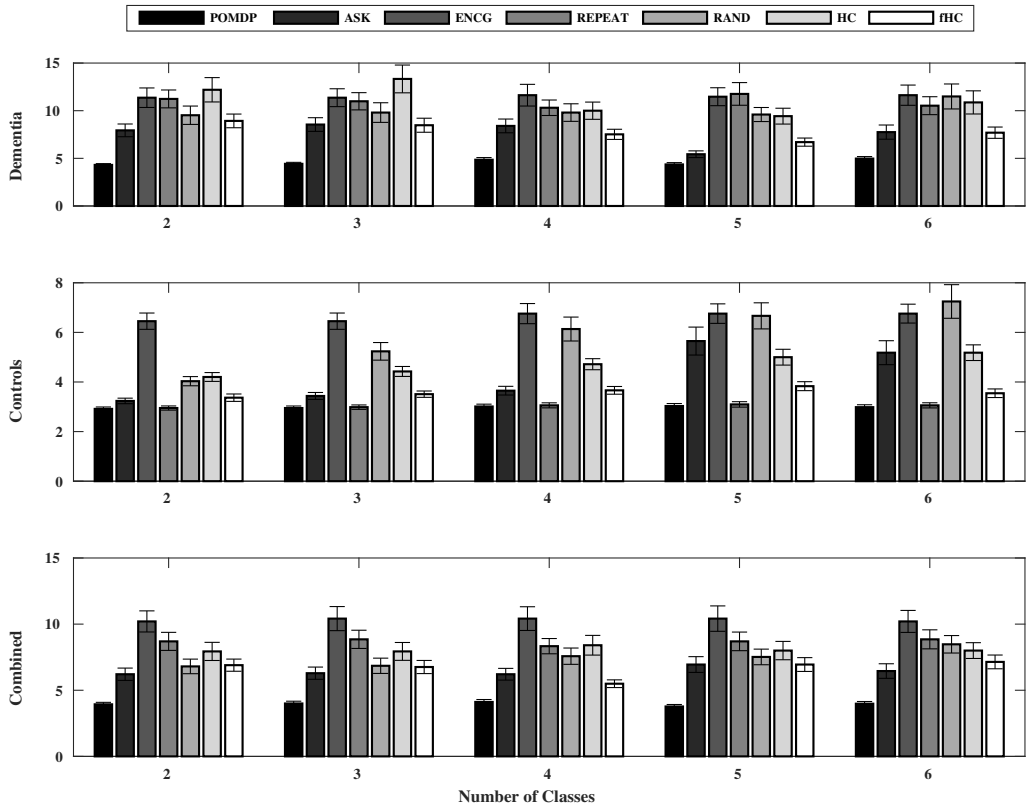


Figure 8
Mean number of dialogue turns in learned POMDP policies, and the baseline policies.

average, for the Dementia POMDP policy to reach the maximum reward, whereas the ASK and Dementia fHC policies take 7.75 (s.d. 8.47) and 6.69 (s.d. 6.75), on average, respectively.

Figures 8 (middle) and 9 (middle) show that, among the Control models, the POMDP and REPEAT policies have the highest performance. The high performance of the REPEAT policy (which asks questions such as "what else do you see in the picture?") could be attributed to the Control population's consistent providing of new information when this action is performed, achieving greater rewards.

Figures 8 (bottom) and 9 (bottom) describe the performance of the policies on the Combined model. The POMDP policy again has the highest performance; when six classes (24 POMDP states) are used, the Combined POMDP on average accumulates 44.60 (13.07) rewards. For the same group, the ASK policy and the Combined fHC policy, on average, collect 34.24 (s.d. 16.57) and 34.61 (s.d. 13.35) rewards, respectively.

Overall, the POMDP policies show better performance than all other policies, based on the mean dialogue turns (until the maximum reward is reached) and the mean of accumulated discounted rewards in simulation runs. In this work, we manually set the reward function, empirically. In the future, we intend to learn this function based on inverse reinforcement learning (Ng and Russell 2000; Choi and Kim 2011; Chinaei and Chaib-Draa 2014).

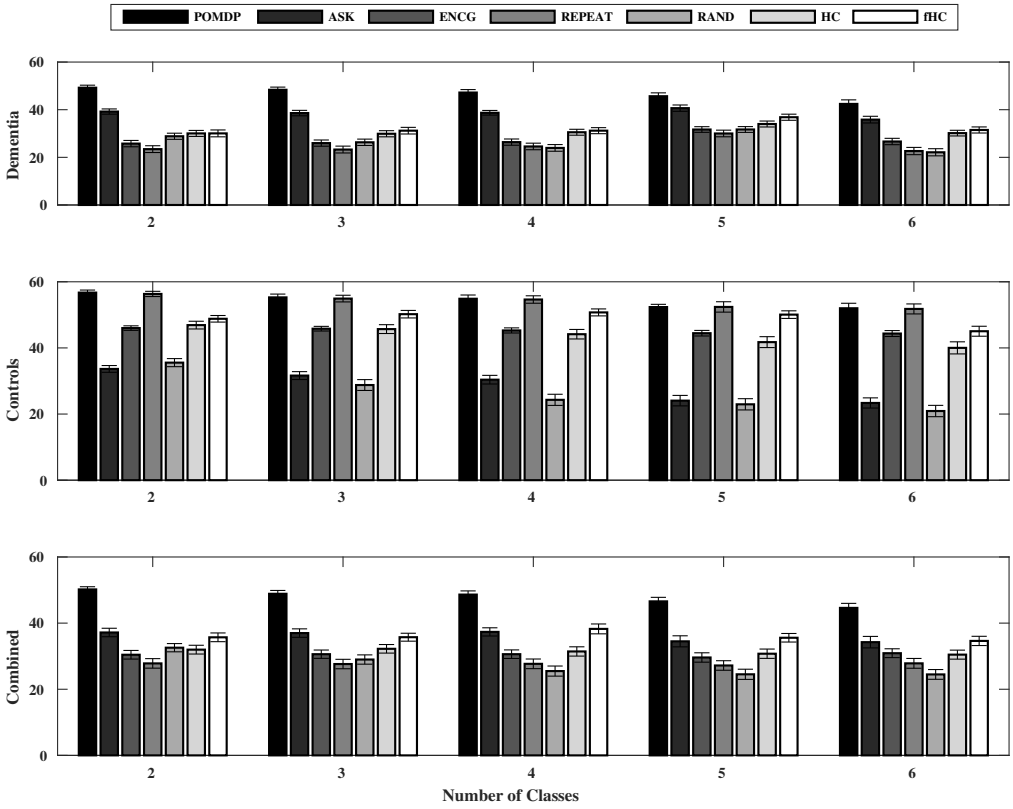


Figure 9
Mean number of accumulated rewards in learned POMDP policies, and the baseline policies.

7. Discussion

This article provides a first assessment of the automatic detection and avoidance of verbal indicators of confusion in dialogue with individuals who have Alzheimer's disease. We have demonstrated that it is possible to identify trouble indicating behaviors (TIBs, both at all, and in specific groups) explicitly in the speech of older adults with AD and matched controls, with accuracies up to 79% and 93%, respectively, using a neural network model and both acoustic and lexico-syntactic features, although the latter appear to be far more indicative. This was validated on a large set of data involving conversations around a standardized picture-description task. Much of this work is based on prior work in which we used similar linguistic features to train machine learning classifiers to distinguish between participants with and without AD (Fraser, Meltzer, and Rudzicz 2016). In that work, we also studied the degree of heterogeneity of linguistic impairments in AD and found four clear factors: semantic impairment, acoustic abnormality, syntactic/fluency impairment, and information/situational impairment. With some modification, assessment of this type (longitudinal or otherwise) can be seamlessly integrated into the conversation-focused systems described in the current article.

Furthermore, we learned several POMDPs using data from individuals with and without AD (both separately and combined) and evaluated those models using

simulation runs and accuracy relative to trained human participants. In simulation, learned POMDPs produce shorter dialogues than four baseline models and accumulate more reword, on average; they also resemble humans far more accurately. This is closely related to our previous work in studying TIBs in older adults with moderate AD interacting with a physical (Wizard-of-Oz) conversational robot in a home setting (Rudzicz et al. 2015). In that work, across the verbal behaviors that indicate confusion, older adults with AD were very likely to simply ignore the robot (TIB 8; lack of uptake), which accounts for over 40% of all such behaviors. However, individuals with AD were much more likely ($t(18) = -4.78, p < 0.0001$) to exhibit *no* TIB when interacting with a robot (18.1% of utterances) than with a non-familiar human (6.7%). The current work completes the largest missing component in that work, namely, the automatic selection of prompts given speech input. Ongoing work involves the use of inverse reinforcement learning to learn the reward function of the POMDP from data.

Given our rapidly aging populations, it is increasingly vital that we develop technological approaches to caring for the elderly in their homes. To a large extent, this will involve speech interaction with augmentative or assistive technologies. Although speech recognition accuracy can be improved in a home-care environment by limiting the vocabulary, and by adapting the acoustic model to the individual (Rudzicz et al. 2015), word error rates remain high in such an environment. Fortunately, while the accuracy of automatic assessment decreases with increased word-error rates, recent work suggests this is weakly correlated ($r = -0.31$) (Zhou, Fraser, and Rudzicz 2016). Also, in practice, several of our conclusions must be tempered by the relatively precise nature of the picture description task. Differences will clearly exist in different interaction types, such as robot-directed activities/exercises, or casual (undirected) conversation. We are currently exploring a variety of hidden variables, including task type and complexity, in our own data collection with a small humanoid robot, but other forms of interaction, including with tablet- or computer-based interfaces and ambient technologies, must also be pursued.

Finally, if assessment and monitoring are increasingly aided by communication software, especially during activities of daily living, then certain bioethics questions arise. Who will bear the responsibility for incorrect decisions if those decisions are the product of statistics and the data from which they are derived? If medically relevant information can be obtained from recordings in one's home, how will individual privacy be maintained, especially if increasingly popular cloud-based speech services are to be used? Inequalities of access and potential loss of privacy are but two risks that must be considered now because—if the benefits of quantifiable, repeatable, and accurate clinical care by machine outweigh those risks—then a fundamental change to healthcare is inevitable.

Acknowledgments

Hamidreza Chinaei is supported by a fellowship from the AGE-WELL NCE and another from the Fonds de recherche du Québec – Nature et technologies. Frank Rudzicz's contribution to this work is funded by an NSERC Discovery grant (RGPIN 435874), by a Young Investigator award by the Alzheimer Society of Canada, and from a project grant from the AGE-WELL NCE. We gratefully thank Maria Yancheva for her insightful and incisive suggested revisions to this work. The original acquisition of the DementiaBank data was supported by NIH grants AG005133 and AG003705 to the University of Pittsburgh, and maintenance of the data archive is supported by NIH-NIDCD grant R01-DC008524 to Carnegie Mellon University.

References

Ahmed, S., A.-M. F. Haigh, C. A. de Jager, and P. Garrard. 2013. Connected speech as a marker of disease progression in

autopsy-proven Alzheimer's disease. *Brain*, 136(12):3727–3737.

Almor, A., D. Kempler, M. C. MacDonald, E. S. Andersen, and L. K Tyler. 1999. Why do Alzheimer patients have difficulty with pronouns? Working memory, semantics, and reference in comprehension and production in Alzheimer's disease. *Brain and Language*, 67(3):202–227.

Batista, G. E. A. P. A. and M. C. Monard. 2002. A study of k-nearest neighbour as an imputation method. In *Second International Conference on Hybrid Intelligent Systems*, volume 87, pages 251–260, IOS Press.

Becker, J. T., F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle. 1994. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *JAMA Neurology*, 51(6):585–594.

Bharucha, A. J., V. Anand, J. Forlizzi, M. A. Dew, C. F. Reynolds III, S. Stevens, and H. Wactlar. 2009. Intelligent assistive technology applications to dementia care: Current capabilities, limitations, and future challenges. *American Journal of Geriatric Psychiatry*, 17(2):88–104.

Bhatnagar, S. C. 2002. *Neuroscience for the study of communication disorders*. Lippincott Wiliams & Wilkins, Baltimore, MD.

Bird, H. S. Franklin and D. Howard. 2001. Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers*, 33(1):73–79.

Bishop, C. M. 1994, Mixture density networks. Technical Report NCRG/94/004, Department of Computer Science and Applied Mathematics, Aston University, Birmingham, UK.

Boller, F. and J. Becker. 2005. *DementiaBank Database Guide*. University of Pittsburgh.

Bortfeld, H., S. D. Leon, J. E. Bloom, M. F. Schober, and S. E. Brennan. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44(2):123–147.

Brysbaert, M. and B. New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Bucks, R. S., S. Singh, J. M. Cuerden, and G. K. Wilcock. 2000. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14:71–91.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chinaei, H. and B. Chaib-Draa. 2014. Dialogue POMDP components (part II): Learning the reward function. *International Journal of Speech Technology*, 17(4):325–340.

Choi, J. and K.-E. Kim. 2011. Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research*, 12:691–730.

Cortese, M. J., D. A. Balota, S. D. Sergent-Marshall, R. L. Buckner, and B. T. Gold. 2006. Consistency and regularity in past-tense verb generation in healthy ageing, Alzheimer's disease, and semantic dementia. *Cognitive Neuropsychology*, 23(6):856–876.

Cummings, J. L. 2004. Alzheimer's disease. *New England Journal of Medicine*, 351(1):56–67.

Ernst, R. L., J. W. Hay, C. Fenn, J. Tinklenberg, and J. A. Yesavage. 1997. Cognitive function and the costs of Alzheimer disease — an exploratory study. *Archives of Neurology*, 54:687–693.

Faber-Langendoen, K., J. C. Morris, J. W. Knesevich, E. LaBarge, J. P. Miller, and L. Berg. 1988. Aphasia in senile dementia of the Alzheimer type. *Annals of Neurology*, 23(4):365–370.

Folstein, M. F., S. E. Folstein, and P. R. McHugh. 1975. Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198.

Fraser, K., F. Rudzicz, and E. Rochon. 2013. Using text and acoustic features to diagnose progressive aphasia and its subtypes. In *Proceedings of Interspeech 2013*, pages 2177–2181. Lyon.

Fraser, K. C., J. A. Meltzer, and F. Rudzicz. 2016. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(3):407–422.

Gaugler, J. E., F. Yu, K. Krichbaum, and J. F. Wyman. 2009. Predictors of nursing home admission for persons with dementia. *Medical Care*, 47(2):191–198.

Godfrey, J. H., E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, pages 517–520, San Francisco, CA.

Goldfarb, R. and M. J. S. Pietro. 2004. Support systems: Older adults with neurogenic communication disorders. *Journal of Ambulatory Care Management*, 27(4):356–365.

Goodglass, H. and E. Kaplan. 1983. *The Assessment of Aphasia and Related Disorders*, ed. 2. Lea and Febiger, Philadelphia, PA.

Gravano, A., J. Hirschberg, and Š. Beňuš. 2012. Affirmative cue words in task-oriented dialogue. *Computational Linguistics*, 38(1):1–39.

Guinn, C. and A. Habash. 2012. Language analysis of speakers with dementia of the Alzheimer's type. *AAAI Fall Symposium Series*, pages 8–13, Menlo Park, CA.

Hirst, G. and V. W. Feng. 2012. Changes in style in authors with Alzheimer's disease. *English Studies*, 93(3):357–370.

Hoey, J., P. Poupart, A. Von Bertoldi, T. Craig, C. Boutilier, and A. Mihailidis. 2010. Automated handwashing assistance for persons with dementia using video and a partially observable Markov decision process. *Computer Vision and Image Understanding*, 114(5):503–519.

Hopper, T. 2001. Indirect interventions to facilitate communication in Alzheimer's disease. *Seminars in Speech and Language*, 22(4):305–315.

Kempler, D. 1995. Language changes in dementia of the Alzheimer type. In R. Lubinski, editor, *Dementia and Communication: Research and Clinical Implications*, pages 98–114, Singular, San Diego, CA.

Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics-Volume 1*, pages 423–430, Sapporo.

Le, X. 2010. Longitudinal detection of dementia through lexical and syntactic changes in writing. Master's thesis, University of Toronto.

Little, M. A., P. McSharry, I. Moroz, and S. Roberts. 2006. Nonlinear, biophysically informed speech pathology detection. In *Proceedings of ICASSP 2006*, pages 1080–1083, Toulouse.

Lu, X. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

MacWhinney, B. 2014. The CHILDES project. `http://childes.psy.cmu.edu/manuals/chat.pdf`.

Marwan, N. and J. Kurths. 2002. Nonlinear analysis of bivariate data with cross recurrence plots. *Physics Letters A*, 302(5–6):299–307.

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of ACM*, 38(11):39–41.

Ng, A. Y. and S. J. Russell. 2000. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning (ICML'00)*, pages 663–670, Haifa.

Nicholas, M., L. K. Obler, M. L. Albert, and N. Helm-Estabrooks. 1985. Empty speech in Alzheimer's disease and fluent aphasia. *Journal of Speech and Hearing Research*, 28:405–410.

Orange, J. B., R. B. Lubinsky, and D. J. Higginbotham. 1996. Conversational repair by individuals with dementia of the Alzheimer's type. *Journal of Speech and Hearing Research*, 39:881–895.

Pakhomov, S. V., G. E. Smith, D. Chacon, Y. Feliciano, N. Graff-Radford, R. Caselli, and D. S. Knopman. 2010. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology*, 23:165–177.

Pineau, J., G. Gordon, and S. Thrun. 2003. Point-based value iteration: An anytime algorithm for POMDPs. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1025–1032, Acapulco.

Pope, C. and B. H. Davis. 2011. Finding a balance: The Carolinas Conversation Collection. *Corpus Linguistics and Linguistic Theory*, 7(1):143–161.

Poupart, P. 2005. *Exploiting Structure to Efficiently Solve Large Scale Partially Observable Markov Decision Processes*, Ph.D. thesis, University of Toronto.

Roark, B., M. Mitchell, J. Hosom, K. Hollingshead, and J. Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2081–2090.

Rochon, E., G. S. Waters, and D. Caplan. 2000. The relationship between measures of working memory and sentence comprehension in patients with Alzheimer's disease. *Journal of Speech, Language, and Hearing Research*, 43:395–413.

Rudzicz, F., R. Wang, M. Begum, and A. Mihailidis. 2015. Speech interaction with personal assistive robots supporting

aging-at-home for individuals with Alzheimer's disease. *ACM Transactions on Accessible Computing*, 7(2):1–22.
Schegloff, E. A., G. Jefferson, and H. Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.
Silva, D., L. Oliveira, and M. Andrea. 2009. Jitter estimation algorithms for detection of pathological voices. *EURASIP Journal on Advances in Signal Processing*, 2009: 1–9.
Small, J. A., E. S. Andersen, and D. Kempler. 1997. Effects of working memory capacity on understanding rate-altered speech. *Aging, Neuropsychology, and Cognition*, 4(2):126–139.
Small, J. A., G. Gutman, S. Makela, and B. Hillhouse. 2003. Effectiveness of communication strategies used by caregivers of persons with Alzheimer's disease during activities of daily living. *Journal of Speech, Language, and Hearing Research*, 46(2):353–367.
Smith, E. A. and R. J. Senter. 1967. Automated readability index. Technical report AMRL-TR-6620, Wright-Patterson AFB:iii.
Spaan, M. T. J. and N. Vlassis. 2005. Perseus: Randomized point-based value iteration for POMDPs. *Journal of Artificial Intelligence Research*, 24(1):195–220.
Stadthagen-Gonzalez, H. and C. J. Davis. 2006. The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, 38(4):598–605.
Teufel, S. and M. Moens. 1997. Sentence extraction as a classification task. In *Proceedings of the ACL Workshop on Intelligent Text Summarization*, pages 58–65, Madrid.
Tomoeda, C. K., K. A. Bayles, D. R. Boone, A. W. Kaszniak, and T. J. Slauson. 1990. Speech rate and syntactic complexity effects on the auditory comprehension of Alzheimer patients. *Journal of Communication Disorders*, 23(2):151–161.
Toutanova, K., D. Klein, C. D. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180, Edmonton.
Watson, Caroline.M. 1999. An analysis of trouble and repair in the natural conversations of people with dementia of the Alzheimer's type. *Aphasiology*, 13(3):195–218.
Weiner, M. F., K. E. Neubecker, M. E. Bret, and L. S. Hynan. 2008. Language in Alzheimer's disease. *Journal of Clinical Psychiatry*, 69:1223–1227.
Wilson, T., J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, Ann Arbor, MI.
Wolters, Maria, Kallirroi Georgila, Johanna D. Moore, Robert H. Logie, Sarah E. MacPherson, and Matthew Watson. 2009. Reducing working memory load in spoken dialogue systems. *Interacting with Computers*, 21(4):276–287.
Young, S., M. Gasic, B. Thomson, and J. D. Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.
Yu, H. 2009. News summarization based on semantic similarity measure. In *Ninth International Conference on Hybrid Intelligent Systems, 2009. HIS '09*, volume 1, pages 180–183, Shenyang.
Zhang, N. L., S. S. Lee, and Zhang W. 1999. A method for speeding up value iteration in partially observable Markov decision processes. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 696–703, Stockholm.
Zhao, S., F. Rudzicz, L. G. Carvalho, C. Márquez-Chin, and S. Livingstone. 2014. Automatic detection of expressed emotion in Parkinson's disease. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP14)*, pages 4813–4817, Firenze.
Zhou, Luke, Kathleen C. Fraser, and Frank Rudzicz. 2016. Speech recognition in Alzheimer's disease and in its assessment. In *Proceedings of Interspeech 2016*, pages 1948–1952, San Fancisco, CA.