Digital Biomarkers

Research Reports – Research Article

Digit Biomark 2023;7:157–166
DOI: 10.1159/000533423

Received: September 6, 2022
Accepted: July 28, 2023
Published online: November 24, 2023

How Much Speech Data Is Needed for Tracking Language Change in Alzheimer's Disease? A Comparison of Random Length, 5-Min, and 1-Min Spontaneous Speech Samples

Ulla Petti[a] Simon Baker[a] Anna Korhonen[a] Jessica Robin[b]

[a]Language Technology Lab, University of Cambridge, Cambridge, UK; [b]Winterlight Labs, Toronto, ON, Canada

Keywords
Digital biomarkers · Natural language processing · Speech · Language · Alzheimer's disease · Dementia · Digital health

Abstract
Introduction: Changes in speech can act as biomarkers of cognitive decline in Alzheimer's disease (AD). While shorter speech samples would promote data collection and analysis, the minimum length of informative speech samples remains debated. This study aims to provide insight into the effect of sample length in analyzing longitudinal recordings of spontaneous speech in AD by comparing the original random length, 5- and 1-minute-long samples. We hope to understand whether capping the audio improves the accuracy of the analysis, and whether an extra 4 min conveys necessary information. ***Methods:*** 110 spontaneous speech samples were collected from decades of Youtube videos of 17 public figures, 9 of whom eventually developed AD. 456 language features were extracted and their text-length-sensitivity, comparability, and ability to capture change over time were analyzed across three different sample lengths. ***Results:*** Capped audio files had advantages over the random length ones. While most extracted features were statistically comparable or highly correlated across the datasets, potential effects of sample length should be acknowledged for some features. The 5-min dataset presented the highest reliability in tracking the evolution of the disease, suggesting that the 4 extra minutes do convey informative data. ***Conclusion:*** Sample length seems to play an important role in extracting the language feature values from speech and tracking disease progress over time. We highlight the importance of further research into optimal sample length and standardization of methods when studying speech in AD.

Introduction

Changes in speech have been identified as one of the earliest biomarkers of Alzheimer's disease (AD) [1–4]. As studies using AI and language-based biomarkers have achieved over 90% accuracy in differentiating the individuals with AD diagnosis from the healthy population [5], contemporary studies are predominantly concerned with prediction and detection of the earliest language changes [6]. However, due to limited data availability [6–8], relatively few studies focus on preclinical [1] and longitudinal changes [4, 9].

Similarly, methods for speech data collection, such as optimal sample length, have not been standardized [8, 10], limiting the comparability across studies and the transferability of research methods. Previous studies have

karger@karger.com
www.karger.com/dib

Correspondence to:
Ulla Petti, ump20@cam.ac.uk

often used random sample length which can lead to bias in feature values due to their text-length-sensitive, especially as people with AD tend to speak less [11–14]. Different strategies have been proposed to cope with text-length-sensitivity, for example, extracting language features as ratios [15–18], however, ratio features can be unreliable as the proportion of specific language features does not change linearly with text length [19, 20]. Another strategy is capping the samples at a maximum length [19], however, the minimum amount of speech data needed to spot AD-related changes remains debated [10]. Previous studies have proposed that 150-word interviews can reflect language impairment in dementia realistically [10], used 1-minute- [21], 4-minute- [22], or 1,400-word-samples [23], capped the transcripts only when analyzing text-length-sensitive features [20] or removed the acoustic features that correlate with duration [24]. All in all, there is no standard method for dealing with text-length-sensitivity, and very different sample lengths have been used in previous research.

Due to limited availability of longitudinal datasets, previous work on standardizing the duration of the samples has not, to the best of our knowledge, considered the ability of different length speech samples to capture language change over time. While longitudinal changes in speech have received little attention compared to classification studies [11, 25] as the datasets are expensive and time-consuming to collect [4, 6, 9], previous literature suggests that changes in language use, such as vocabulary richness [20, 23, 26, 27], syntax [2, 20], and pausation patterns affected by word retrieval difficulties [28, 29] begin before the clinical diagnosis and worsen over time. We aim to investigate how the length of speech sample affects their ability to capture these expected changes.

The goal of the current study is to explore the effect of audio capping in analyzing speech in AD and to understand whether 1-min samples carry enough information in comparison to 5-min samples. Using shorter samples would promote data collection and analysis by requiring less effort from vulnerable subjects and minimizing computation time, but the samples must be long enough to be informative.

Our dataset consists of transcripts of public interviews with famous individuals, some of whom eventually develop AD. We have constructed three different length datasets from the same voice samples by cutting the transcripts and audio: original random length, 5-min, and 1-min dataset. While this dataset cannot replace clinical studies, it has several advantages: (1) collecting longitudinal data from prerecorded interviews in the public domain provides an insight into language change in AD without expert labeling and extensive privacy concerns, unlike clinical data [8], and allows to explore how speech duration affects the robustness of the extracted features; (2) while cognitive tasks conducted in laboratory setting can have methodological advantages, spontaneous speech is considered to provide a more realistic reflection of cognitive abilities and promotes longitudinal data collection due to natural setting [30]; (3) using multiple speech samples recorded over several decades from a number of participants allows to explore longitudinal change instead of the difference at a single time point, contributing to finding the earliest markers and tracking disease progress; (4) using over 450 language features contributes to detailed analysis of changes in speech patterns in a wide range of language areas, and automated feature calculation improves objectivity of the analysis [8].

Materials and Methods

Participants

The original dataset consisted of 405 recordings from 9 AD – healthy control (HC) participant pairs of public figures. The public figures with AD were identified based on internet searches and the number of available voice recordings. As we did not have access to the clinical information of the public figures, the time of diagnosis was based on Wikipedia and media entries. HC participants were paired with the AD participants based on demographic information (gender, age, education level, origin, and occupation where possible) to minimize the difference in life experiences and background which may affect language use.

As the design of the current study required the length of the samples to be at least 5 min long, all shorter samples were excluded from the original dataset, resulting in 110 speech samples from 17 individuals. The included speech samples were public interviews and monologues recorded during TV and radio performances, including contexts such as talk shows, documentary interviews, press conferences, public speeches, etc. All speech samples were available on Youtube. See Table 1 for details of the individual participants and the recordings included in the final dataset.

Materials

Speaker diarization was done manually by trained transcriptionists, separating speech segments based on speaker identity, and allowing us to only include the audio of the speaker of interest. Speech was manually transcribed by Winterlight Laboratories employees, following CHAT protocol.

The same speech samples were used to create three different length datasets. All included speech samples in the original random length dataset were at least 5-min long after diarization, but their length was not capped, ranging from 5 to 21 min (mean = 9.75), and 479–6,665 words (mean = 1,893).

To create the 5-min dataset, a cut-off point was applied at the 300-s mark of these samples. The number of words across the 110 samples in the 5-min dataset ranged from 479 to 1,339 (mean = 908).

Table 1. Individual participant and recording information in the final dataset

Participant	Number of recordings	Age range over recordings	Sex	Education	Age at diagnosis
AD_1	12	43–76	M	12	75
HC_1	9	44–88	M	12	–
AD_2	7	46–74	M	17	80
HC_2	4	34–77	M	16	–
AD_3	7	50–77	M	14	79
HC_3	8	35–81	M	14	–
AD_4	3	65–75	M	18	80
HC_4	7	32–81	M	13	–
AD_5	3	68–71	M	14	83
HC_5	8	44–86	M	16	–
AD_6	8	64–72	M	16	78
HC_6	9	70–75	M	18	–
AD_7	2	47–59	F	17	58
HC_7	7	55–63	F	17	–
AD_8	1	63	F	12	63
HC_8	7	37–62	F	12	–
AD_9	8	75–81	M	16	83

To avoid re-transcription and potential between-transcriber differences, the 1-min cut-off point was established at the nearest utterance boundary of the 60-s mark, resulting in ±5-s variation in sample length. The transcript length in the 1-min dataset ranged from 102 to 278 words (mean = 179).

This process resulted in three different length datasets consisting of the same speech samples, so that the first minute was the same in all datasets, and the first 5 min was the same in the 5-min and original random length dataset.

456 identical linguistic and acoustic features were automatically extracted from each dataset. Linguistic features were related to coherence (such as cosine similarity between utterances), syntactic structures (such as t-units – the shortest grammatical sentences into which text can be split), lexical features (such as age of acquisition [AoA] or level of arousal), and vocabulary richness (such as type:token ratio [TTR] – the number of different words divided by the number of total words; moving average type:token ratio (MATTR) [31] – TTR in different moving window sizes of consecutive words; Brunet index [32] – based on text length and vocabulary size [16]; Honoré statistic [33] – based on hapax legomena and an assumption that growth in their use is constant to the logarithm of text size [34]).

Acoustic features included speech tempo- and pause-related features, Mel-frequency cepstral coefficient (MFCC) measures, and zero-crossing rate statistics. Acoustic features were extracted using Praat and Parselmouth software [35, 36].

Procedure

We conducted 3 experiments. Experiment 1 explored the text-length-sensitivity of the language features in the three different datasets by measuring Spearman correlations between the language feature values and the number of words. This was done separately in each dataset (n = 3), between each feature (n = 456) and text length, resulting in 3 × 456 correlations using 110 datapoints. Bonferroni correction was applied to account for multiple comparisons, resulting in a significant p value of $p = 0.0001$. For informative analysis of speech in AD, we would expect the feature values not to be extensively affected by transcript length. Experiment 2 investigated the comparability of feature values across datasets using Kruskal-Wallis test to identify the number of features that differed in the three datasets, and Dunn test for post hoc analysis to identify the datasets where these differences occurred. Spearman correlation was used to analyze whether the values of the features that differed in the two datasets correlated with each other. We hypothesized that if the datasets of different lengths capture the same information, the feature values across the datasets should be comparable or strongly correlated. In Experiment 3, we focused on language change over time and investigated the relationship between participant age, dataset length, and language features. We first explored the number of significant Spearman correlations between participant's age and the language feature in each dataset, comparing the AD and HC groups. We expected an informative dataset to capture the largest number of significant correlations with age in the AD group as a representation of a more rapid decline in language due to progressing cognitive difficulties, and the correlations between age and language feature values in the HC group to be less significant. Second, we used linear mixed effect model to better understand whether the age impact on the language features differs depending on dataset length in the AD group.

Results

Experiment 1

We found 120 text-length-sensitive features in the original random length, 30 in the 5-min, and 22 in the 1-min dataset. 104 text-length-sensitive features in the original random length dataset were acoustic MFCC features, compared to 0 and 1 in the other two datasets. Figure 1 shows the details of the text-length-sensitivity of the features, with MFCC features removed due to strong correlations with each other. See Table 2 for examples of feature categories and their text-length-sensitivity.

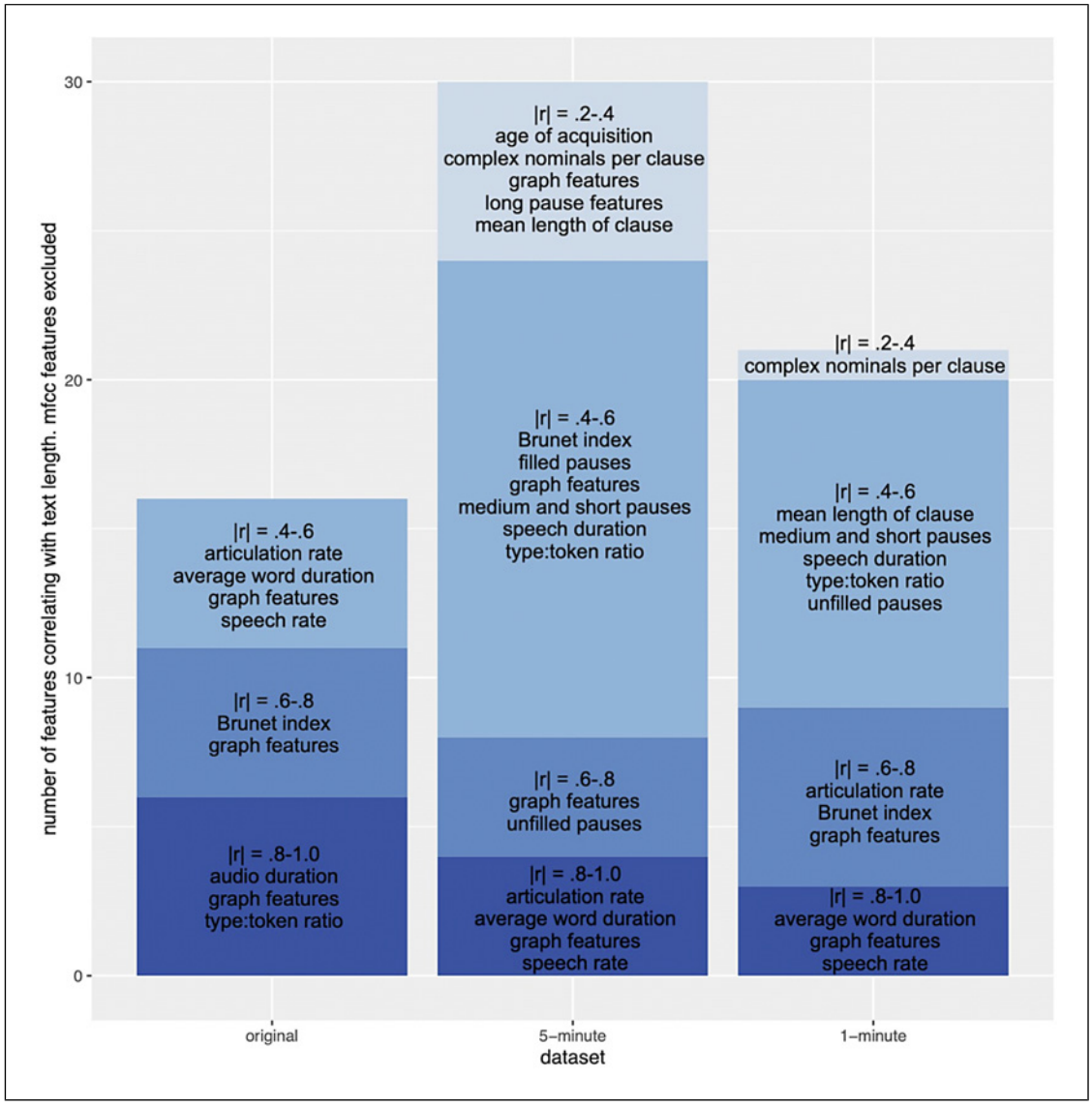

Fig. 1. Number of features correlating with text length in each dataset, with MFCC features excluded.

Experiment 2

226 language features showed significant differences across datasets in Kruskal-Wallis test (118 acoustic, 55 syntactic, 18 vocabulary richness and parts-of-speech proportions, 15 coherence, 11 graph, and 9 fluency features). Table 3 shows a breakdown of post hoc analysis exploring the difference between dataset pairs using Dunn test, as well as the number of features correlating significantly between the datasets using Spearman correlation, and the number of features that remained incomparable (significantly different and not correlated).

Experiment 3

The 5-min dataset captured a significant correlation with age in six language features in the AD group: AoA of words, AoA of nouns, average word duration, articulation rate, speech rate, and total words uttered. The original random length dataset captured 3 features that correlated significantly with age in the AD group: nonwords and incomprehensible words, AoA of words, and AoA of nouns. The 1-min dataset did not capture significant change in any features in the AD group. No feature correlated significantly with age in any dataset in the HC group.

Table 2. Examples of language features according to their category and text-length-sensitivity

Category	Not text-length-sensitive	Text-length-sensitive
Discourse	Cosine similarity between utterances	Graph features
Syntax	Phrase and sentence constructions	–
Lexical	Familiarity, imageability	Age of acquisition
Vocabulary richness	Honoré statistic, moving average type:token ratio	Type:token ratio, brunet index
Acoustic	Zero-crossing rate features	Mel-frequency cepstral coefficient features
Speech timing	–	Unfilled pauses, hesitation, pause count and duration, word duration, articulation rate

Table 3. Dataset comparability – number of features that show significant differences in the Kruskal-Wallis tests, Dunn test, and Spearman correlation across different length datasets

	1-min versus 5-min dataset	5-min versus original random length dataset	1-min versus original random length dataset
Number of significantly different features (Dunn test)	176	154	214
Out of the different features, the number of features that correlate across the two datasets (Spearman correlation)	159	71	81
Number of critical features that differ and do not correlate	17	83	133
Number of critical acoustic features	9	64	64
Number of critical linguistic features	8	19	69
Breakdown of critical linguistic features	6 syntactic 2 coherence	8 syntactic 8 coherence 3 graph	44 syntactic 14 coherence 7 graph 3 vocabulary richness 1 fluency

LMER showed that the effect of age in the AD group significantly differed in 6 language features depending on whether the dataset was 1-minute-long or of random length: the number of complex nominals, the length of t-units, verb familiarity, function words, average shortest path in a graph, and sentiment arousal. The 5-min dataset did not differ significantly.

As an illustration, Figures 2 and 3 show examples of average group and individual language feature values at different time points in different datasets. The feature choices are based on the results of Experiments 1–3; the one individual AD participant was chosen based on most available samples.

Discussion

This study examined the role of audio duration in capturing the changes in language in AD. Understanding the role of sample length contributes to standardizing the methods of language-based cognitive decline analysis, making them more transferrable and increasing the comparability in future studies, as well as the efficiency of clinical applications. We compared 1-min, 5-min, and original random length spontaneous speech samples recorded over several decades, focusing on text-length-sensitivity of the language features, the comparability of the information captured by different length datasets, and language change with time.

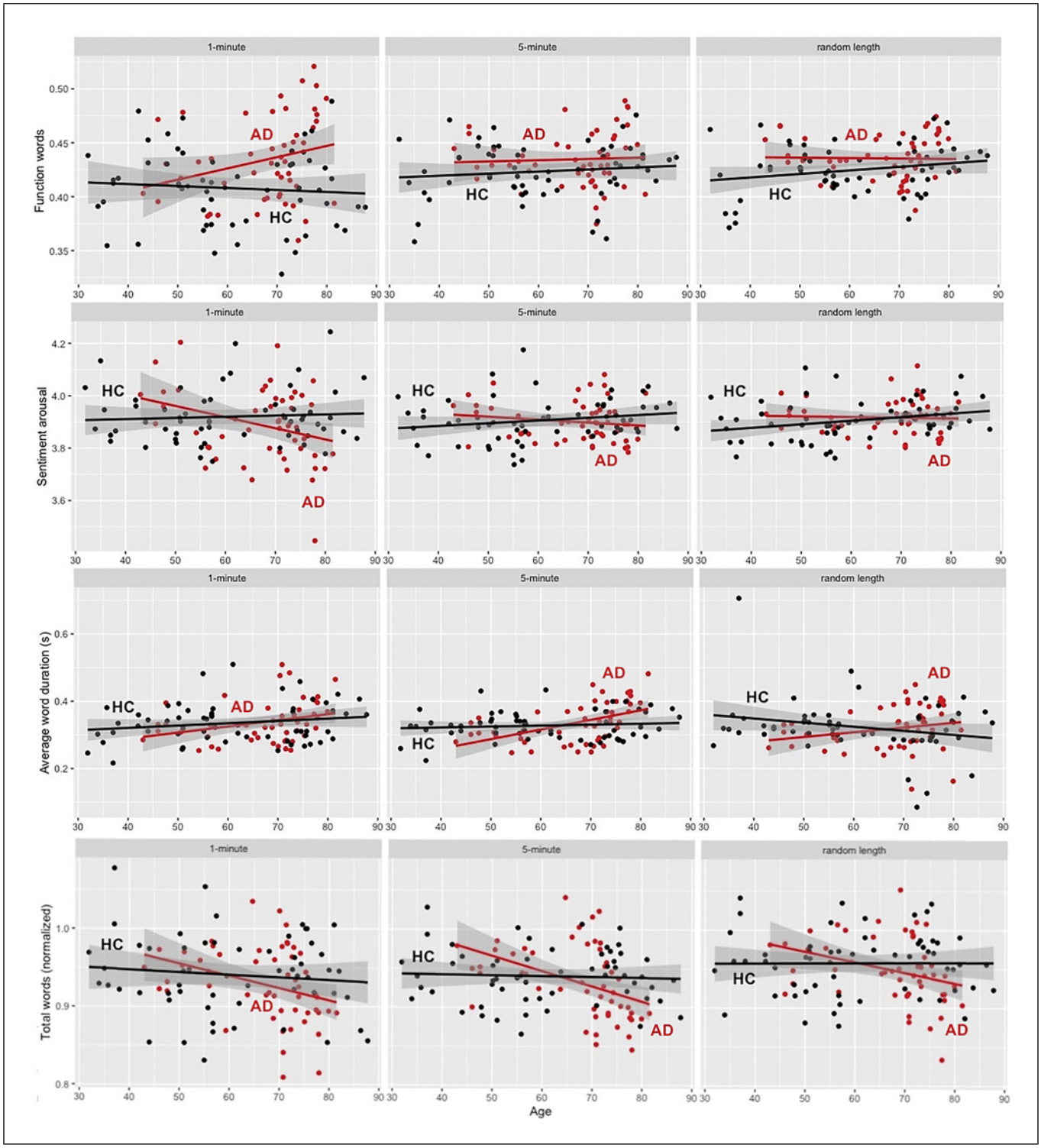

Fig. 2. Average group values of function words, arousal, word duration, and the number of words at different time points across 3 datasets and AD and HC participant groups (AD, Alzheimer's disease – red, HC, healthy control – black).

Digit Biomark 2023;7:157–166
DOI: 10.1159/000533423

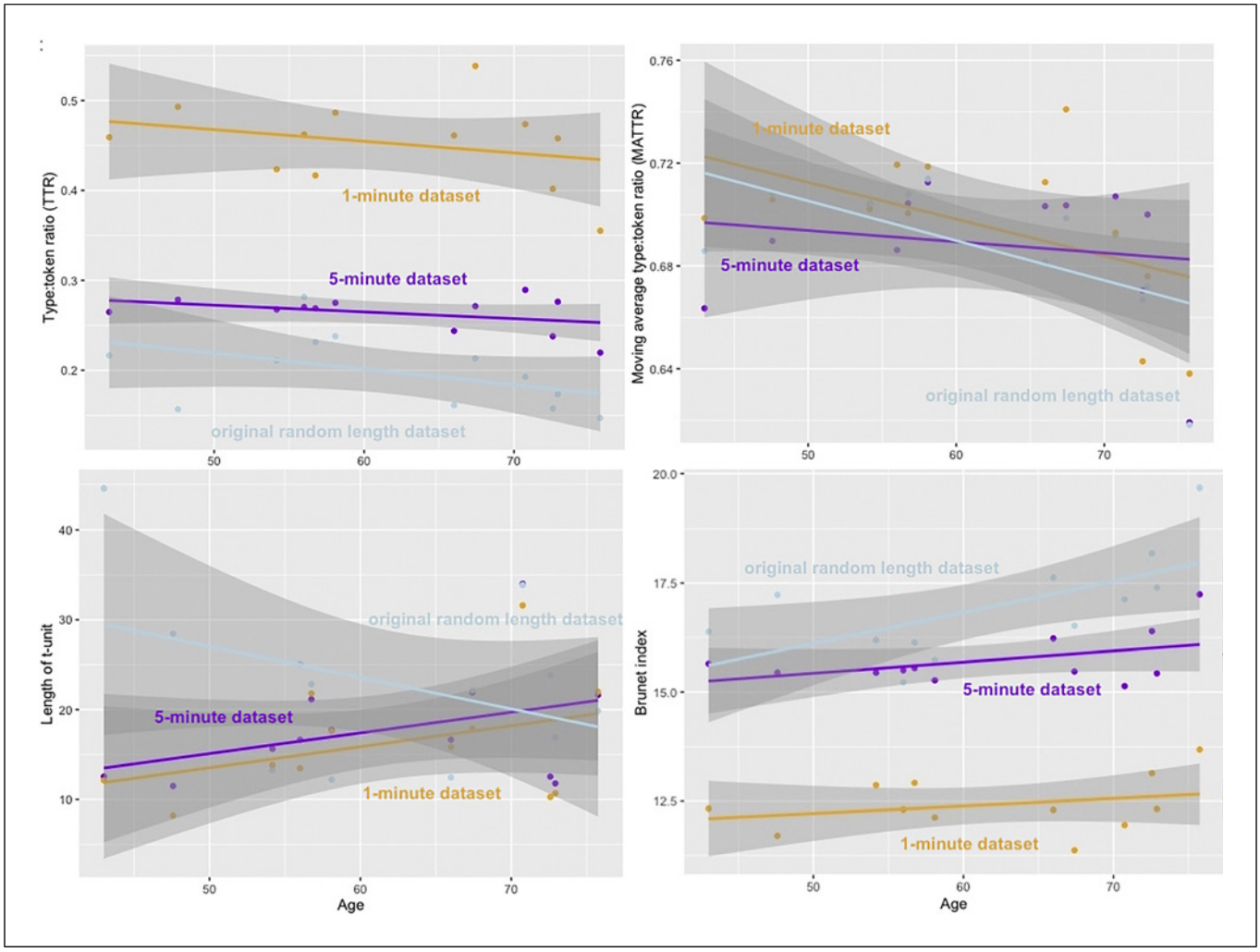

Fig. 3. Individual participant's feature values of type:token ratio, moving average type:token ratio, average length of t-units, and Brunet index at different time points across 3 datasets (1-min dataset – orange, 5-min dataset – purple, original random length dataset – light blue).

As expected, capping audio eliminated text-length-sensitivity of the acoustic features, supporting the standardization of sample length in acoustic analysis. The impact on linguistic features was less clear, as a similar relatively small number of features correlated with sample length in all three datasets (see Fig. 1). While most of these correlations were straightforward, vocabulary richness has been discussed more in previous literature [8, 34, 37]. Many studies suggest that as a function of the total tokens, TTR is likely to be affected by sample length [8, 16, 18, 37], and MATTR has been proposed as an alternative, less text-length-dependent metric [31]. The results of Experiment 1 (Fig. 1) and the illustration on Figure 3 support these claims. While numerous sources suggest that Brunet index is independent of sample length [1, 8, 18, 38], the results of Experiment 1 showed significant correlations with text length in all 3 datasets, in line with the illustration on Figure 3. Both Brunet index and TTR values suggest poorer vocabulary in longer datasets, explainable by the number of unique words not increasing linearly with the number of tokens [8, 20, 31]. As vocabulary richness is often used in studies concerned with speech in AD, it is important to keep the potential effect of sample length in mind and apply appropriate preventative measures, such as using MATTR instead of TTR, controlling for (the correlation with) sample length [34, 37], and reporting the length of the samples to increase comparability across studies.

In line with previous studies arguing that there are no major differences between the first and the second 150 words of an interview [39], or the first 150 and 600 words [10], we found that the information captured by 1-min and 5-min dataset was mostly comparable. A few acoustic and linguistic features suggested differences in the content captured: some complex syntactic structures may not have appeared in the shorter 1-min samples [39], and some coherence metrics suggest that 5-min samples understandably convey more diverse content. As expected, the random length samples differed the most from the other 2 datasets.

Earlier literature suggests that changes in vocabulary richness and verbal memory [20, 23, 26, 27], syntax complexity [2, 20], and fluency [28, 29] manifest before AD diagnosis and continue declining as the condition worsens. In the current study, the 5-min dataset suggested significant age correlation in six language features related to vocabulary complexity (AoA) and speech fluency (speech rate, word duration, total words) in the AD group. It has been proposed that familiarity-based memory is likely to be preserved in early stages of AD [27], potentially explaining the changes in AoA features. Similarly, word retrieval difficulties have been linked to pause proportion in speech, affecting speech rate, word duration, and the number of words [28] – the percentage of voiceless segments has been proposed to be one of the most informative features of the decline in language ability in AD with a potential to provide a low-cost solution to early AD detection [29].

The impact of age on six language features related to vocabulary complexity and familiarity, and syntax differed significantly between the 1-min and original random length dataset, suggesting that dataset length plays an important role in capturing language change in AD. While the 1-min dataset seems to capture changes in some language features most clearly (see, for example, function words on Fig. 3) and could therefore have a potential to differentiate between the AD and HC participants from early on, the feature values tend to be more randomly distributed and further from the predicted values than the same feature values in the 5-min samples, potentially indicating that the first minute is too short to provide reliable and stable results. The downside of the random length samples seems to be the appearance of the odd outliers, potentially due to some drastically longer or shorter samples. The age impact of the 5-min dataset did not differ significantly from the other two datasets, potentially suggesting that 5 min is long enough to capture the content comparable to the longer random length samples while, potentially due to standardized length, also remaining comparable to the 1-min samples.

Based on these findings, the 5-min dataset seems to provide most stable feature values and capture the expected decline in language most consistently. However, when interpreting the results, it must be remembered that this study is based on a very small sample and the findings are descriptive in nature and could only act as indicator, suggesting directions for further research. Dataset size and heterogeneity have been identified as the main limitation of the studies exploring speech in AD [7], contributing to potential overfitting. To avoid overfitting, it is also important to consider the number of necessary features extracted from the limited amount of data (see [40] for more detailed discussion).

Other limitations include the lack of clinical and confounder information, potential scriptedness of the recordings, uncontrolled content, intervals, and setting which limit the comparability of the samples. Further research using larger sample size, real participants, standardized recording conditions and intervals, and reliable clinical information is needed to make stronger conclusions. It must also be acknowledged that while the current study uses 5-min or longer speech samples, it can be challenging to find or collect such data in naturally occurring environments. Additionally, free speech and interviews might place different cognitive demands on the participants, and different tasks and lengths of speech can capture different changes. Therefore, the research design decisions on the type and length of the data should be made based on the research question.

Future studies could investigate whether the amount of speech data needed to capture change in language reduces as the participant's condition worsens, for example, whether less data could be used for more severe cases due to the impairment being more easily detectable. This would help reduce testing-related stress of the participants with more severe dementia and encourage participation. Another direction could be exploring the trade-off between the number and the length of the samples and analyze whether it is best to collect more shorter samples or fewer longer samples.

To sum up, the current study demonstrates that the length of the recordings plays an important role in analyzing speech changes in AD. The findings suggest that capped audio files have advantages over the random length ones, and while the 1-min and 5-min dataset convey largely comparable information, the stability of the feature values and the ability to track language change over time in AD advocate for using 5-min samples.

Acknowledgments

We thank the employees at Winterlight Laboratories who helped with data collection, transcription, and processing.

Statement of Ethics

An ethics statement is not applicable because this study is based exclusively on public data freely available online, based on recordings of televised appearances.

Conflict of Interest Statement

Jessica Robin is an employee of Winterlight Laboratories, Inc. Ulla Petti has previously been an intern at Winterlight Laboratories, Inc.

Funding Sources

This work was supported by Economic and Social Research Council (ESRC) Cambridge Doctoral Training Partnership (DTP) grant number ES/P000738/1.

Author Contribution

Jessica Robin and Ulla Petti contributed to the conception of the manuscript and design of the work. Ulla Petti performed data analysis and drafted the original manuscript. Simon Baker, Anna Korhonen, and Jessica Robin revised the manuscript and contributed to the interpretation of data. All authors approved the final version of the manuscript.

Data Availability Statement

All data generated or analyzed during this study are included in this article. Further inquiries can be directed to the corresponding author.

References

1 Calzà L, Gagliardi G, Rossini Favretti R, Tamburini F. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. Computer Speech Lang. 2021;65:101113.
2 Fang C, Janwattanapong P, Martin H, Cabrerizo M, Barreto A, Loewenstein D, et al. Computerized neuropsychological assessment in mild cognitive impairment based on natural language processing-oriented feature extraction. 2017 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE; 2017. p. 543–6.
3 Forbes-McKay KE, Venneri A. Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. Neurol Sci. 2005;26(4):243–54.
4 Yancheva M, Fraser KC, Rudzicz F. Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. In: Proceedings of SLPAT 2015: 6th workshop on speech and language processing for assistive technologies; 2015. p. 134–9.
5 Petti U, Baker S, Korhonen A. A systematic literature review of automatic Alzheimer's disease detection from speech and language. J Am Med Inform Assoc. 2020;27(11):1784–97.
6 Meltzer JA. Towards early prediction of Alzheimer's disease through language samples. EClinicalMedicine. 2020:29.
7 Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. Alzheimer's dementia recognition through spontaneous speech: the ADReSS challenge. arXiv preprint; 2020. arXiv:2004.06833.
8 Voleti R, Liss JM, Berisha V. A review of automated speech and language features for assessment of cognitive and thought disorders. IEEE J Sel Top Signal Process. 2020;14(2):282–98.
9 Robin J, Xu M, Kaufman LD, Simpson W. Using digital speech assessments to detect early signs of cognitive impairment. Front Digit Health. 2021;3:749758.
10 Sajjadi SA, Patterson K, Tomek M, Nestor PJ. Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. Aphasiology. 2012;26(6):847–66.
11 Lopez-de-Ipina K, Martinez-de-Lizarduy U, Calvo PM, Mekyska J, Beitia B, Barroso N, et al. Advances on automatic speech analysis for early detection of Alzheimer disease: a non-linear multi-task approach. Curr Alzheimer Res. 2018;15(2):139–48.
12 Ash S, Moore P, Antani S, McCawley G, Work M, Grossman M. Trying to tell a tale: discourse impairments in progressive aphasia and frontotemporal dementia. Neurology. 2006;66(9):1405–13.
13 Knibb JA, Woollams AM, Hodges JR, Patterson K. Making sense of progressive non-fluent aphasia: an analysis of conversational speech. Brain. 2009;132(Pt 10):2734–46.
14 Graham NL, Emery T, Hodges JR. Distinctive cognitive profiles in Alzheimer's disease and subcortical vascular dementia. J Neurol Neurosurg Psychiatry. 2004;75(1):61–71.
15 Beltrami D, Gagliardi G, Rossini Favretti R, Ghidoni E, Tamburini F, Calzà L. Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline? Front Aging Neurosci. 2018;10:369.
16 Bucks RS, Singh S, Cuerden JM, Wilcock GK. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: evaluation of an objective technique for analysing lexical performance. Aphasiology. 2000;14(1):71–91.
17 Gosztolya G, Vincze V, Tóth L, Pákáski M, Kálmán J, Hoffmann I. Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. Computer Speech Lang. 2019;53:181–97.
18 Guinn C, Singer B, Habash A. A comparison of syntax, semantics, and pragmatics in spoken language among residents with Alzheimer's disease in managed-care facilities. 2014 IEEE symposium on computational intelligence in healthcare and e-health (CICARE). IEEE; 2014. p. 98–103.
19 Durán P, Malvern D, Richards B, Chipere N. Developmental trends in lexical diversity. Appl Linguistics. 2004;25(2):220–42.
20 Le X, Lancashire I, Hirst G, Jokel R. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. Lit Ling Comput. 2011;26(4):435–61.
21 Chien YW, Hong SY, Cheah WT, Fu LC, Chang YL. An assessment system for Alzheimer's disease based on speech using a novel feature sequence design and recurrent neural network. 2018 IEEE international conference on systems, man, and cybernetics (SMC). IEEE; 2018. p. 3289–94.
22 Romero B, Kurz A. Deterioration of spontaneous speech in AD patients during a 1-year follow-up: homogeneity of profiles and factors associated with progression. Dement. 1996;7(1):35–40.

23 Berisha V, Wang S, LaCross A, Liss J. Tracking discourse complexity preceding Alzheimer's disease diagnosis: a case study comparing the press conferences of presidents ronald reagan and george herbert walker bush. J Alzheimers Dis. 2015;45(3):959–63.
24 Haider F, De La Fuente S, Luz S. An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. IEEE J Sel Top Signal Process. 2020;14(2):272–81.
25 Luz S, Haider F, de la Fuente Garcia S, Fromm D, MacWhinney B. Editorial: Alzheimer's dementia recognition through spontaneous speech. Front Comput Sci. 2021;3:780169.
26 Fox NC, Warrington EK, Seiffer AL, Agnew SK, Rossor MN. Presymptomatic cognitive deficits in individuals at risk of familial Alzheimer's disease. A longitudinal prospective study. Brain. 1998;121 (Pt 9)(9):1631–9.
27 Simon J, Bastin C, Salmon E, Willems S. Increasing the salience of fluency cues does not reduce the recognition memory impairment in Alzheimer's disease!. J Neuropsychol. 2018;12(2):216–30.
28 Pistono A, Jucla M, Barbeau EJ, Saint-Aubert L, Lemesle B, Calvet B, et al. Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimer's disease. J Alzheimers Dis. 2016;50(3):687–98.
29 Meilán JJG, Martínez-Sánchez F, Carro J, Sánchez JA, Pérez E. Acoustic markers associated with impairment in language processing in Alzheimer's disease. Span J Psychol. 2012;15(2):487–94.
30 de la Fuente Garcia S, Ritchie CW, Luz S. Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. J Alzheimers Dis. 2020;78(4):1547–74.
31 Covington MA, McFall JD. Cutting the Gordian knot: the moving-average type – token ratio (MATTR). J quantitative linguistics. 2010;17(2):94–100.
32 Brunet É. Le vocabulaire de Jean Giraudoux structure et évolution. Slatkine; 1978.
33 Honoré A. Some simple measures of richness of vocabulary. Assoc Lit Ling Comput Bull. 1979;7(2):172–7.
34 Tweedie FJ, Baayen RH. How variable may a constant be? Measures of lexical richness in perspective. Comput Humanities. 1998;32(5):323–52.
35 Boersma P, Weenink D. Praat. Doing phonetics by computer (version 5.2. 20) [Software].
36 Jadoul Y, Thompson B, De Boer B. Introducing parselmouth: a python interface to praat. J Phonetics. 2018;71:1–15.
37 Durán P, Malvern D, Richards B, Chipere N. Developmental trends in lexical diversity. Appl Linguistics. 2004;25(2):220–42.
38 Hernández-Domínguez L, Ratté S, Sierra-Martínez G, Roche-Bergua A. Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. Alzheimer's & Dementia: diagnosis. Alz Dem Diag Ass Dis Mo. 2018;10:260–8.
39 Saffran EM, Berndt RS, Schwartz MF. The quantitative analysis of agrammatic production: procedure and data. Brain Lang. 1989;37(3):440–79.
40 Berisha V, Krantsevich C, Hahn PR, Hahn S, Dasarathy G, Turaga P, et al. Digital medicine and the curse of dimensionality. NPJ Digit Med. 2021;4(1):153.