A Method for Analysis of Patient Speech in Dialogue for Dementia Detection

Saturnino Luz, Sofia de la Fuente, Pierre Albert
Usher Institute of Population Health Sciences & Informatics
Edinburgh Medical School
The University of Edinburgh, Scotland, UK
{s.luz,sofia.delafuente,pierre.albert}@ed.ac.uk

Abstract

We present an approach to automatic detection of Alzheimer's type dementia based on characteristics of spontaneous spoken language dialogue consisting of interviews recorded in natural settings. The proposed method employs additive logistic regression (a machine learning boosting method) on content-free features extracted from dialogical interaction to build a predictive model. The model training data consisted of 21 dialogues between patients with Alzheimer's and interviewers, and 17 dialogues between patients with other health conditions and interviewers. Features analysed included speech rate, turn-taking patterns and other speech parameters. Despite relying solely on content-free features, our method obtains overall accuracy of 86.5%, a result comparable to those of state-of-the-art methods that employ more complex lexical, syntactic and semantic features. While further investigation is needed, the fact that we were able to obtain promising results using only features that can be easily extracted from spontaneous dialogues suggests the possibility of designing non-invasive and low-cost mental health monitoring tools for use at scale.

Keywords: Dementia diagnosis and prediction, Alzheimer's disease, dialogue analysis, speech features, vocalisation graphs, content-free analysis.

1. Introduction

Research into early detection of Alzheimer's disease (AD) has intensified in the last few years, driven by the realisation that in order to implement effective measures for secondary prevention of Alzheimer's type dementia (ATD) it may be necessary to detect AD pathology decades before a clinical diagnosis of dementia is made (Ritchie et al., 2017). While imaging (PET, MRI scans) and cerebrospinal fluid analysis provides accurate diagnostic methods, there is an acknowledged need for alternative, less invasive and more cost-effective tools for AD screening and diagnostics (Laske et al., 2015). A number of neuropsychological tests have been developed which can identify signs of AD with varying levels of accuracy (Mortamais et al., 2017; Ritchie et al., 2017). However, the proliferation of technologies that enable personal health monitoring in daily life points towards the possibility of developing tools to predict AD based on processing of behavioural signals.

Speech is relatively easy to elicit and has proven to be a valuable source of clinical information. It is closely related to cognitive status, having been used as the primary input in a number of applications to mental health assessment. It is also ubiquitous and can be seamlessly acquired. In recent years, combinations of signal processing, machine learning, and natural language processing have been proposed for the diagnosis of AD based on the patient's speech and language (Fraser et al., 2016). Models built on phonetic, lexical and syntactic features have borne out the observation that these linguistic processes are increasingly affected as the disease progresses (Kirshner, 2012). However, most machine learning research in this area has employed either recorded narrative speech (Lopez-De-Ipiña et al., 2012), or recorded scene descriptions (Luz, 2017; Fraser et al., 2016) collected as part of a neuropsychological assessment test, such as the Boston "cookie theft" picture description task (Becker et al., 1994).

In contrast to those methods, our approach employs spontaneous conversational data, exploring patterns of dialogue as basic input features. Content-free interaction patterns of this kind were first used in the characterisation of psychopathology by Jaffe and Feldstein (1970), who represented therapist-patient dialogues as Markov chains. Here, we build on these ideas to analyse patient data from the Carolina Conversations Collections (CCC) (Pope and Davis, 2011). We trained machine learning models on these data to differentiate AD and non-AD speech. This work is, to the best of our knowledge, the first to employ low-level dialogue interaction data (as opposed to lexical features, or data from narrations other forms of monologue) as a basis for AD detection on spontaneous speech.

2. Background

One of the greatest challenges facing developed countries, and increasingly the developing world, is the challenge of improving the quality of life of older people. In 2015, the First Ministerial Conference of the WHO on Global Action Against Dementia estimated that there are 47.5 million cases of this condition in the world. Cohort studies show between 10 and 15 new cases per each thousand people every year for dementia, and between 5 and 8 for Alzheimer's Disease. Prognosis is usually poor, with an average life expectancy of 7 years from diagnosis. Less than 3% diagnosed live longer than 14 years. Current statistics predict that the population aged over 65 is expected to triple between years 2000 and 2050 (World Health Organization and others, 2015). This will lead to structural and societal changes, accentuating what is already becoming a highly demanding issue for health care systems.

Dementia is therefore set to become a very common cause of disability which places a heavy burden on carers and patients alike. While there are currently neither a cure nor a way to entirely prevent the progress of the disease, it is hoped that a better understanding of language and communication patterns will contribute to secondary prevention. A characterisation of communication patterns and their relation to cognitive functioning and decline could be useful in the design of assistive technologies such as adaptive interfaces and social robotics (Wada et al., 2008). These technologies might help provide respite to carers, and stimulate cognitive, physical and social activity, which can slow disease progression and improve the patient's quality of life (Middleton and Yaffe, 2009). Collecting relevant real life observational data and assembly of prior and current knowledge (Wada et al., 2008) could lead to new effective and personalised interventions.

Assessing people's behaviour in natural settings might also contribute to earlier detection (Parsey and Schmitter-Edgecombe, 2013; Mortamais et al., 2017). Language impairment is a common feature of dementia, implying signs such as word-finding and understanding difficulties, blurred speech or disrupted coherence (American Psychiatric Association, 2000). Although language is a good source of clinical information regarding cognitive status, manual analysis of language by mental health professionals for diagnostic purposes is challenging and time-consuming. Advances in speech and language technology could help by providing tools for detecting reliable differences between patients with dementia and controls (Bucks et al., 2000), distinguishing among dementia stages (Thomas et al., 2005) and differentiating various types of dementia (Fraser et al., 2016).

Features such as grammatical constituents, vocabulary richness, syntactic complexity, psycholinguistics, information content, repetitiveness, acoustics, speech coherence and prosody, have been explored in conjunction with machine learning methods to identify Alzheimer's and other types of dementia through the patient's speech. This is not only because language is impaired in these patients, but also because language relies on other cognitive functions, such as executive functions, which allow us to interact in a sound and meaningful way. These functions are responsible for decision making, strategy planning, foreseeing consequences and problem solving, which are essential to successful communication, but are impaired by AD (Fraser et al., 2016; Marklund et al., 2009; Satt et al., 2013). Although hardly perceptible to the speakers themselves, patterns of impairment are thought to occur even in informal and spontaneous conversations (Bucks et al., 2000; Cohen and Elvevåg, 2014).

Our hypothesis in this paper is that people with an AD diagnosis will show identifiable patterns during dialogue interactions. These patterns include disrupted turn taking and differences in speech rate. These indices relate to the fact that, in general, patients with AD show poorer conversation abilities and their normal turn-taking is repeatedly interrupted. Therefore, we expect less conversational fluidity overall in the AD group dialogues, as compared to non-AD group. Our approach, which does not rely on transcription but only on speech-silence patterns and basic prosodic information, obtains levels of accuracy comparable to state-of-the-art systems that rely on more complex feature sets.

3. Related work

Potential applications of the kind of speech technology described in this paper include the development of interactive assistive technologies, and monitoring of users for signs of cognitive decline with a view to mitigating further decline. From the perspective of potential applications of automatic speech analysis to technology-assisted care, there is evidence (Rudzicz et al., 2014b) that it is psychologically more acceptable for a user to be aided by another person or a robot than from ambient sensors and devices which are unable to offer meaningful interaction. Therefore, the development of such assistive applications involves research on speech processing for natural conversations rather than scripted speech or monologues (Conway and O'Connor, 2016).

From the perspective of monitoring for early detection, it is known that AD leads to disruption of one's ability to follow dialogues, even in simple, routine interactions. At later stages of the disease, failure to perform meaningful interactions appears (Watson, 1999). This has a negative impact on tasks such as following instructions regarding household activities and medication, as well as preventing rewarding social interactions. Here, once again the focus should be on natural interaction data, as scripted talk cannot be compared to spontaneous conversation in terms of information richness and external validity of results (Kato et al., 2013). Over the last decades, different approaches have targeted early detection of AD on spontaneously generated data through automatic and non-invasive intelligent methods. Some of these approaches have focused on speech parameters analysis: automatic spontaneous speech analysis (ASSA), emotional temperature (ET), (Lopez-De-Ipiña et al., 2012), voiceless segments, and phonological fluency have been shown to explain significant variance in neuropsychological test results (García Meilán et al., 2012). These methods are not only non-invasive and free from side-effects, but also relatively cheap in time and in terms of resources. Another approach that rely on easily extracted acoustic features, such as the ones we propose in this paper, though not in dialogical or spontaneous speech settings is presented by Satt et Al. (2013). This approach extracts a number of voice features (voiced segments, average utterance duration, etc.) from recordings of picture description, sentence repetition, and repeated pronunciation of three syllables used in diadochokinetic tests in succession. The method achieves accuracy levels of over 80% in detection of AD and mild cognitive impairment (MCI).

Other approaches have used time-aligned transcripts and syntactic parsing, extracting speech features and using them for classifying healthy elderly subjects from subjects suffering AD or MCI, as well as other tasks. This classification has been done either by comparing impaired to healthy speech performance (speech quality in terms of lexicon, coherence, etc.), or by comparing classifier performance when only neuropsychological tests are included against performance when such tests are used together with speech features, generally with statistically significant improvements (Roark et al., 2011; Fraser et al., 2016).

Analysis performed on similar corpora provide good insight of the performances achieved using different features.

A first analysis (Fraser et al., 2016), based on a monologue corpus (DementiaBank), identified four different linguistic factors as main descriptors: syntactic, semantic, and information impairments, and acoustic abnormality. They achieved accuracy of up to 92.05% using full scale analysis of 25 features, selected amongst an original feature set of 370 features after extensive experimentation.

An analysis of the CCC corpus by Guinn et al (Guinn and Habash, 2012) used similar linguistic features. Unlike the work presented in this paper, Guinn's analysis was focused on the differences between interviewers and subjects in the subset of patients with AD. They achieved a combined accuracy of 75-79.5 % using decision trees, with a large discrepancy between AD (38-42 %) and non-AD (74-100 %) recognition accuracy.

Works on dialogue so far have identified features such as conversational confusion (AD increases confusion rates, and this relates to slower and shorter speech; (Rudzicz et al., 2014a), prosodic measures (Gonzalez-Moreira et al., 2015), and emotion (Devillers et al., 2005). These studies used machine learning methods (neural networks, Naïve Bayes, and random forests, respectively), reporting accuracy in the 70-90 % range. Although these results are promising, they are difficult to generalise. This is because they are primarily content dependent. That is, they employ lexical, and sometimes syntactic information, which present a number of potential disadvantages. The content of a conversation is likely to change greatly depending on whether a participant belongs to the control group or to the group with Alzheimer's Disease, especially if the conversational partner is their doctor. In addition, such content is difficult to acquire in spontaneous speech settings. Despite the advances in automatic speech recognition, recognition (word) error rates in unconstrained settings are still over 11%, even for fairly clear, telephone dialogues (Xiong et al., 2016). Another difficulty with these approaches is the fact that they are language-dependent, and therefore require building different models for different languages, which in the context of global mental health could be a major shortcoming. Therefore, these models should aim to be as content-independent as possible to be generalisable (Satt et al., 2013). In contrast to content-based approaches, our method focuses on the interaction patterns themselves, rather than on characteristics of the speech and language content as such.

4. Methods

4.1. Dataset

We have conducted our analysis using the Carolina Conversations Collection (Pope and Davis, 2011). The dataset is a digital collection of recordings of conversations about health, including both audio and video data, with corresponding transcriptions. The corpus consists of natural conversations involving an older person (over the age of 65) with a medical condition. Several demographic and clinical variables are also available, including: age range, gender, occupation prior to retirement, disease diagnosed, and level of education (in years). The interviewers were gerontology and linguistic students or researchers to whom the patients spoke at least twice a year. A unique alias was assigned to each patient to protect their identity.

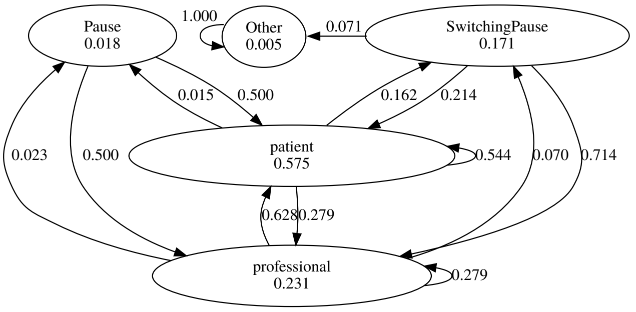

Figure 1: Vocalisation diagram for a patient dialogue.

Access to the data was provided after complying with the ethical requirements of the University of Edinburgh and the Medical University of South Carolina. In order to ensure that the results described here are reproducible we will provide, on request, the identifiers for the dialogues used in our experiments so that interested researchers can recreate our dataset upon being granted access to the CCC. The source code used for processing the data is available at a University of Edinburgh gitlab server[1].

For the research described here we selected a total of 38 patient dialogues: 21 patients had a diagnosis of Alzheimer's disease (15 females, 6 males), and 17 patients (12 females, 5 males) had other diseases (diabetes, cardiac issues, etc., excluding neuropsychological conditions), but not AD. These groups were selected for matching age ranges and gender frequencies so as to avoid statistical bias. The dataset also included time-aligned transcripts, which we did not use except for the computation of an alternative speech rate feature as described below.

4.2. Data Preparation

The speech data selected as previously described were preprocessed in order to generate *vocalisation graphs* — that is, Markov diagrams encoding the first-order conditional transition probabilities between vocalisation events and steady-state probabilities (Luz, 2013).Vocalisation events are classified as speech by either the patient or the interviewer/others, joint talk (overlapping speech), or silence events (also known as 'floor' events, which are further in the diagrams as pauses and switching pauses, according to whether the floor is taken by the same speaker or another speaker, respectively). An example of vocalisation graph is shown in Figure 1.

Vocalisation and pause patterns have been successfully employed in the analysis of dialogues in a mental-health context (Jaffe and Feldstein, 1970), segmentation (Luz and Su, 2010) and classification of dialogues, and more recently on characterisation of participant role and performance in collaborative tasks (Luz, 2013). Models that employ basic turn-taking statistics have also been proposed for dementia diagnosis (Mirheidari et al., 2016), though not in a systematic content-fee framework as in our proposed method.

The distributions of event counts according to vocalisation

[1] https://cybermat.tardis.ed.ac.uk/pial/CCCdataset

events is shown in Figure 2. It can be observed that patients with AD tend to produce more vocalisation events than their interviewers (and, consequently, produce more silence events). This is consistent with findings in the literature on language changes in AD (American Psychiatric Association, 2000).

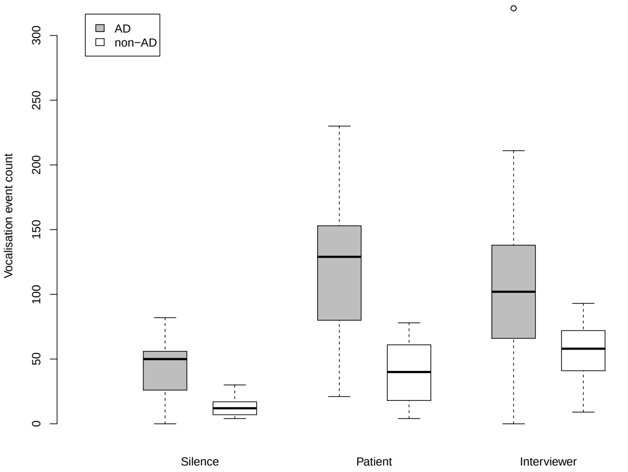

Figure 2: Distribution of vocalisation event counts for patients with and without AD in CCC dialogues.

Speech rate was estimated using De Jong's syllable nuclei detection algorithm (Jong and Wempe, 2009), which is an unsupervised method – that is, it can be applied directly to the acoustic signal, with no need of human annotation. However, as the audio quality of the CCC recordings is uneven, and as the dataset provides no gold standard against which one could assess syllable count, we decided to validate the use of De Jong's method against the time-stamped transcripts provided. Using these transcripts one could, in principle, estimate average words per minute (WPM) for individual utterances, as is sometimes done (Hayakawa et al., 2017). However, this method of measuring WPM based on transcription has a number of limitations. Words have variable length, and their articulation can vary greatly due to a number of speech-related phenomena, such as phonological stress, frequency, contextual predictability, and repetition (Bell et al., 2009). In order to mitigate these problems, we instead produced *speech rate ratio* estimates normalised through a speech synthesizer, employing the methods proposed by Hayakawa et al. Hayakawa et al. (2017). These estimates represent deviations from a "normalised" pace of 160 words per minute (WPM) synthesised using the MaryTTS system (Schröder and Trouvain, 2003). We therefore computed the ratio of the synthesised speech to the actual duration of the patient's speech. The speech rate ratio correlated well with the syllable per minute rate extracted using only the recorded audio ($\rho = 0.502$, $t(30) = 3.19, p = 0.003$) indicating that speech rate can be estimated with an acceptable level of reliability through the unsupervised method, even in fairly noisy settings.

A Python script was employed to extract basic speaker turn time stamps, speaker role information, and transcriptions from the original XML-encoded CCC data. The resulting data were then processed using the R language in order to detect silence intervals, and categorise turn transitions and pause events.

Table 1: Descriptive statistics on dialogue turn-taking (duration given in seconds).

Feature	non-AD	AD
Dialogue duration	4107.3	7628.4
Dialogue duration TTS	7618.8	7618.8
Avg turn duration	97.3	255.8
Total turn duration	1654.3	4348.3
Norm. total turn duration	3.0	4.1
Avg turn duration TTS	107.6	238.0
Total turn duration TTS	1829.7	4046.1
Norm. total turn duration TTS	3.0	4.2
Avg number of words	314.6	742.5
Total number of words	5348.0	12622.0
Avg words per minute	155.9	166.5

Some descriptive statistics on the dialogues can be seen in Table 1. These statistics include: average turn duration (how many seconds a participant speaks on average), total turn duration (how many seconds did the participant's turns lasted in total), normalised turn duration (the ratio of a participant's turn duration to the total duration of AD or non-AD dialogues, according the participant's class), number of words generated (total per class and on average per class' participant), and number of words per minute (average per class participant).

Contrary to our expectations, we did not observe a statistically significant difference between the speech rate in syllables per minute between patients with and without AD (Welch two sample t-test $t(30.5) = 1.15, p = 0.28$), even though the mean for non-AD ($M = 180.8$ syllables/min, $sd = 28.4$) was higher than that for patients with AD ($M = 168$ syllables/min, $sd = 35.6$).

Two alternative data representations were generated. The first (henceforth referred to as VGO) was based on the vocalisation graphs only. That is, VGO encodes the probabilities of each possible pair of transitions, including self-transitions, which tend to dominate Markov chains sampled, and the steady-state probabilities for each vocalisation event. The second form of representation (VGS) simply consists of the VGO with information about the participant's speech rate (mean and variance) added to the vocalisation statistics. With the exception of speech rate ratio, which necessitates transcription, all the information needed to build VGO and VGS instances can be extracted through straightforward signal processing methods.

4.3. Machine learning

The data instances in the two alternative representation schemes were annotated for presence or absence of Alzheimer's Disease (AD). A supervised learning procedure was employed in order to train classifiers to predict such annotations on unseen data.

We trained a boosting model (Schapire and Freund, 2014) using decision stumps (i.e. decision trees with a single split node) as weak learners. The training process consisted of 10 iterations whereby, for each training instance (x_i), a weak classifier $\hat{f}_m$ was fitted using weights on the data which were iteratively computed so that the instances misclassified in the preceding step had their weights increased by a factor proportional to the weighted training error. In this case class probability estimates $P(ad = 1|data)$ were used to compute these weights and to weigh the final classification decision (additive logistic regression) following the Real Adaboost algorithm (Friedman et al., 2000):

$$\hat{F}(x) \quad = \quad sign\left[\sum_{m=1}^{M} \hat{f}_m(x)\right] \qquad (1)$$

Classification performance was assessed through a 10-fold cross validation procedure. As the dataset is reasonably balanced, results were assessed in terms of accuracy, precision (the ratio of the number of true positives to the number of instances classified as AD), recall (or sensitivity, the ratio of true positives to the number of AD cases) and F_1 score (the harmonic mean of precision and recall). Micro (μ) and macro (M) averages for these scores are given by taking means over the entire set of classification decisions and over individual classifiers respectively, across the 10 folds. As the data set is fairly small, we also ran a leave-one-out cross validation (LOOCV) procedure to obtain better estimates of generalisation accuracy. This consisted of selecting one instance for testing, and building a classification model on the remaining instances, and iterating this procedure until all instances have been selected as testing instances. Macro averages are uninformative in LOOCV, so we only report overall accuracy figures for this procedure.

ROC curves showing the relationship between true positive and false positive rates as the classification threshold is varied were also plotted. Simulation was employed in order to smooth these ROC curves by running 10 rounds of 10-fold cross validation tests with a randomised selection of instances making up the hold-out sets.

5. Results

Our first approach, based on the VGO data representation scheme, produced promising results. Accuracy levels were well above the baseline, with overall accuracy reaching 81.1%, showing that turn taking patterns can provide useful cues to the detection of AD in dialogues. The results for the VGO-based classification are shown in Table 2. The corresponding ROC curve is shown in Figure 3.

Adding speech rate information (VGS representation) contributed to further enhancing AD detection, bringing the overall accuracy score to about 86.5%. Detailed evaluation metrics are shown in Table 3. The ROC curve for the VGS-based classification approach is shown in Figure 4. It can be seen that the addition of features for mean and variance of speech rate ratio over dialogues had the effect of improving classification trade-offs, particularly reducing the false positives while increasing the true positives at low threshold cut-offs.

Table 2: AD detection results for the VGO data representation scheme.

AD		non-AD	
Accuracy$_\mu$	0.812	Accuracy$_\mu$	0.714
Precision$_\mu$	0.765	Precision$_\mu$	0.769
Recall$_\mu$	0.812	Recall$_\mu$	0.714
$F_{1,\mu}$	0.788	$F_{1,\mu}$	0.741
Precision$_M$	0.667	Precision$_M$	0.792
Recall$_M$	0.722	Recall$_M$	0.729
$F_{1,M}$	0.685	$F_{1,M}$	0.721
Overall accuracy (LOOCV): 0.811			

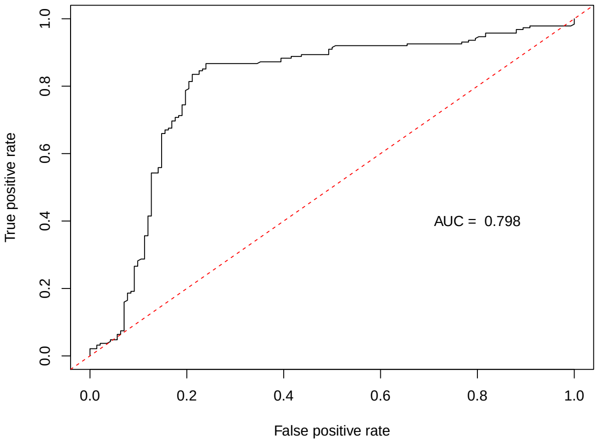

Figure 3: ROC curve for VGO-based classifiers.

For comparison we ran the same testing procedure using some of the other classifiers employed in the literature reviewed in section 3., namely, logistic regression, naïve Bayes (Gaussian kernel), decision trees (C4.5 algorithm), SVM trained using sequential minimal optimisation, with a polynomial kernel (Platt, 1998), and random forests (Breiman, 2001), Weka implementation (Hall et al., 2009). The overall (LOOCV) accuracy figures are shown in Table 4. There is little difference in performance between our chosen method (Real Adaboost) and other methods used in the literature, except for logistic regression, which underperforms the machine learning methods. Real Adaboost slightly outperforms SVM and random forests classifiers, and matches C4.5 decision trees, with a slight advantage over the latter on the target AD class ($F_m = 0.878$ vs. $F_m = 0.872$).

Although there is considerable room for improvement upon this level of classification performance, the levels obtained with these simple models are comparable to the accuracy

Table 3: Results for the VGS data representation scheme.

AD		non-AD	
Accuracy$_\mu$	0.882	Accuracy$_\mu$	0.769
Precision$_\mu$	0.833	Precision$_\mu$	0.833
Recall$_\mu$	0.882	Recall$_\mu$	0.769
$F_{1,\mu}$	0.857	$F_{1,\mu}$	0.800
Precision$_M$	0.796	Precision$_M$	0.708
Recall$_M$	0.833	Recall$_M$	0.708
$F_{1,M}$	0.811	$F_{1,M}$	0.700
Overall accuracy (LOOCV): 0.865			

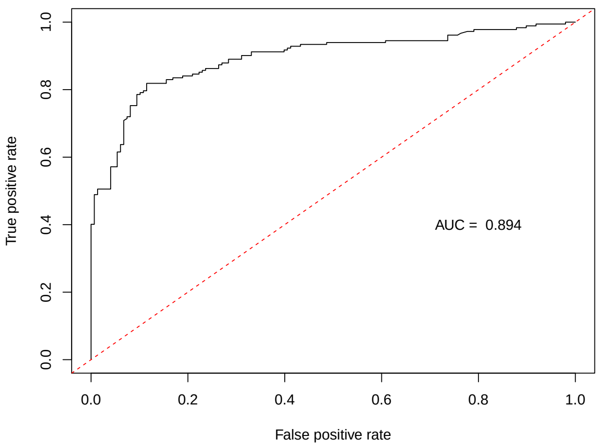

Figure 4: ROC curve for VGS-based classifiers.

Table 4: Compared accuracy results obtained with different classification algorithms, on VGS-based datasets.

Classification method	Accuracy (LOOCV)
Logistic regression	75.7%
Real Adaboost	86.5%
Decision trees	86.5%
SVM	83.7%
Random forests	81.1%

of approaches that employ more detailed linguistic information, which are presumably harder to acquire in everyday conversational situations, as they would involve a level of speech recognition accuracy which is beyond the capabilities of current systems for spontaneous speech in noisy environments.

6. Conclusion and Further Work

Dementia prevention and life quality in elderly care are important societal challenges. Automatic detection of signs of AD in speech can provide useful tools for the design of technologies for care-giving and cognitive health monitoring to help address these challenges.

This paper presented initial results of a new method to automatically recognise the first signs of disrupted communication using dialogue features. This method obtained an overall accuracy of 0.83, with a micro F-measure of 0.83 and a macro F-measure of 0.76 on the classification of patients as "AD" and "non-AD". Although it is difficult to compare these results directly to related works (Fraser et al., 2016; Guinn and Habash, 2012), our accuracy figures are situated within a similar range, 0.70-0.80, with a smaller discrepancy between the classification of the two groups, while relying on features that can be more robustly extracted from spontaneous speech.

Thanks to the increasingly important role of social technology, longitudinal studies may become richer in terms of the amount of variables measured, frequency of measurements and places where measures are taken (living settings), allowing for larger datasets. As more data are gathered in natural settings, we expect to obtain more reliable and generalisable results.

There are several linguistic parameters that are promising for the assessment of cognitive functioning. In current approaches, these features have been typically extracted from data collected through structured interviews, storytelling or picture descriptions. The work presented here contributes a new perspective to feature extraction by focusing on spontaneous dialogues. Dialogue processing provides a convenient framework for the analysis of natural conversations, in which readily available predictors, such as turn taking behaviour, have already yielded satisfactory results. We plan to further analyse verbal and non-verbal parameters to obtain a better characterisations of AD in order to infer neurosychological assessment results through speech and language processing, and subsequently to combine such features with actual neuropsychological evaluations and other relevant variables, building accurate models to achieve detection of AD at the time of onset.

The data set used in the present study has some limitations. Due to its constraints, the study was performed on a restricted subset of 21+17 sessions. In addition, the interview setting includes a degree of bias, as the interviewer's objective is to get the patient to perform a certain task (e.g. description of a picture, driving the discussion) therefore influencing the patient's speech. In order to mitigate these limitations, we plan to collect further data in more spontaneous dialogue in the near future.

7. Acknowledgements

Sofia De la Fuente and Pierre Albert are supported by the Medical Research Council (MRC). The authors would like to acknowledge Charlene Pope and Boyd H. Davis, from the Medical University of South Carolina, who host the Carolinas Conversation Collection, for providing access to the dataset and help in completing the required procedures.

8. Bibliographical References

American Psychiatric Association. (2000). Delirium, dementia, and amnestic and other cognitive disorders. In American Psychiatric Association, editor, *Diagnostic and Statistical Manual of Mental Disorders, Text Revision (DSM-IV-TR)*, chapter 2. Arlington, VA, 4^{th} edition.

Becker, J., Boiler, F., Lopez, O., Saxton, J., and McGonigle, K. (1994). The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. 60(1):92–111.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Bucks, R., Singh, S., Cuerden, J., and Wilcock, G. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91.

Cohen, A. S. and Elvevåg, B. (2014). Automated Computerized Analysis of Speech in Psychiatric Disorders. *Current opinion in psychiatry*, 27(3):203–209.

Conway, M. and O'Connor, D. (2016). Social media, big data, and mental health: Current advances and ethical implications. *Current Opinion in Psychology*, 9:77–82.

Devillers, L., Vidrascu, L., and Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422.

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease*, 49(2):407–422, October.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, April.

García Meilán, J. J., Martínez-Sánchez, F., Carro, J., Sánchez, J. a., and Pérez, E. (2012). Acoustic Markers Associated with Impairment in Language Processing in Alzheimer's Disease. *The Spanish Journal of Psychology*, 15(2):487–494.

Gonzalez-Moreira, E., Torres-Boza, D., Kairuz, H., Ferrer, C., Garcia-Zamora, M., Espinoza-Cuadros, F., and Hernandez-Gómez, L. (2015). Automatic prosodic analysis to identify mild dementia. *BioMed Research International*.

Guinn, C. I. and Habash, A. (2012). Language analysis of speakers with dementia of the alzheimer's type. In *AAAI Fall Symposium: Artificial Intelligence for Gerontechnology*, pages 8–13.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hayakawa, A., Vogel, C., Luz, S., and Campbell, N. (2017). Speech rate comparison when talking to a system and talking to a human: A study from a speech-to-speech, machine translation mediated map task. In *Proc. Interspeech 2017*, pages 3286–3290.

Jaffe, J. and Feldstein, S. (1970). *Rhythms of dialogue*. Personality and Psychopathology. Academic Press, New York.

Jong, N. H. d. and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390, May.

Kato, S., Endo, H., Homma, A., Sakuma, T., and Watanabe, K. (2013). Early detection of cognitive impairment in the elderly based on Bayesian mining using speech prosody and cerebral blood flow activation. *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2013:5813–6.

Kirshner, H. S. (2012). Primary Progressive Aphasia and Alzheimer's Disease: Brief History, Recent Evidence. *Current Neurology and Neuroscience Reports*, 12(6):709–714.

Laske, C., Sohrabi, H. R., Frost, S. M., de Ipiña, K. L., Garrard, P., Buscema, M., Dauwels, J., Soekadar, S. R., Mueller, S., Linnemann, C., Bridenbaugh, S. A., Kanagasingam, Y., Martins, R. N., and O'Bryant, S. E. (2015). Innovative diagnostic tools for early detection of alzheimer's disease. *Alzheimer's & Dementia*, 11(5):561–578.

Lopez-De-Ipiña, K., Alonso, J., Solé-Casals, J., Barroso, N., Faundez, M., Ecay, M., Travieso, C., Ezeiza, A., and Estanga, A. (2012). Alzheimer disease diagnosis based on automatic spontaneous speech analysis. In *Proceedings of the 4th International Joint Conference on Computational Intelligence*, pages 698–705.

Luz, S. and Su, J. (2010). The relevance of timing, pauses and overlaps in dialogues: Detecting topic changes in scenario based meetings. In *Proceedings of INTERSPEECH 2010*, pages 1369–1372, Chiba, Japan. ISCA.

Luz, S. (2013). Automatic Identification of Experts and Performance Prediction in the Multimodal Math Data Corpus through Analysis of Speech Interaction. *Proceedings of the 15th ACM on International conference on multimodal interaction, ICMI'13*, pages 575–582.

Luz, S. (2017). Longitudinal monitoring and detection of Alzheimer's type dementia from spontaneous speech data. In *Computer Based Medical Systems*, pages 45–46. IEEE Press.

Marklund, P., Sikström, S., Bååth, R., and Nilsson, L. G. (2009). Age effects on semantic coherence: Latent Semantic Analysis applied to letter fluency data. *3rd International Conference on Advances in Semantic Processing - SEMAPRO 2009*, pages 73–76.

Middleton, L. E. and Yaffe, K. (2009). Promising strategies for the prevention of dementia. *Arch Neurol*, 66(10):1210–1215.

Mirheidari, B., Blackburn, D., Reuber, M., Walker, T., and Christensen, H. (2016). Diagnosing people with dementia using automatic conversation analysis. In *Proceedings of Interspeech 2016*, pages 1220–1224. ISCA.

Mortamais, M., Ash, J. A., Harrison, J., Kaye, J., Kramer, J., Randolph, C., Pose, C., Albala, B., Ropacki, M., Ritchie, C. W., and Ritchie, K. (2017). Detecting cognitive changes in preclinical Alzheimer's disease: A review of its feasibility. *Alzheimer's & Dementia*, 13(4):468–492.

Parsey, C. M. and Schmitter-Edgecombe, M. (2013). Applications of technology in neuropsychological assessment. *The Clinical neuropsychologist*, 27(8):1328–1361.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, et al., editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

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

Ritchie, K., Carrière, I., Su, L., O'Brien, J. T., Lovestone, S., Wells, K., and Ritchie, C. W. (2017). The midlife cognitive profiles of adults at high risk of late-onset alzheimer's disease: The PREVENT study. *Alzheimer's & Dementia*, 13(10):1089–1097.

Roark, B., Mithcell, M., Hosom, J.-P., Hollingshead, K., and Kaye, J. (2011). Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. *The New England journal of medicine*, 19(7):2081–2090.

Rudzicz, F., Chan Currie, L., Danks, A., Mehta, T., and Zhao, S. (2014a). Automatically Identifying Trouble-indicating Speech Behaviors in Alzheimer's Disease. In *16th International ACM SIGACCESS Conference on Computers & Accessibility*, pages 241–242.

Rudzicz, F., Wang, R., Begum, M., and Mihailidis, A. (2014b). Speech recognition in Alzheimer's disease with personal assistive robots. *Proceedings of the 5th Workshop on Speech and Language Processing for Assistive Technologies*, pages 20–28.

Satt, A., Sorin, A., Toledo-Ronen, O., Barkan, O., Kompatsiaris, I., Kokonozi, A., and Tsolaki, M. (2013). Evaluation of speech-based protocol for detection of early-stage dementia. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (August):1692–1696.

Schapire, R. E. and Freund, Y. (2014). *Boosting: Foundations and Algorithms*. The MIT Press, January.

Schröder, M. and Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377.

Thomas, C., Keselj, V., Cercone, N., Rockwood, K., and Asp, E. (2005). Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. *IEEE International Conference Mechatronics and Automation, 2005*, 3(February):1569–1574.

Wada, K., Shibata, T., Musha, T., and Kimura, S. (2008). Robot Therapy for Elders affected by Dementia. (August).

Watson, C. 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.

World Health Organization et al. (2015). First who ministerial conference on global action against dementia: meeting report, who headquarters, geneva, switzerland, 16-17 march 2015.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. (2016). Achieving human parity in conversational speech recognition. *CoRR*, abs/1610.05256.