TOWARDS DISFLUENCY FEATURES FOR SPEECH TECHNOLOGY BASED AUTOMATIC DEMENTIA CLASSIFICATION

Megan Thomas[1*], Samuel Hollands[1*], Daniel Blackburn[2], Heidi Christensen[1]

[1]Department of Computer Science, University of Sheffield, UK

[2]Sheffield Institute for Translational Neuroscience, University of Sheffield, UK
[*]*Authors Contributed Equally*
{mcthomas2, shollands1, d.blackburn, heidi.christensen}@sheffield.ac.uk

ABSTRACT

Automatically detecting cognitive decline using speech and language technologies is a field of growing research. This domain deploys AI-based classification technologies into clinical environments to assist traditional diagnostic methods. This paper explores a new series of disfluency features to assist the widely adopted *acoustic-linguistic* elements of contemporary automatic dementia diagnosis.

Utilising automatic speech recognition, phoneme recognition, and voice activity detection technologies on the Sheffield IVA data [1], we show improved accuracy from a baseline of 78.4% to 83.2% using automatically calculated disfluency features. This paper contributes disfluency features as a method of enhancing current feature sets used in the automatic detection of cognitive decline, and highlights potential approaches towards reliably extracting these disfluency features automatically. Features that are less dependent on raw acoustic information are potentially more generalisable and robust for dementia classification tasks in varied environments, a key obstacle in this domain.

Keywords: Dementia, Disfluency, Speech Analytics, Classification

1. INTRODUCTION

Cognitive decline (CD) and diseases such as Alzheimer's Dementia (AD) pose a significant challenge to health services worldwide, with an estimated 139 million people worldwide living with dementia by 2030 [2]. As the need for early and accurate CD detection increases, a growing body of work within the speech and language technologies (SLTs) domain is looking at assisting doctors and clinicians in the diagnosis of CD through the use of automatic cognitive decline detection (ACDD) systems, specifically through the automatic analysis of speech [3]. Speech is an effective biomarker for this task as it is readily available, easy to collect, and contains features that correlate well to different levels of CD [4, 5, 6, 7].

Numerous studies have highlighted the effectiveness of different temporal measures of speech for discriminating between people with CD and healthy controls. Some temporal measures such as speech rate are even salient enough to differentiate between prodromal dementia and early stages of AD [8]. Pauses are another common feature frequently investigated for the purpose of automatically detecting CD, with correlations found between the number and duration of filled and unfilled pauses and levels of CD (for example in [9] and [10]). Despite this, little research has been conducted into the feasibility of using other disfluencies to aid in the detection of CD. This is due, in part, to the difficulties of including disfluencies in automatically produced transcripts.

This study, building on work from [11], explores the efficacy of using manually annotated disfluency features to aid in the automatic detection of cognitive decline. The capability of these features to enhance the performance of traditional *acoustic-linguistic* based systems is demonstrated, and initial experiments to automatically calculate these features are presented. Automation of these features is imperative in order to apply them to speech analytics based classification systems.

2. BACKGROUND

Recent successful ACDD systems have consisted of the following components. A virtual interface asks a patient some pre-written prompts and records the response [12]. An automatic speech recognition (ASR) system is used to transcribe their utterance into textual information. Two sets of features are typically generated; acoustic (non-verbal) and linguistic (verbal) [13]. Acoustic features include

Table 1: Disfluency features used within this study and their descriptions with constituent elements

Disfluency Feature	Description
Unfilled Pauses	Silence lasting >200ms
Filled Pauses	Sustained sound, typically a vowel or a sound that is not part of a word, lasting >200ms
Repetitions	Part word, whole word, or phrase repetitions
Prolongations	Prolongation of a phone lasting >200ms
Repairs	When a speaker alters what they had been saying
Speech Errors	Phonetic additions, substitutions, or deletions. selection errors, retrieval errors, and blends

metrics extracted directly from the audio recordings, and linguistic features include parameters extracted from a transcript of the speech to be analysed. These features are then fed into a classification system which is trained to discriminate between classes (or conditions).

Recent work from [11] suggests that disfluency information could be valuable in a CD detection task. A manual disfluency analysis found that features such as number of word repetitions and prolongations were significantly different between people at different stages of cognitive decline. The present study aims to use this disfluency information to inform an ACDD system alongside *acoustic-linguistic* features. However, automating this process is no simple feat. Typical ASR systems are designed to ignore disfluencies in speech and therefore disfluencies are not typically present in automatically created transcripts. Unlike "traditional" features used for similar tasks, disfluency features have the added benefit of being easily explainable. The issue of explainability and interpretability is especially relevant when working in the medical domain, as discussed in [14].

3. DATA

A subset of the Sheffield IVA dataset [1] was used for this project. This data is collected via a virtual agent as part of an ACDD system that poses a number of language tasks to patients and records their answers. The task used for this study was the Cookie Theft picture description task [15]. Participants belonged to one of three different diagnosis groups; healthy controls (HC), mild cognitive impairment (MCI), and neurodegenerative dementia (ND). The ND group contains participants with various types of dementia, which could also include AD. Each group contained 18 recordings. Recordings were selected for this study based on the intelligibility of the speech from the participants. Classification tasks within this study use 15 HC and 10 ND participants. In the larger IVA dataset there may be multiple recordings of the same speaker. Our sample contains only one recording for each participant.

4. METHODOLOGY

4.1. Feature Selection & Extraction

4.1.1. Manual Disfluency Features

A disfluency schema was created to determine exactly which disfluencies should be investigated. This was based largely on the work of [16]. After the work conducted in [11], minor changes were made to the disfluency schema resulting in the schema outlined in Table 1 above.

Praat [17] was used to manually annotate all instances of the above disfluencies within the recordings. Diagnosis labels were hidden from the annotator, a trained phonetician, to avoid any potential bias.

4.1.2. Automatic Disfluency Features

A Jupyter Notebook was created to calculate the disfluency features automatically. Voice activity detection (VAD) [18, 19] with a threshold of 50ms onset/offset and a 50ms minimum speech threshold (informed by the average length of voiced and unvoiced British English syllables [20]) was used to identify and parse the speech components of each file. A convolutional recurrent deep neural network (CRDNN) based Commonvoice [21] ASR system was used; this was preferred over a transformer based system given the potential for improved performance on non Southern British English accent groups [22]. For the syllabic (and particularly non-lexical) aspects of the task both phoneme recognition (PR) and grapheme to phoneme (G2P) technologies were explored. PR (using CMUSphinx) transcribes a phone given an audio signal input. G2P takes the output of an ASR system and transcribes back into phones. PR was found to outperform G2P and was therefore used for syllable parsing and all non-lexical feature

Table 2: Linguistic features used within this study [12], the number of features (n) within each category, and the automated elements needed for each feature to function

Tok: Tokeniser, POS: Part-of-Speech Tagger, W_List: Word List, CoRef: Co-reference Tagger, Sem_Tag: Semantic Tagger, POS_Pat: POS Pattern Matcher, Tree: Syntactic Tree Parser

Feature(s)	n=#	Automated Components
Content Density	1	Tok, POS, W_List
Part-of-Speech Rate	45	Tok, POS
Reference Rate to Reality	1	Tok, POS
Personal, Spatial and Temporal Deixis Rate	3	Tok, POS, W_List, CoRef
Relative Pronouns and Negative Adverbs Rate	2	Tok, POS, W_List
Lexical Richness	3	Tok
Action Verbs Rate	1	Tok, POS, Sem_Tag
Frequency-of-Use Tagging	1	Tok, W_List
Propositional Idea Density	1	Tok, POS, POS_Pat
Mean Number of Words in Utterance	1	Tok
Number of Dependent Elements Linked to the Noun	2	Tok, POS, Tree
Global Dependency Distance	2	Tok, POS, Tree
Syntactic Complexity	1	Tok, POS, Tree
Syntactic Embeddedness	2	Tok, POS, Tree
Utterance Length	2	Tok

calculations. The disjoint between the ASR and PR provided some inaccuracy; a robust ASR system for Northern British English with a phone level output would have resulted in better accuracy in this area.

4.1.3. Automatic Acoustic Features

The openSMILE toolkit [23] was used in this study for acoustic feature extraction using the Python API. eGeMAPSv02, Emobase, and ComParE were used as feature sets that have been well explored within the literature. All three feature sets were calculated for each of the participant recordings. 42 features containing SMB contour smoothing within ComParE were redacted due to a strong negative impact on support vector machine (SVM) accuracy due to the values being calculated as near binary measurements (distance away from min|max $< 10^{-5}$ when normalised).

4.1.4. Automatic Linguistic Features

Table 2 highlights the linguistic features extracted for this research [12]. These features were extracted using a variety of different tools highlighted within the *Automated Components* column. The libraries used were spaCy [24] (word tokenisation, POS tagging, coreference resolution, and semantic tagging), NLTK [25] (sentence tokenisation and tree parsing), and Re(gex) (POS pattern matching). A Jupyter Notebook was developed using Python 3.10 containing custom written functions to extract each feature.

4.2. Baseline System

The baseline classifier in this study was developed using exclusively interpretable metrics; we did not use features such as word embeddings as it was counter-productive to the ambition of finding interpretable solutions. A baseline two way classifier was built to discriminate HC-MCI/ND. An SVM was trained using acoustic feature sets; eGeMAPSv02, Emobase, and ComParE using the following hyperparameters {“C”: 100, “gamma”: 0.001, “kernel”: “rbf”}. All SVMs were developed using 5-fold cross validation. Baseline performance was 78.4% for HC-MCI/ND. The baseline system used exclusively acoustic features for reasons discussed in Section 5.

For our disfluency features, normalisation was explored using three options; relative to speech duration, instances per 100 syllables, and relative to word frequency. Word frequency was the most

underperformant normalisation metric hindering performance of the disfluency features to below the baseline. This may be due to the syllabic nature of disfluencies and that an ASR system does not provide a reliable output for features such as speech errors and non-speech events, limiting the utility of word frequency as a way of measuring the length of discourse. Speaker duration worked well, achieving performance slightly above the baseline at 79.2% (HC-MCI/ND). Syllabic relative frequency, a common method for counting disfluencies, was the most effective outperforming the baseline with 83.2% (HC-MCI/ND). Syllabic relative frequency was not found to be beneficial for the normalisation of textual features, therefore the final SVMs utilised conventional normalisation as detailed in [12] for these features.

5. RESULTS & DISCUSSION

The linguistic features used in this study were developed for what is arguably the most popular corpus of ND speech; the DementiaBank Pitt Corpus [26]. Tested on Pitt Cookie Theft, an SVM was built with an accuracy of 78.8% on a 2-class HC-MCI/ND discrimination task; however these features perform with an accuracy of 49.6% on the Sheffield IVA data. There is a large limitation in this area of research caused by data scarcity and mono-corpus challenge based research that results in difficulties in determining the generalisability of feature sets across different corpora. These results highlight the need to test ACDD systems on unseen corpora to ensure reliable efficacy outside of a single data set. For this reason, linguistic features were not included in our baseline. A preliminary series of experiments on the Pitt Corpus demonstrated enhanced performance with linguistic features and disfluency features combined. Therefore these features were retained for the final Sheffield IVA classification system to acknowledge the potential of these features to increase generalisability.

Table 3: HC-MCI/ND SVM Accuracy Results (5-fold cross validation)
{M-#: Manual, A-#: Automatic}

Model	Accuracy
Acoustic Only	77.6%
Linguistic Only	57.6%
Acoustic + Linguistic	68.8%
Manual Disfluency	**88.8%**
Automatic Disfluency	78.4%
Ac + Ling + Auto Disfluency	83.2%

An SVM trained exclusively on the manually annotated disfluency features yielded a baseline performance of 88.8%. This strongly suggests human annotated disfluency features have the capacity to enhance the performance of existing models. In order to automate disfluency feature extraction certain features were omitted. Blends and substitutions (categorised as speech errors in the above schema) were removed from the feature bank due to falling below a threshold of 10 instances within our subset of the Sheffield IVA dataset.

Table 3 demonstrates improved performance of 5.6% from the acoustic only SVM baseline trained on the Sheffield IVA data when using an SVM enhanced with automatically calculated disfluency features. The performance of the manually annotated features remains substantially higher at 88.8%. Overall these results demonstrate that disfluency features provide meaningful performance improvements to ACDD SVMs in the context of HC-MCI/ND discrimination when using conventional and interpretable linguistic/acoustic feature sets.

6. CONCLUSION

This study has demonstrated the potential for enhancing ACDD systems through the automatic extraction of disfluency features. The difference between manual and automatic feature performance remains large; whilst this paper has demonstrated the potential for boosting system efficacy more research is needed to optimise approaches for extracting each feature. Future research may wish to focus on a larger scale analysis of disfluency feature generalisability. Our future research will focus on the behaviour of linguistic features to supplement this work on disfluency features and aim to inform greater intuition about the behaviour of different modes of automated features in real world environments.

7. ACKNOWLEDGEMENTS

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]

8. REFERENCES

[1] B. Mirheidari, D. Blackburn, R. O'Malley, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Computational cognitive assessment: Investigating the use of an intelligent

virtual agent for the detection of early signs of dementia," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2732–2736.
[2] W. H. Organization *et al.*, "Global status report on the public health response to dementia," 2021.
[3] M. L. B. Pulido, J. B. A. Hernández, M. Á. F. Ballester, C. M. T. González, J. Mekyska, and Z. Smékal, "Alzheimer's disease and automatic speech analysis: a review," *Expert systems with applications*, vol. 150, p. 113213, 2020.
[4] K. Croot, J. R. Hodges, J. Xuereb, and K. Patterson, "Phonological and articulatory impairment in alzheimer's disease: a case series," *Brain and language*, vol. 75, no. 2, pp. 277–309, 2000.
[5] P. Gómez-Vilda, V. Rodellar-Biarge, V. Nieto-Lluis, K. L. de Ipiña, A. Álvarez-Marquina, R. Martínez-Olalla, M. Ecay-Torres, and P. Martínez-Lage, "Phonation biomechanic analysis of alzheimer's disease cases," *Neurocomputing*, vol. 167, pp. 83–93, 2015.
[6] N. Biassou, M. Grossman, K. Onishi, J. Mickanin, E. Hughes, K. Robinson, and M. D'Esposito, "Phonologic processing deficits in alzheimer's disease," *Neurology*, vol. 45, no. 12, pp. 2165–2169, 1995.
[7] A. J. Astell and T. A. Harley, "Tip-of-the-tongue states and lexical access in dementia," *Brain and language*, vol. 54, no. 2, pp. 196–215, 1996.
[8] V. Vincze, G. Szatlóczki, L. Tóth, G. Gosztolya, M. Pákáski, I. Hoffmann, and J. Kálmán, "Telltale silence: temporal speech parameters discriminate between prodromal dementia and mild alzheimer's disease," *Clinical Linguistics & Phonetics*, vol. 35, no. 8, pp. 727–742, 2021.
[9] V. Vincze, G. Gosztolya, L. Tóth, I. Hoffmann, G. Szatlóczki, Z. Bánréti, M. Pákáski, and J. Kálmán, "Detecting mild cognitive impairment by exploiting linguistic information from transcripts," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016, pp. 181–187.
[10] K. López-de Ipiña, J.-B. Alonso, C. M. Travieso, J. Solé-Casals, H. Egiraun, M. Faundez-Zanuy, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage *et al.*, "On the selection of non-invasive methods based on speech analysis oriented to automatic alzheimer disease diagnosis," *Sensors*, vol. 13, no. 5, pp. 6730–6745, 2013.
[11] M. Thomas, N. Pevy, and T. Walker, "Disfluencies and cognitive decline: An investigative study," in *ICPLA 2023- Conference of the International Clinical Phonetics and Linguistics Association*, 2023.
[12] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.
[13] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, "Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 342–356, 2020.
[14] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.
[15] H. Goodglass and E. Kaplan, *The assessment of aphasia and related disorders*. Lea & Febiger, 1972.
[16] K. McDougall and M. Duckworth, "Profiling fluency: An analysis of individual variation in disfluencies in adult males," *Speech Communication*, vol. 95, pp. 16–27, 2017.
[17] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer." [Online]. Available: http://www.praat.org/
[18] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.
[19] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Interspeech 2021*, Brno, Czech Republic, August 2021.
[20] U. Gut, "Analysing phonetic and phonological variation on the suprasegmental level," *Research Methods in Language Variation and Change*, pp. 244–259, 2013.
[21] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.
[22] S. Hollands, D. Blackburn, and H. Christensen, "Evaluating the Performance of State-of-the-Art ASR Systems on Non-Native English using Corpora with Extensive Language Background Variation," in *Proc. Interspeech 2022*, 2022, pp. 3958–3962.
[23] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.
[24] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020.
[25] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.
[26] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of neurology*, vol. 51, no. 6, pp. 585–594, 1994.