interspeech 2008 incorporating SST 08

Speech as a means of monitoring cognitive function of elderly speakers

Shona D'Arcy [1], Viliam Rapcan[1], Nils Penard[2], Margaret E Morris[3], Ian H Robertson[2], Richard B Reilly[4]

[1] School of Electrical, Electronic and Mechanical Engineering, University College Dublin, Rep of Ireland

[2] Trinity College Institute of Neuroscience (TCIN), Trinity College Dublin, Rep of Ireland

[3]Digital Health Group, Intel Corporation, Beaverton, Oregon, USA

[4]Department of Electronic and Electrical Engineering, Trinity College Dublin, Rep of Ireland

shona.darcy@tcd.ie, viliam.rapcan@tcd.ie, penardn@tcd.ie, margaret.morris@intel.com, ian.Robertson@tcd.ie, richard.reilly@tcd.ie

Abstract

This study investigates the use of speech as an indicator of the onset of cognitive decline in the elderly. The analysis found features that correlate with the results of a clinical measure of cognitive function. Using a combination of temporal language features, such as pause and utterance duration, 76% classification accuracy was achieved. While no significant results were found for ASR experiments, vowel duration, on average, increased by 17% for subjects with cognitive impairment compared to those without. The results from this study introduce the concept of longitudinal studies into aging using speech as a window into cognitive function.

Index Terms: cognitive decline; elderly, temporal features

1. Introduction

The world's population is growing older. In the next 50 years, the number of older people is forecast to quadruple, growing from about 600 million to almost 2 billion people [1]. With the prospect of people living longer there has been concerted efforts to provide healthcare at appropriate times to allow the elderly to live as independently as possible, for as long as possible [2], [3]. Remote monitoring of older people in their homes may enable early detection of some of the problems associated with old age, such as propensity to falls and social isolation along with decline in cognitive function. Once Mild Cognitive Impairment (MCI) has set in it is usually too late to implement any cognitive training schema that could help.

The biggest limiting factor to independence in the elderly is impaired cognitive function and its consequences. Such consequences include: accident proneness (falls, burns, bruising and cuts), self-neglect (missed medication, poor nutrition, poor hygiene), loss of initiative, diminished repertoire of activities and low mood. The assessment of cognitive function is expensive and labour intensive and hence non-viable for all but a tiny fraction of the elderly. In addition the elderly is a population traditionally highly averse to using new technology, which makes technical solutions to the assessment of cognitive function more difficult. Speech may provide a cost effective, easily implementable means of monitoring and assessing cognitive function.

The disproportionate vulnerability of the brain's frontal lobe to aging means that attention and executive function predominate in cognitive impairment having major detrimental effects on memory, attention, planning, initiative, mood vigilance as well as self-, safety-, and environmental awareness. One of the standard clinic tools used by psychologists for assessing cognitive function is the Mini Mental State Examination (MMSE) [5]. This consists of a questionnaire that rates cognitive function out of 30. MMSE scores of less than 27 are generally considered as possibly cognitively impaired by clinicians.

Previous studies have found spoken language markers to be indicative of Mild Cognitive Impairment (MCI) [6]. Roark et al found standardized pause rate (the ratio of words uttered to the number of pauses uttered) to be statistically different for two sets of elderly speakers; using a Clinical Dementia Rating (CDR), the cohort was separated into those with MCI and those without. While other speech duration parameters were investigated in the study including verbal rate (number of words per second), phonation rate (proportion of total time spent speaking) and mean duration of pauses, none of these were found be significantly different for the two populations.
Speech has also been shown to be an indicator of other neuro-degenerative diseases, such as schizophrenia [7]. Stassen et al looked at temporal parameters extracted from read speech from known sufferers of the disease and matched controls. In this case the correlation with this type of cognitive impairment and temporal features such as mean pause duration is attributed to the negative symptoms of the disease. These include affective flattening, blunted affect emotional dullness, poverty of speech and psycho-motor retardation.

A recent study carried out by Lieberman et al [8], investigated audio recorded from climbers ascending Mount Everest. The effect of altitude on the cognitive function of climbers was used to simulate that experienced by astronauts in space. Astronauts' cognition is impacted by hypoxic and cosmic ray-induced insult to the brain, as well as degraded cognitive performance resulting from task difficulty. The authors found acoustic measures of temporal characteristics of speech can be used to monitor cognitive impairment.

This study focuses on identifying the onset of cognitive decline from features that can readily be extracted from speech. Experimentation includes investigating speech from a temporal feature perspective and from an acoustic perspective. Speech is a non-intrusive means of data collection and speech centred applications, particularly telephony based, can be implemented using technology that is familiar to the potential users.

2. Data

The subjects for this study are recruited from the TRIL clinic in St James hospital, Dublin [4]. Volunteers attending the clinic underwent a battery of cognitive tests, which include audio recordings of read and spontaneous speech tasks.

Initially each subject completed an MMSE, to provide a quantitative measure of his or her cognitive status. The corpus currently contains speech data from 87 patients from 62 to 92 years of age; 62% female and 48% male. There are 23 patients with MMSE scores of 26 or less (26%) there are a further 14 subjects with an MMSE score of 27 and the remainder of the corpus (57%) who have MMSE scores that are considered to be in the normal range.

2.1. Demographic information

Personal information relevant to the study was collected.

- Gender
- Handedness
- Age
- Existing medical condition (Y/N)
- Drinker (light/heavy/ex/never)
- Prone to falls (Y/N)
- Smoker (Y/N/ex)
- Years of education (none/primary/half secondary/ full secondary/third level/postgrad).

These characteristics may or may not be contributing factors in cognitive decline but it was considered important to gather this information. However, it is known that gender, age, smoking, and medical conditions affect individual acoustic characteristics. While years of education, IQ, cognitive function and social interaction intuitively have implications for speech and language fluency, their significance has yet to be quantitatively assessed.

The first quantitative measure of each subject is their MMSE score, these range from 24 to 30. Although the corpus collection did not target any specific cognitive population, patients with severe cognitive impairment were not included in the corpus, resulting in an uneven distribution of the MMSE scores. The corpus was divided into two groups: one contained patients with MMSE scores of 27 and below (low MMSE group), and the other contained patients with MMSE scores greater than 27 (High MMSE group).

2.2. Cognitive Battery

All of the patients participated in a thorough evaluation of their cognitive and mood status in St James's Hospital. A first subset of the neuropsychological tests measured their cognitive functions. The MMSE gave a general overview of their cognition and the NART provided a pre-morbid I.Q. Standard word list learning and recall (Word Recall: immediate and delayed), as well as Digit Span (forward and backward) tested the patients' memory. Finally, the executive functions of the patients were tested using Animal Naming (category fluency) [9], and an in-house task called Picture Taboo. Executive functions refer to a set of cognitive abilities generally associated with the frontal lobes. These functions are monitoring and supervising more automated behaviors. Sustained attention, inhibitory control and planning are examples of executive functions. This last task consisted of a set of five pictures, each presented with a pair of words. The patients had one minute to describe each of the pictures without using the associated words. This task tapped into the executive functions by forcing the patients to inhibit the use of the taboo words while they are looking for alternative descriptor words. In a second subset of tests, the CESD and the HADS measured the mood status of the patients.

2.3. Audio recordings

For the audio recording portion of the data collection each subject wore a head-mounted microphone [Sennheiser PC 20] connected to a laptop. All efforts were made to keep the microphone at the same distance from the mouth and to keep volume levels of recordings consistent.

The subjects were recorded while carrying out the Picture Taboo and animal naming tasks and reading the following texts;

- A list of words that contain vowels in consistent context, i.e. /h/ vowel /d/. The list includes 2 examples of 19 vowels (11 monothongs and 8 diphthongs).
- The ``Heidi passage'', which is a short paragraph, extracted from a children's story. This passage is considered to be emotionally neutral and has been used in similar studies [7].
- 20 Scribe sentences [10], these short sentences that have been designed to provide a wide coverage of all the phones in the English language. There are 460 sentences of varying length. A statistics tool box for Natural language processing, developed by Hu [11] was used to grade these sentences by their complexity. The Fog readability index (Eqn 1) is used to measure how complex a sentence is;

$$Fog = 0.4\left(\frac{\#words}{\#sentences}\right) + 100\left(\frac{\#complexwords}{\#words}\right) \quad (1)$$

The number of syllables in a word is used to decide if it is a complex word or not. Each subject read 10 sentences each of varying complexity; the order of the sentences is randomized so that a practice effect does not occur.

3. Temporal Parameters

The hypothesis for this set of experiments is that cognitive impairment affects speaking behaviors and voice sound characteristics and these can be measured through temporal features. Cognitive decline affects neurological processing and it is the time taken to process speech and generate language is longer for cognitively impaired speakers.

Stassen et al [7] investigated 45 hospitalized acute schizophrenic patients and matched controls. Based on speech analysis of subjects reading a fixed passage of text 85% of the patients and controls were correctly classified based on a set of 12 temporal speech variables. Roark et al [6] carried out a similar experiment to measure cognitive function in the elderly using pause frequency and grammatical complexity on spontaneous speech. However, the results are preliminary, specifically looking at pauses greater than one-second duration. In terms of spoken language this would be considered a long duration. Both of these studies employed manual transcription of the audio for feature extraction.

3.1. Syntactic temporal features

Similarly to Stassen [7], the following temporal and acoustic features were then extracted from the read Heidi passage and the spontaneous Picture Taboo task;

- The total number of pauses
- The mean pause duration
- The mean pause duration per second
- The mean utterance duration
- The total recording time

- The total length of pauses
- The total length of utterances

Pauses longer than 250ms were removed at the beginning and the end of each acoustic recording and all speech signals were inspected visually, so that disturbed parts could be removed before data analysis.

The first objective of this analysis is to identify what parts of the signal were speech and what parts were non-speech. An average amplitude for noisy portions of the audio was calculated and this value set as the threshold for voicing Only segments of audio that fell below this threshold for durations greater than 250ms were included, while segments below this threshold lasting less than 250ms were added to voiced parts of signal. Feature extraction was carried out using MATLAB executed processing algorithms.

The Linear Discriminant Analysis (LDA) classifier was trained using all features (the pause and utterance features). The models were then tested using the same test set used in the ASR experiments above. LDA is method used in statistics and machine learning to find the linear combination of features which best separate two or more classes of object or events. The resulting combination may be used as a linear classifier or, more commonly, for dimensionality reduction before later classification.

3.2. Results

10 subject Test set	Read Heidi	Picture Taboo
Total number of pauses Mean pause duration Mean pause duration/s	80%	60%
Total recording time Total length of utterances Mean utterance duration Total recording time (read speech)	70%	40%
All features	100%	90%

Table 1 LDA classification performance for test set using syntactic temporal features

All subjects test set	Read Heidi	Picture Taboo
Total number of pauses Mean pause duration Mean pause duration/s	79.07%	60.47%
Total recording time Total length of utterances Mean utterance duration Total recording time (read speech)	69.77%	58.14%
All features	76.74%	62.79%

Table 2 LDA classification performance for all female speakers using syntactic temporal features

Tables 1 shows the LDA classification performance for the test set used in the ASR experiments; this contained 10 subjects, 3 with low MMSE and 7 with high MMSE scores. Table 2 shows the LDA classification for a larger test set of 32 female speakers. The first observation is that the features extracted from the read Heidi passage perform better than the spontaneous speech task; this is not surprising given the constrained nature of this task. Speakers are also more likely to be comfortable with the reading task compared to the unusual Picture Taboo task. It is also clear that it is the pause features that are more discriminative than the utterance features.

Another important result from this is the number of subjects with an MMSE score of 27 who have been mis-classified. Taking the pause feature results, of the 8 examples of subjects with an MMSE score of 27, 37% of these were classified as 'high'. While all 5 of the lower MMSE subjects are correctly classified and 16% the higher MMSE subjects are classed incorrectly. In the smaller test set it is consistently the subject with an MMSE of 27 that is mis-classified. These results support that a MMSE score below 27 is a likely cut-off point for cognitive impairment.

3.3. Acoustic Temporal Features

Leiberman et al measured vowel durations in speech generated by climbers at high altitude; the climbers were carrying out contextual learning tests that reflect hippocampal function in the brain. The hippocampus forms a part of the limbic system in the brain and plays a part in long-term memory and spatial navigation. This is particularly relevant for our cohort of elderly speakers as the hippocampus is one of the first regions of the brain to suffer damage in Alzheimer's disease [12]. They found increased vowel duration under these conditions and a hit rate in discriminating impaired from unimpaired performance on Mini-Cog tests of working memory or vigilance of 85%.

Using the /H vowel D/ words recorded, vowel alignment was found by implementing a forced alignment schema using HTK. The whole database was used to train 3-state monophone HMMs, with the exception of the /hh/ vowel which was reduced to a 1 state HMM (hand checking the resultant MLF file found a large improvement in alignment).

3.4. Results

Average vowel durations for 18 monophones were extracted from the phonetic transcription of 16 subjects, 8 with MMSE scores of 27 and less and 8 with MMSE score of 28-30. Applying a statistical significance test (T-test) to these results found the distributions of durations for 7 vowels to be significantly different for the two groups. The duration of individual vowels were consistently found to be longer for the group with a lower MMSE than the group with the higher MMSE, Table 3. The average vowel duration (over all vowels) was 17% longer for the lower MMSE group compared to the higher MMSE group.

vowel	MMSE <28	MMSE >28	vowel	MMSE <28	MMSE >28
aa	34.44	28.31	ia	28.09	22.94
ae	25.82	22.00	ih	25.40	21.38
ah	24.67	20.56	iy	31.30	29.44
ao	28.48	23.88	oh	25.85	24.13
aw	31.13	27.50	ow	29.20	25.56
ay	31.04	28.31	oy	32.04	29.31
ea	28.41	27.50	ua	30.32	26.80
eh	28.31	21.44	uh	27.23	20.13
ey	30.42	26.19	uw	29.73	24.38

Table 3: Comparing vowel durations (ms) for MMSE groups

The validity of including subjects with MMSE scores of 27 in the lower MMSE category has yet to be tested, the current data set makes this very difficult with so few examples of subjects below this threshold[1]. However there is a 13%

[1] Patients with MMSE scores of less than 24 were excluded after observation made by the ethics comittee after initial recordings session

increase in average vowel duration between the 6 examples of subjects with MMSE scores of 27 and subjects with MMSE scores of 28 and above. Comparing this group to those with an MMSE score of 26 and below a 9% increase in vowel duration is observed. These results suggest that subjects with MMSE scores of 27 are more similar to those patients with the lower MMSE scores.

4. Automatic speech recognition experiments

It is hypothesized that the dynamics of speech production differ between those with impaired cognitive function and those aging healthily. To test this hypothesis, a recognizer was trained for both speaker populations and performance compared for matched and mis-matched ASR scenarios.

Speaker Identification experiments were carried out using the Hidden Markov Model Toolkit (HTK) [13]. The data was divided into training sets corresponding to low and high MMSE score (27 was considered the threshold), each training set contained speech from 13 subjects reading 22 SCRIBE sentences and a set of 8-mixture Gaussian triphone models were trained using this data. Forced alignment of the read Scribe sentences by Viterbi decoding was carried out on the matched and mis-matched recognizers and the recognizer with the highest log likelihood taken as the correct classification.

4.1. Results

The test set contained 10 subjects, 3 with MMSE scores of 27 or less and 10 with MMSE scores of 28 and above. No significant classification was observed for this method, all of the low MMSE subjects and 4 of the 7 high MMSE subjects performed better for the mis-matched.

The small amount of data available is certainly a factor is the poor resolution of the models, particularly since the source of variation is very subtle. It is difficult for even a trained listener to determine whether a person appears cognitively impaired from only their speech. However using ASR for speaker dependent monitoring is anticipated to provide improved results, by looking for changes in an individual's speech.

5. Conclusions

The results presented in this paper suggest that there are elements of speech that can be used to monitor cognitive decline in the elderly.

Although the current ASR experiments have not shown statistically significant results the data collection is on going, and further experiments are currently assessing the use of ASR for quantifying speech insertions and deletions and non-speech insertions in read speech.

The classification experiments based on the temporal characteristics provide a more reliable method of classifying speakers in terms of high and low MMSE scores and certainly suggests that it is what is *not* said and not what *is* said that is an important feature of speech from people who are cognitively declining.

Both these experiments rely on the assumption that people with MMSE scores above 27 are considered cognitively normal and those with an MMSE of 27 and below are starting to develop some aspect of cognitive decline. It is generally accepted that the threshold of 27 is not an absolute threshold, and it is this group of people where their cognitive rating is most likely to be impacted by other factors such as age, years of education and perhaps ones job (be it skilled or non skilled) can affect literacy.

While there has been some work done to monitor cognitive [14] status of the elderly using speech and/or language, this usually requires expert interviews that must be carried out individually as well as hand analysis of the data. These can be very time consuming and intrusive. The methods developed in this study allow the automation of this process by focusing on speech parameters that can be reliably and automatically extracted and analyzed making them attractive for use in real time applications.

Cognitive decline is progressive, so in order to properly assess cognitive function and its degradation a longitudinal study is imperative. This type of study will allow intra-speaker comparisons of features which we know to be indicative of cognitive decline. The methods developed in this study offer the basis for a longitudinal study of cognitive function in older people. There are currently 2 longitudinal studies of the elderly in the design phase; the experiments presented in this paper will be replicated for the data collected.

6. Acknowledgments

We would like to thank all the clinical and TRIL staff at St James Hospital, Dublin for facilitating this study and in particular Ms Kate O'Sullivan and Ms Chiara Besani for data collection. The authors also acknowledge the TRIL Centre, www.trilcentre.org for funding this study.

7. References

[1] United Nations. Report of the Second World Assembly on Aging. Madrid, Spain: United Nations, April 8--12, 2002.

[2] Goldman D P, Baoping S, Gaber A M, Battacharya J, Hurd M, Joyce G F, Lakdawalla D N, Panis C, Shekelle P G. "Consequences of Health Trends and Medical Innovation For the Future Elderly", Health Affairs, Web exclusive, Sept. 2005

[3] Purser J L, Weinberger M, Cohen H J, Pieper C F, Morey M C, Li T, Williams G R, Lapuerta P. "Walking speech predicts health staus and hospital costs for frail elderly male veterans" Journal of rehabilitation Research and Development, 42:535-546, July, 2005

[4] www.trilcentre.org

[5] Folstein, M., Folstein, S., & McHugh, P. R. ``Mini-Mental State: a practical method for grading the cognitive state of patients for the clinician." Journal of Psychiatric Research, 12, 189-198, 1975

[6] Roark, B., Hosom, J. P., Mitchell, M., and Kaye, J. A., "Automatically Derived Spoken Language Markers for Detecting Mild Cognitive Impairment," in Proceedings of the 2nd International Conference on Technology and Aging (ICTA), Toronto, Canada, Jun. 2007.

[7] Stassen H.H., Albers M., Puschel J., Scharfetter C., Tewesmeier M. Woggon B. ``Speaking behavior and voice sound characteristics associated with negative schizophrenia", Journal of Psychiatric Research, Vol 29, pp 277-296, July 1995

[8] Lieberman P, Morey A, Hochstadt J, Larson M, Mather S. ``Mount Everest: a space analogue for speech monitoring of cognitive deficits and stress". Aviat Space Environ Med. Jun 2005

[9] Spreen, O. Strauss, E. " A Compendium of Neuropsychological Tests", Oxford University Press : New York, 1991

[10] SCRIBE (1989). SCRIBE (Spoken Corpus Recordings in British English) Manual. ttp://www.phon.ucl.ac.uk/resource/scribe/scribe-manual.htm

[11] Hu C, "Text Statistics Toolbox for Natural Language Processing", 2003

[12] Kennard, M. L., Feldman, H., Yamada, T., and Jefferies, W. A. (1996) Serum levels of the iron binding protein p97 are elevated in Alzheimer's disease. *Nat. Med.* **2**, 1230–1235

[13] Hidden Markov Model Toolkit (HTK), Cambridge University Engineering Department (CUED), http://htk.eng.cam.ac.uk/.

[14] Ancelin M L, Artero S, Portet F, Dupuy A,Touchon J, Ritchie K, "Non-degenerative mild cognitive impairment in elderly people and use of anticholinergic drugs: longitudinal cohort study". BMJ, 332:455-459, 25 February, 2006