INTERSPEECH 2009 BRIGHTON

The Use of Telephone Speech Recordings for Assessment and Monitoring of Cognitive Function in Elderly People

Viliam Rapcan[1], *Shona D'Arcy* [1], *Nils Penard*[2], *Ian H. Robertson*[2], *Richard B. Reilly*[1,2]

[1] Trinity Centre for Bioengineering, Trinity College Dublin, Ireland
[2]Trinity College Institute of Neuroscience, Trinity College Dublin, Ireland

rapcanv@tcd.ie, shona.darcy@tcd.ie, penardn@tcd.ie, ian.robertson@tcd.ie, richard.reilly@tcd.ie

Abstract

Cognitive assessment in clinic represents time consuming and expensive task. Speech may be employed as a means of monitoring cognitive function in elderly people. Extraction of speech characteristics from speech recorded remotely over a telephone was investigated and compared to speech characteristics extracted from recordings made in controlled environment. Results demonstrate that speech characteristics can be, with little changes in feature extraction algorithm, reliably (with overall accuracy of 93.2%) extracted from telephone quality speech. With further development of a fully automated IVR system, an early screening system for cognitive decline may be easily realized.

Index Terms: cognitive function, speech analysis, telephone recording, remote assessment and monitoring

1. Introduction

Deficits in cognitive function do not appear over night. Cognitive decline occurs gradually over varying periods of time. Regular monitoring of an aging person can identify the onset of cognitive decline and facilitate intervention that could address and hopefully arrest this decline. However while this monitoring in a clinic environment is highly desirable, it is both very expensive and time consuming and needs to be implemented by a trained expert.

Speech has been shown to be an indicator of certain brain disorders, such as schizophrenia. Stassen et al [1] investigated temporal parameters extracted from read speech recorded from 42 known sufferers of the disease and 42 matched controls. Correlation was found between temporal characteristics of speech (mean pause duration) and the negative symptoms of the disease, including affective flattening, blunted affect emotional dullness, poverty of speech and psycho-motor retardation. In a study of Mild Cognitive Impairment (MCI), Roark et al [2] demonstrated that standardized pause rate (the ratio of words uttered to the number of pauses in sample) to be statistically different between elderly speakers with and without MCI. A study carried out by D'Arcy et al [3] applied a similar approach to a cohort of elderly speakers with different levels of cognitive impairment. This study investigated temporal features of read and spontaneous speech along with several standard clinical measures of cognitive function. Pause and utterance duration were found to correlate with cognitive impairment and provide an indicators of a speaker's ability to read a sentence and continue the logical train of thought to its conclusion.

Using speech, remotely recorded over a telephone, may provide an unobtrusive, time saving and cost reducing way of assessment and monitoring of cognitive function. The study reported in here presents changes needed to be performed on the feature extraction algorithm for reliable extraction of speech characteristics form telephone recordings and results of applying a feature extraction algorithm to speech recorded over the telephone.

2. Methods

2.1. Participants

The subjects were recruited from the St. James's Hospital, Dublin, Ireland as part of a larger study on the elderly - Technology Research for Independent Living (TRIL) [4]. During a large battery of cognitive tests subjects were recorded reading aloud a fixed passage of text. The passage was an extract from a children's story (Heidi) contained 390 words and took on average 3-4 minutes to read completely. All audio files were recorded in a quiet room on a laptop via an external soundcard. Direct digital 16-bit sampling was used, at a sampling rate of 44.1 kHz. Speech was recorded in an uncompressed format. All audio files were high-pass filtered at 80Hz to improve signal-to-noise ratio.

Later, in the home deployment study, the recordings were moved from the clinic environment to participant's homes. Cognitive interviews were recorded with subjects, who had undergone the clinic assessment, in their own homes over the phone. Each participant was called by a research assistant and the conversation recorded onto a laptop via a USB recorder to which the phone line was connected. The recordings were sampled at 8 kHz, with 16-bit resolution and CCITT μ-LAW compression.

To assess cognitive function, all participants underwent a series of neuropsychological tests. Mini Mental State Examination (MMSE) [5] which gave a general overview of their cognition. The NART provided a pre-morbid I.Q. Participants' memory was tested using the standard word list learning, immediate and delayed word recall and Digit Span (forward and backward) test. Finally, the executive functions of the participants were tested using Animal Naming task. [6]

Recordings of the text passage in the clinic and at home are available for 19 subjects, allowing comparison of clinic and home recordings.

2.2. Feature extraction algorithm for high quality speech

Initial processing of the recorded audio data requires segmentation into speech and non-speech sections. Non speech segments were considered silences, breaths and other non-speech artifacts such as clicks, knocks and coughs.

2.2.1. Speech/Non-speech threshold estimation

A Speech/Non-speech threshold is estimated at the beginning of feature extraction process, and is subsequently employed to separate the audio data to speech and non-speech samples.

The stages for Speech/Non-speech threshold estimation are as follows.

1) Full-wave rectification is performed on the audio signal. Full-wave rectification is defined as

$$x(n) = abs(x(n)) \quad (1)$$

2) The rectified signal is divided into non-overlapping frames of 50ms duration and the energy is calculated in each frame.

3) Based on observations of the recordings of the text passage, typically at least 15%t of the frames represents no speech samples, 15% of frames with the lowest value of energy are selected for Speech/Non-speech Threshold estimation and maximum amplitude value in each of these frames is stored.

4) Finally, the Speech/Non-speech Threshold is calculated as a mean value of the stored maximum amplitude values.

2.2.2. Gradient threshold

The hypothesis of this study is that pauses are employed by speakers to deal with the cognitive load of reading aloud. An increase in the number of pauses indicates that the speaker needs extra time to process their thoughts, in order to fulfill the task in front of them (in this case a reading task). The duration of pauses is an indication of the cognitive load being experienced by the subject.

Breath sounds can be considered a normal part of speech, but breaths can be elongated by subjects attempting to fill pauses while trying to process their current cognitive load. It is for this reason that classing breaths as pauses may lead to a more discriminative feature of cognitive function.

In order to do this they must first be identified. If one inspects the amplitude envelope of breath sounds and compares them to the amplitude envelope of speech sounds, a difference in gradient can be observed. See Figure 1.

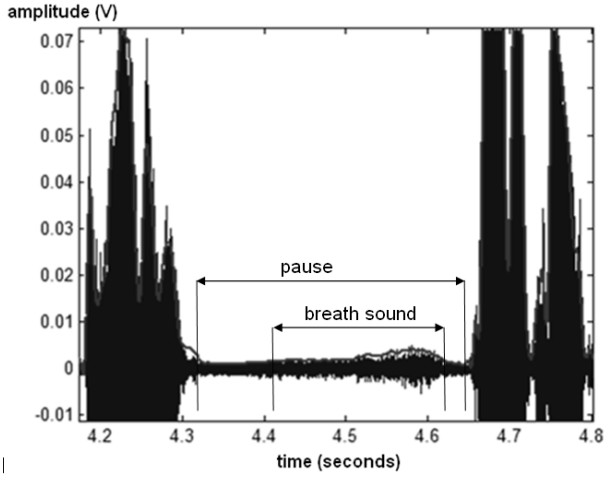

Figure 1: *Difference in amplitude envelope between breath sound and speech signal*

2.2.3. Estimation of the Reference Value

The Gradient Threshold differs for each speaker, and so a Reference Value was calculated which can be employed to estimate this Gradient Threshold. This Reference Value was estimated experimentally from randomly selecting 10 speech files from all available recordings. For every file the Gradient Threshold which detected all breath sounds in the recording was estimated. The Reference Value, to be used for all speech recordings, was calculated by dividing the Gradient Threshold by the Speech/Non-Speech Threshold value for that particular recording. Finally, the Reference Value was calculated as an average value of all 10 recording's reference values.

2.2.4. Breath sounds detection

Initial identification of breath sounds is based on an amplitude envelope detection schema. The stages for Breath Sounds Detection estimation are as follows:

1) Estimating the amplitude envelope of the waveform is calculated by performing a Hilbert transform and full-wave rectification on the audio signal. The sampling rate is then decreased by decimation of the signal and the signal is low-pass filtered at 33 Hz.

The original sampling frequency of 44.1 kHz requires the signal to be decimated by a factor of 30 in order to achieve a smooth envelope.

2) The gradient between each adjacent sample of the envelope is calculated. The additional advantage of measuring envelope gradients rather than an absolute amplitude values is that if one changes the volume level during recording it will change the value of amplitude, however the value of difference between two neighbouring samples will not be significantly affected.

3) The two conditions that must be satisfied so that the sample could be marked as a breath sample are:

a) The value of the gradient has to be below the Gradient Threshold.

b) False positives can arise from individual gradient values (particularly at the peak of envelopes) being less than the Gradient Threshold. The amplitude value of the sample must be below 120% of the Speech/Non-Speech Threshold to be considered a breath sound.

The Speech/Non-Speech Threshold is calculated on the segments of the smallest amplitude; however breath sounds have higher amplitude than the mean amplitude of noise and thus 120% was found experimentally to compensate for this difference.

4) Breath sounds are typically longer than 100ms [7], thus segments identified as possible breath sound shorter than 100ms are considered to be speech segments.

2.2.5. Edge detection algorithm

Edge detection is the process of maximizing temporal resolution in order to find the boundaries of breath sounds.

The stages for Edge Detection are as follows;

1) Breath sound detection, as described above does not always detect the entire breath sound and therefore, in the first step of edge detection, the breath sound segment is extended by 200ms on both sides. The extended segment is split into 10ms frames, in which the energy is calculated.

2) Energy contour is estimated by calculating the log energy in every frame. The energy of a discrete time signal [8] is defined as

$$E_x \overset{def}{=} \sum_{n=-\infty}^{\infty} |x(n)|^2 \quad (2)$$

Where, log energy is defined as

$$E_{\log} = \log_{10}(E_x) \quad (3)$$

A three-point running average filter is applied to smooth the energy contour.

3) A maximum energy value (Emax) is estimated over the original (non-extended) breath sound segment.

4) A minimum energy value (Emin) is estimated over the extended breath sound segment.

5) The Energy range (Erange) is calculated as absolute value of Emax – Emin.

6) Local minimums are located over the Energy contour and non-significant minimums are excluded. A minimum is considered non-significant if the difference between Emax and the energy value of minimum is less than 70% (found by experiment) of Erange.

7) The last step new edges of the breath segment are calculated based on the position of significant minimums closest to the position of the maximum of the original breath segment.

After edge detection, the breath segments are muted (replaced with zeros) in the recording and the recording is passed for the feature extraction phase.

2.2.6. Feature extraction

Based on the hypothesis that pause and utterance durations are indicators of one's ability to read a sentence and continue the logical train of thought to its conclusion, the following seven features were extracted from speech;

- Number of pauses
- Mean pause duration
- Proportion of recording in silence
- Mean utterance duration
- Total recording time
- Total length of pauses
- Total length of utterances

The result of this process is a segmentation of the audio recording into speech or non-speech. This segmentation provided pause onset and offset timing, enabling calculation of the number and duration of pauses. Analogously the utterance duration can be extracted. Proportion of silence is an indication of the proportion of non-speech in the recording.

2.3. Feature extraction algorithm for telephone speech

The telephone recordings differ from clinic recordings in sampling frequency, noise level, have variable volume level and contain more perturbations in form of "clicks", "knocks". As a result of these differences, for telephone recordings two changes in feature extraction algorithm were required.

Increased number of perturbations (clicks, knocks) in the recordings caused feature extraction algorithm to misinterpret pauses as speech segments, therefore, as a first step, all telephone recordings were filtered with seven-point running average filter to smooth the signal which helped with suppressing of these unwanted effects. Further, different noise level between the clinic and telephone recordings required the Reference Value (see Section 2.2.3) to be re-estimated for the telephone recordings.

2.4. Assessment of performance of the feature extraction

The ability of the feature extraction algorithm to correctly detect pauses and speech in telephone quality recording was investigated. Performance was measured using sensitivity, specificity, positive predictivity, negative predictivity and the overall accuracy. These measures were calculated as per the definition of true positives (TP) - entire pause marked as pauses by the feature extraction algorithm, true negatives (TN) - speech segment marked as a speech segment, false positives (FP) - part or entire speech segment marked as pause, and false negatives (FN) - part or entire pause marked as speech segment. Where, performance metrics are defined as follows;

$$Overall\ Accuracy = (TP + TN) \Big/ (TP + TN + FP + FN) \quad (4)$$

$$Sensitivity = TP \Big/ (TP + FN) \quad (5)$$

$$Specificity = TN \Big/ TN + FP \quad (6)$$

$$Positive\ Predictivity = TP \Big/ (TP + FP) \quad (7)$$

$$Negative\ Predictivity = TN \Big/ (TN + FN) \quad (8)$$

2.5. Results

Following performance of the feature extraction algorithm was achieved;

- Overall accuracy: 93.2%
- Sensitivity: 97.3%
- Specificity: 89.5%
- Positive predictivity: 91.1%
- Negative predictivity: 97.1%

Table 1: *Average values of the extracted features from clinic and telephone recordings*

	Clinic Recordings	Telephone Recordings
Number of Pauses	47	50
Mean Pause Duration (s)	0.63	0.65
Proportion of Recording in Silence	0.20	0.21
Mean Utterance Duration (s)	2.51	2.48
Total Recording Time (s)	144.64	151.76
Total Length of Pauses (s)	29.57	32.35
Total Length of Utterances (s)	115.08	119.41

Further, features presented in Section 2.2.6 were extracted for 19 subjects from the clinic-based recordings and from the telephone recordings. Differences between the average values of the extracted features from the two recording environments are presented in Table 1.

3. Discussion

The monitoring of cognitive function of elderly people in clinic environment is labour intensive, very expensive and time consuming. It also only provides a snapshot of an individual's cognitive function. Therefore, a system that is able to provide continuous cognitive assessment and reduce cost associated with that assessment would be very beneficial in tackling the problems of an ageing population and promoting timely interventions. The use of speech characteristics for assessment of cognitive decline may represent this solution.

The hypothesis is that cognitive impairment affects speaking behaviours and voice sound characteristics and these can be measured through temporal features [2]. Cognitive decline affects neurological processing and it is the time taken to process speech and generate language is longer for cognitively impaired speakers. In a study by D'Arcy et al [3], use of a combination of temporal features yielded to 76% accuracy of classifying elderly people into group with MMSE score of 27 and lower, and group with MMSE score of 28 and higher. Where, MMSE scores of less than 27 are generally considered as possibly cognitively impaired by clinicians.

Speech can be easily remotely recorded over a telephone and, as shown it Section 2.4, speech features can be with little modifications to feature extraction algorithm reliably (overall accuracy of 93.2%) extracted from the telephone recordings. Average differences between features extracted from clinic environment recordings and telephone recordings for 19 subjects are under 15% for all features, except the Total Length of Pauses where the difference is around 19%. These differences were found not to be statistically significant, which supports our hypothesis of using speech recorded over telephone in studies investigating changes in cognitive function. As future work, changes in the features over time will be investigated in a longitudinal study, where participants' speech will be recorded remotely over a telephone every two months.

All recordings were conducted by a trainee psychologists calling the participants and asking the participants to perform different tasks to assess their level of cognitive function. Further development of the protocol and automation of the tasks may lead to fully automated Interactive Voice Response (IVR) system, similar to that presented in study of vocal fold pathologies by Moran et al [9] and Wormald et al [10]. Such a system would not necessitate the need for trained human experts, allowing expertise to be redirected towards intervention as opposed to assessment of cognitive function, and will allow monitoring of cognitive function of large scale cohorts.

The elderly people are willing to use technology if that is properly designed, i.e. works well in human settings, honours people's needs, does not remove person's autonomy, doesn't invades the privacy of its users [11]. Almost everybody is familiar with telephones and doesn't feel uncomfortable while using them, which may be another advantage of telephony based system for remote monitoring and assessment of cognitive function.

In this study, a re-estimation of the Reference Value (used for calculation of the Gradient Threshold in the detection of breath sounds) was needed prior extraction of features from the speech recordings made remotely over a telephone line. To avoid the need to re-estimate the Reference Value for every different recording system, research is currently developing automatic Reference Value estimation.

4. Conclusions

Features that are known to be used in the assessment of cognitive function can be reliably extracted from telephone quality speech. Fully automated IVR system may be developed that will decrease the cost associated with assessment and monitoring of cognitive function.

5. Acknowledgements

This research was supported by the TRIL Centre, www.trilcentre.org. The authors would like to acknowledge the efforts of Sian Counihan, Michelle Marie McCormick and Liam Quaide carrying out the data collection exercise both in the TRIL clinic at St James Hospital, Dublin and remotely by the telephone.

6. References

[1] Stassen, H. H., Albers, M., Püschel, J., Scharfetter, C., Tewesmeier, M., and Woggon, B.: 'Speaking behavior and voice sound characteristics associated with negative schizophrenia', Journal of Psychiatric Research, 1995, 29, (4), pp. 277-296

[2] 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.

[3] D'Arcy, S., Rapcan, V., Penard, N., Morris, M. E., Robertson, I. H., Reilly, R. B., "Speech as a Means of Monitoring Cognitive Function of Elderly Speakers" Proceedings of Interspeech, Brisbane, Australia, 2008.

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

[7] Ruinskiy, D., Lavner, Y., "An Effective Algorithm for Automatic Detection and Exact Demarcation of Breath Sounds in Speech and Song Signals", IEEE Transactions on Audio, Speech, and Language Processing, 15, (3), pp. 838-850, 2007.

[8] Deller, J. R., Proakis, J. G., and Hansen, J. H. L., Discrete-Time Processing of Speech Signals, Prentice-Hall, 1993.

[9] Moran, R.J., Reilly, R.B., de Chazal, P., and Lacy, P.D.: 'Telephony-based voice pathology assessment using automated speech analysis', Biomedical Engineering, IEEE Transactions on, 2006, 53, (3), pp. 468-477

[10] Wormald, R.N., Moran, R.J., Reilly, R.B., and Lacy, P.D.: 'Performance of an automated, remote system to detect vocal fold paralysis', Ann Otol Rhinol Laryngol, 2008, 117, (11), pp. 834-838

[11] Bailey, C., Sheehan, C., Mclean, A.: 'Older Adults, ICT and Technology: Users' Perspectives' conference proceedings from Centre for Excellence in Universal Design, Launch Conference, Dublin 2007 WP-ICSG-2008-20