34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Dynamic Minimum Pause Threshold Estimation for Speech Analysis in Studies of Cognitive Function in Ageing

Ivan Rochford, Viliam Rapcan, *Student Member, IEEE*, Shona D'Arcy, *Member, IEEE*, and Richard B. Reilly, *Senior Member, IEEE*

Abstract— **Cognitive decline represents the biggest limiting factor to independence in older adults. Speech analysis has emerged as an alternative to standard cognitive assessment tools. Temporal segmentation of speech is reported in many studies and typically employs a static threshold to define a pause. This study investigated the effect of using pause and utterance duration distribution data in differentiating between cognitively healthy and impaired older adults. Three sets of features were extracted from 187 speech recordings: temporal features using a static 250ms threshold; temporal features using a dynamic threshold; and pause and utterance duration distribution parameters. The ability of each of these sets to differentiate between cognitively healthy and cognitively impaired participants was investigated using a Linear Discriminant Analysis (LDA) classifier. Improvements of 0.22% (to 64.20%) in sensitivity, 6.33% (73.12%) in specificity, and 3.27% (68.66%) in overall accuracy were observed in the performance of the classifier using the pause and utterance duration distribution parameters when compared to the static temporal features. The use of the dynamic threshold had a negative impact on the classifier performance, with a decrease of 5.73% (to 58.25%) in sensitivity, 1.10% (65.69%) in specificity, and 3.42% (61.97%) in accuracy.**

I. Introduction

One of the negative aspects of ageing is the natural process of cognitive decline, which follows a linear trajectory over adulthood, accelerating into old and very-old age [1]. Subtle changes in cognitive function, however, can also be symptomatic of a progression towards mild cognitive impairment or dementia [2].

A debilitating condition predominantly occurring in older people, dementia is caused by disease of the brain, and is characterized by a progressive global deterioration in intellect including memory, learning, orientation, language, comprehension and judgment [3]. Early detection of dementia is widely accepted as being beneficial both to those with dementia and their carers and evidence suggests that treatments are likely to have maximum effect in the early stages of the condition [4].

*This work was completed as part of a wider program of research within the TRIL (Technology Research for Independent Living) Centre. The TRIL Centre is funded by Intel Coorporation, GE Healthcare and the Industrial Development Agency (IDA), Ireland.

I. Rochford, S. D'Arcy, V. Rapcan are with the Trinity Centre for Bioengineering, Trinity College Dublin, Ireland (e-mail: rochfori@tcd.ie, shona.darcy@tcd.ie, corresponding author - phone: +353-1-8964214; fax: +353-1-6795554; e-mail: rapcanv@tcd.ie).

R.B. Reilly is with the Trinity Centre for Bioengineering & School of Medicine, Trinity College Dublin, Ireland (e-mail: richard.reilly@tcd.ie).

Screening for cognitive decline is typically performed using the Mini-Mental State Exam (MMSE), a 30-point assessment tool typically administered in a clinical environment. The MMSE however is inherently flawed with reports of associated learning effects, ceiling effects and floor effects [5]. These effects significantly limit both the precision of the test and the frequency with which it can be administered, rendering it redundant in detecting subtle changes in cognitive function over time.

Speech analysis has emerged as a robust alternative to the MMSE in assessing cognitive function [6][7]. A study conducted by D'Arcy et al [6] investigated seven temporal features relating primarily to pause and utterance duration. Pause related features produced the strongest discrimination between cognitively healthy and cognitively impaired subjects. The conclusion being that it is what is *not* said rather than what *is* said that is an important feature of speech from people who are cognitively declining. This observation was also reported by Rapcan et al [8], Hird & Kirsner [9], and Pakhomov et al [7], and in studies reported by Roark et al [10].

Despite the apparent significance of pause-oriented features, much of the research into pause detection has employed criteria that is either poorly defined or based on the speech performance of neurologically intact adults [11]. Minimum pause duration is consistently used as a criterion for detection of pause boundaries [11] and is furthermore one that varies considerably across the literature, reflecting the arbitrary nature in which most researchers have employed it [12] [13]. Goldmann-Eisler [14] advocated the use of a minimum pause duration of 250ms, and this threshold has proliferated many of the studies into pausing behavior including Stassen et al [15], D'Arcy et al [6], and Rapcan et al [16]. Furthermore there are numerous examples of studies that deviated from this 250ms value for minimum pause duration – a review of relevant publications conducted by Kirsner et al [12] yielded 32 different values, ranging from 100 up to 300ms and with a median of 250ms. Minimum pause duration thresholds encountered as part of this study included 40ms [17], 150ms [7], 270ms [18], and 1000ms [10]. Given the variability in pause threshold values, it is not immediately evident what is an appropriate value for minimum pause duration.

Initiated by Kirsner et al [12], this traditional approach to pause identification has been readdressed, and emerging from this renewed interest is evidence that pause duration exhibits a two-component mixed lognormal distribution, one component associated with short pauses – products of articulatory processes and another associated with long

pauses – associated with cognitive processes. Having successfully fitted this distribution it is possible to determine an optimal threshold value for each recording for differentiating between the short-pause component and long-pause component. As this threshold value can vary from speaker to speaker it is dynamic in nature.

Recent studies by Rosen et al [19], and Hird & Kirsner [8] investigated the correlation between the pause distribution parameters and Friedrich's ataxia (FRDA) and brain-damage induced aphasia respectively. Rosen et al [19] found a significant difference between the parameters for those with FRDA and control subjects and Hird & Kirsner [8] graphically demonstrated the variation between brain-damaged subjects and control subjects. Given these findings and the findings of D'Arcy et al [6] and Roark et al [10] that pause-related features are correlated with cognitive function, there is clearly potential in using pause distribution parameters to discriminate between cognitively healthy and cognitively impaired subjects.

This study readdresses the approach of D'Arcy et al [6], replacing the use of a static 250ms minimum pause duration with this dynamic threshold. The parameters of the pause and utterance distributions are furthermore assessed in their ability to discriminate between cognitively healthy and cognitively impaired speakers.

II. Methods

A. Participants

Recruited from St. James's Hospital in Dublin as part of the Technology Research for Independent (TRIL) program, 187 older adults participated in this study, 73 (39.04%) of whom were male and 114 (60.96%) female, with a mean age of 72.44 (SD 7.01, range 60 - 80) years. All participants underwent cognitive assessment using the MMSE, the scores ranging from 20 to 30 with a mean MMSE score of 27.68 (SD 2.00). Based on their MMSE scores, the participants were segmented into two groups. With an MMSE score of 27 or higher, 150 of the participants (80.21%) were classified as cognitively healthy, and 37, with a score of 26 or lower (19.79%), were classified as cognitively impaired.

B. Audio Corpus

The audio corpus used for this project consists of 187 recordings of participants reading aloud. Recordings were made using 16-bit direct digital sampling at a sampling rate of 44.1 kHz. The recordings were saved in an uncompressed format. All recording were high-pass filtered at 80Hz to remove low frequency noise.

C. Speech Task

Each participant was required to read a short text passage from a children's story. This passage has been used in previous studies by Rapcan et al [8][16], D'Arcy et al [6] and also by Stassen et al [15]. The text passage is considered to be emotionally neutral and to exhibit verbal and semantic simplicity [15].The text is relatively short with most recordings being between two and three minutes in length.

D. Analysis

1) Pause and Utterance Detection

Pauses and utterance detection was performed employing three threshold values: minimum pause duration, minimum utterance duration, and minimum signal amplitude [11]. Breath detection and removal was implemented according to the algorithm described by Rapcan et al [20] prior to the application of these thresholds to prevent their misclassification as speech.

A minimum signal amplitude was calculated for each recording by performing full-wave rectification on the signal, and segmenting it into 50ms non-overlapping windows. Based on the empirical estimation made by Rapcan et al [16] that 15-20% of such windows can be classed as silence, a conservative 15% of the windows with the lowest energy were selected and the average value of the maximum amplitude in each window was determined – giving the minimum amplitude threshold. This threshold was then used to perform the initial speech/non-speech segmentation with all portions of the signal exceeding the threshold classed as speech and all those below the threshold classed as non-speech.

Following this initial segmentation temporal thresholds were applied to more accurately identify both pauses and utterances. Initial thresholds of 100ms for minimum utterance duration and 250ms for minimum pause duration were employed to provide the secondary segmentation of the signal into long utterances and pauses. Thresholds of 30ms and 20ms were then applied to the long utterances to break them into short utterances and pauses, providing the tertiary and final segmentation of the signal into speech and pauses. In this manner all pauses of duration greater than or equal to 20ms and all utterances of duration greater than or equal to 30ms were identified.

2) Pause and Utterance Distributions

Two-component mixed lognormal distributions were fitted to the pause duration data using maximum likelihood estimation (MLE). The two components of the distribution were termed the short-pause component and the long-pause component. The point at which the detection error was minimized for each component was found to be at the intersection of the two components. This intersection point was employed as the 'dynamic' pause threshold and was used to discriminate between the long-pauses, which were of interest and the short-pauses which were to be discarded. MLE was also employed to fit a unimodal lognormal distribution to utterance duration data for each recording.

The following features were then extracted from the pause and utterance distributions:

1. Pause Mixing Proportion
2. Short-Pause Mean
3. Long-Pause Mean
4. Short-Pause Standard Deviation
5. Long-Pause Standard Deviation
6. Utterance Mean
7. Utterance Standard Deviation

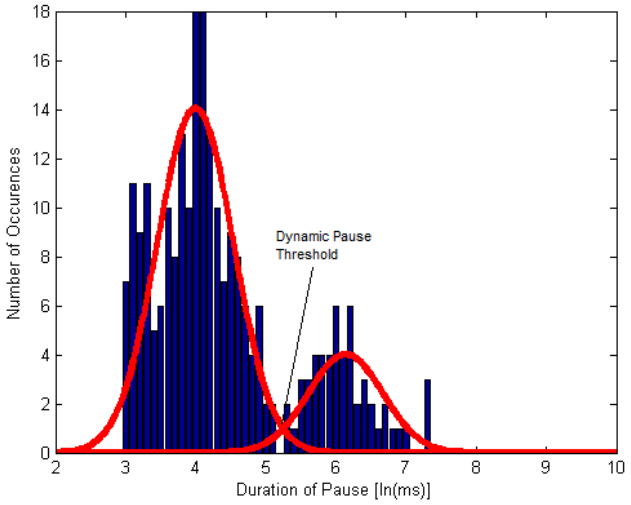

Figure 1. Pause Mixed Lognormal Distribution

3) Temporal Features

Using both the dynamic pause duration threshold from the pause distribution, and the static 250ms threshold employed by D'Arcy et al [6], seven temporal features were extracted:

1. Number of Pauses
2. Mean Pause Duration
3. Mean Pause Duration per Second
4. Mean Utterance Duration
5. Total Length of Pause
6. Total Length of Utterances
7. Total Recording Time

By employing both the static and dynamic thresholds it was possible to compare the performance of both thresholds in discriminating between cognitively healthy and cognitively impaired participants.

4) Statistical analysis

The performance of these features in discriminating between cognitively healthy (MMSE >=27) and cognitively impaired (MMSE<27) participants was assessed using a combination of Student's t-Test, Welch's t-Test and Wilcoxon's Rank-Sum Test. The selection of the appropriate test for each feature was based on the results of normality tests, variance tests, and visual inspection of quantile-quantile plots. The feature set under analysis comprised the parameters of the pause distributions and utterance distributions, and the temporal features for the static and dynamic thresholds. The features that emerged as statistically significant at a significance level of $\alpha = 0.05$ were subsequently used in the classification procedure.

5) Classification

A Linear Discriminant Analysis (LDA) classifier [21] was chosen to differentiate between cognitively healthy and impaired groups. Cross-fold validation [22] was used to maximize training and determine classification accuracies.

III. Results

From the three sets of features (static temporal features, dynamic temporal features, and distribution features) the variance analysis yielded seven statistically significant features (see Table I and II). Inspection of the mean values of these features (see Table III) revealed that the participants of the cognitively impaired group generated more pauses for the static threshold case, with a mean of 28.76 seconds compared to 24.01 for the healthy group. The '(Dynamic) Total Length of Pauses' feature indicated that the impaired group paused on average 6.5 seconds more than the healthy group. Both static and dynamic 'Total Length of Utterances' demonstrated that the impaired group had longer utterances than the healthy group.

TABLE I. Student's t-test

Speech Feature	Student's t-test		
	t	*df*	*p*
(Distribution) Utterance Mean	2.4824	185	0.0139
(Dynamic) Total Length of Utterances	2.1258	185	0.0348
(Static) Mean Pause Duration per Second	-2.5708	185	0.0109
(Static) Total Length of Utterances	2.2888	185	0.0260

Distribution – features extracted from lognormal distributions, Dynamic – temporal features extracted employing dynamically estimated pause threshold, Static – temporal features extracted employing static pause threshold

TABLE II. Wilcoxon Rank-sum Test

Speech Feature	Rank-Sum Test		
	U	*z*	*p*
(Distribution) Pause Mixing Proportion	3214	-2.0883	0.0368
(Dynamic) Total Length of Pauses	4488	2.0687	0.0386
(Static) Number of Pauses	4514	2.1556	0.0311

Distribution – features extracted from lognormal distributions, Dynamic – temporal features extracted employing dynamically estimated pause threshold, Static – temporal features extracted employing static pause threshold

TABLE III. Mean Feature Values for Both Participants' Groups

Speech Feature	Mean Feature Value	
	Cognitively Healthy Group	*Cognitively Impaired Group*
(Distribution) Utterance Mean (ln(ms))	5.77	5.93
(Distribution) Pause Mixing Proportion	0.76	0.71
(Static) Number of Pauses	24.01	28.76
(Static) Mean Pause Duration per Second (s)	0.13	0.15
(Static) Total Length of Utterances (s)	110.64	122.03
(Dynamic) Total Length of Pauses (s)	29.08	35.58
(Dynamic) Total Length of Utterances (s)	109.66	118.44

Distribution – features extracted from lognormal distributions, Dynamic – temporal features extracted employing dynamically estimated pause threshold, Static – temporal features extracted employing static pause threshold

TABLE IV. LINEAR DISCRIMINANT ANALYSIS CLASSIFICATION

Classification Performance	Feature set		
	(Static) Temporal Features	*(Dynamic) Temporal Features*	*Distribution Features*
Overall Accuracy (%)	65.39	61.97	68.66
Sensitivity (%)	63.98	58.25	64.20
Specificity (%)	66.79	65.69	73.12
ROC Area	0.69	0.58	0.74

ROC Area – Area under the Receiver Operating Characteristics (ROC) curve

Classification of the participants was then performed using these features, the results of which can be seen in Table IV. Employing the dynamic temporal features yielded decreases of 5.73% (to 58.25%) in the sensitivity, 1.10% (65.69%) in the specificity, and 3.42% (61.97%) in the accuracy when compared with the performance of the static temporal features. When the classifier was trained using the pause and utterance distribution features, the classification performance increased. The sensitivity of the LDA classifier increased by 0.22% (to 64.20%), specificity by 6.33% (73.12%) and the overall accuracy by 3.27% (68.66%).

IV. DISCUSSION

The results of this study reaffirm the potential present in speech analysis for assessing cognitive function of older subjects. From the investigation of pause and utterance duration distribution data and their impact on the performance of an LDA classifier, two distributional parameters, pause mixing proportion and utterance mean, were found to be statistically significant and encouragingly outperformed the temporal features in classifying the participants according to their level of cognitive function. Contrary to expectations however the use of a dynamic threshold derived from the distributional data had a negative impact on the classification performance of the temporal features. These findings would suggest that the distributional data are best employed directly, rather than using them to tailor the temporal features for each individual speaker via the dynamic thresholding.

V. CONCLUSION

Despite dynamic thresholding yielding no improvement on traditional methods, this study did highlight the potential in pause and utterance duration distribution parameters in classifying people according to their cognitive function. Because speech can be acquired remotely (e.g. via telephone) the results of the study also indicate that temporal speech analysis may help in the development of systems for the remote detection of cognitive decline.

ACKNOWLEDGMENT

The authors would like to thank the members of the TRIL clinic in St James's hospital Dublin and all participants for their contribution to this study.

REFERENCES

[1] D. Zimprich et al., "Cognitive Abilities in Old Age: Results from the Zurich Longitudinal Study on Cognitive Aging," Swiss Journal of Psychology 67 (3), 2008, pp. 177-195.

[2] A. Levey, J. Lah, F. Goldstein, Kyle Steenland, and Donald Bliwise, "Mild Cognitive Impairment: An Opportunity to Identify Patients at High Risk for Progression to Alzheimer's Disease," Clinical Therapeutics (28) 7, 2006, pp. 991 - 1001.

[3] M. Prince et al., "The World Alzheimer Report 2009," London, UK, 2009.

[4] A. Milne, "Dementia Screening and Early Diagnosis: The Case For and Against," Health, Risk & Society, vol. 12, no. 1, 2010, pp. 65 – 76.

[5] H. J. Woodford and J. George, "Cognitive Assessment in the Elderly: A Review of Clinical Methods," Quarterly Journal of Medicine, vol. 100, 2007, pp. 469 – 484.

[6] S. D'Arcy et al., "Speech as a Means of Monitoring Cognitive Function of Elderly Subjects," Proceedings of Interspeech, Brisbane, Australia, 2008, pp. 2230 - 2233.

[7] V. Rapcan, S. D'Arcy, N. Pénard, I. H. Roberston, and R. B. Reilly, "The Use of Telephone Speech Recordings for Assessment and Monitoring of Cognitive Function in Elderly People," Proceedings of Interspeech 2009, Brighton, England, 2009.

[8] K. Hird and K. Kirsner, "Objective Measurement of Fluency in Natural Language Production: A Dynamic Systems Approach," Journal of Neurolinguistics 23, 2010, pp. 518 - 530.

[9] S. V. S. Pakhomov et al., "Effects of Age and Dementia on Temporal Cycles in Spontaneous Speech Fluency," Journal of Neurolinguistics 24, 2011, pp. 619 - 635.

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

[11] J. R. Green, D. R. Beukelman, and L. J. Ball, "Algorithmic Estimation of Pauses in Extended Speech Samples of Dysarthric and Typical Speech," Journal of Medical Speech-Language Pathology, vol. 12, no. 4, 2004, pp. 149 – 154.

[12] K. Kirsner, J. Dunn, K. Hird, T. Parkin, and C. Clark, "Time for a Pause.," in Proceedings of the 9th Australian Conference on Speech Science & Technology, Melbourne, Australia, 2002, pp. 52 - 57.

[13] A. E. Hieke, S. Kowal, and D. C. O'Connell, "The Trouble with "Ariculatory Pauses"," Language and Speech, vol. 26, no. 3, 1983, pp. 203 – 214.

[14] F. Goldmann-Eisler, Psycholinguistics: Experiments in Spontaneous Speech. New York, USA: Academic Press, 1968.

[15] H. H. Stassen et al., "Speaking Behavior and Voice Sound Characteristics Associated with Negative Schizophrenia," Journal of Psychiatric Research (29) 4, 1995, pp. 277 – 296.

[16] V. Rapcan et al., "Acoustic and Temporal Analysis of Speech: A Potential Biomarker for Schizophrenia," Medical Engineering & Physics (32), 2010, pp. 1074 - 179.

[17] L. Castro and J. A. Moraes, "The Temporal Structure of Professional Speaking Styles in Brazilian Portuguese," Proceedings of ITRW on Experimental Linguistics, Athens, Greece, 2008, pp. 57 - 60.

[18] S. Kowal, D. C. O'Connell, E. A. O'Brien, and E. T. Bryant, "Temporal Aspects of Reading Aloud and Speaking: Three Experiments," The American Journal of Psychology, vol. 88, no. 4, 1975, pp. 549 – 569.

[19] K. Rosen et al., "Automatic Method of Pause Measurement for Normal and Dysarthric Speech," *Clinical Linguistics & Phonetics*, vol. 24, no. 2, 2010, pp. 141 – 154.

[20] V. Rapcan, S. D'Arcy, and R. B. Reilly, "Automatic Breath Sound Detection and Removal for Cognitive Studies of Speech and Language," Proceedings of the Irish Signal and Systems Conference, Dublin, Ireland, 2009.

[21] R. Duda, P. Hart, and D. Stork, in Pattern Classif., S. Edition, Ed., ed: Wiley-Interscience, 2000.

[22] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2, Montreal, Quebec, Canada, 1995.