Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Anne Rowling Neurological Speech Corpus: clinically annotated longitudinal dataset for developing speech biomarkers in neurodegenerative disorders

Johnny Tam[1−4]*, Christine Weaver*[1−4]*, Oliver Watts*[1,5]*, Siddharthan Chandran*[1−4]*, Suvankar Pal*[1−4]*, Rowling Speech Consortium*[1−5]

[1]Anne Rowling Regenerative Neurology Clinic, Edinburgh, UK [2]Centre for Clinical Brain Sciences, University of Edinburgh, UK [3]Euan MacDonald Centre for MND Research, University of Edinburgh, UK [4]UK Dementia Research Institute, UK [5]SpeakUnique Ltd, Edinburgh, UK

jtam@ed.ac.uk

Abstract

There is urgent need for scalable, non-invasive and quantifiable biomarkers in neurodegenerative disorders. Speech is an attractive candidate with potential for remote and cheap assessments. Progress is limited by a lack of high quality clinically annotated speech data. We present a longitudinal speech corpus including speakers with dementia, motor neuron disease, Parkinson's disease, progressive multiple sclerosis, and healthy individuals. Participants complete standardised recordings on an app co-produced with patients, aligned to contemporaneous phenotyping (clinical rating scales, cognitive tests and blood-based biomarkers). 780 participants have provided 5169 recordings in 1033 assessments. Benchmark classification and regression models show promising performance, and predictions on non-speech segments demonstrate limited bias from recording conditions. We continue to upscale data collection and analysis across larger diverse populations to accelerate clinical translation.

Index Terms: cognition, dementia, dysarthria, health, memory, neurodegeneration, speech pathology

1. Introduction

1.1. Background

Neurodegenerative disorders including dementia, Parkinson's disease (PD), motor neuron disease (MND), and progressive multiple sclerosis (MS) constitute a major global public health challenge that exerts significant personal and societal burden [1]. These conditions are clinically heterogeneous, progressive, often associated with persistent, pervasive, and functionally impairing motor and cognitive symptoms, and have limited treatment options. Individuals are intermittently assessed in the clinical setting and there is an urgent unmet need for the discovery and validation of biomarkers to accelerate i) improved understanding of pathophysiology of disease, ii) early and accurate diagnosis, iii) subtyping of heterogeneous disease groups, iv) prognostication, v) disease monitoring and vi) outcome measures and engagement evaluation in clinical trials.

1.2. Speech as a marker of neurodegeneration

Acoustic and linguistic speech changes occur in all neurodegenerative disorders. Digital technologies for acquisition and analysis of speech have potential to deliver more reliable, objective, remote, scalable, and sensitive measures of change than routinely used clinical rating scales, and may be useful as prognostic and diagnostic tools. In affected individuals, distinctive patterns of speech dysfunction can be ascertained in different neurodegenerative diseases to aid the diagnostic process. As disease progresses, change in speech function can be monitored as a sensitive marker of disease status. In the prodromal, or 'at-risk' phase, speech is often also affected , thereby offering a marker for early detection.

Existing state of the art methods in speech signal processing and machine learning show promising performance in detecting disease and predicting disease severity [2, 3]. With robust clinical validation, these novel methods hold potential to produce an accurate, sensitive, and generalisable digital biomarker. Publicly available neurological speech datasets such as DementiaBank[1] and TORGO [4] have driven recent advances in this area. However, most of these datasets are cross-sectional, focus on single disorders, and are limited in participant numbers and ground-truth clinical data that can be used for development and clinical validation. Previous studies have also shown that some of these datasets suffer from a significant amount of bias from recording conditions [5]. A large volume of longitudinal and clinical annotated data is required for robust methods to be developed, and for proper clinical adoption of these technologies to occur.

1.3. Aims & objectives

We aim to develop the Anne Rowling Neurological Speech Corpus (ARNSC), a large longitudinal dataset of speech from people living with multiple neurodegenerative disorders, accompanied with detailed and contemporaneous clinical annotations. We describe the current speech and clinical data characteristics of the corpus, and present early benchmark results that showcase the informative signals present for development of clinical speech systems.

2. Methods

2.1. Data sources

Data for the ARNSC are collected in multiple clinical studies at the Anne Rowling Regenerative Neurology Clinic, where we have implemented and integrated a unified speech recording pipeline. Current active studies are:

- The Scottish National MND Register (CARE-MND) [6].
- NeuroCARE Research Database[2], a research registry for people living with neurological disorders and their carers and relatives.
- Digital App for Speech & Health Monitoring (DASH)[3], a prospective longitudinal digital observational study.

[1]https://dementia.talkbank.org/

[2]https://www.annerowlingclinic.org/research-trials/find-project/neurocare

[3]https://www.annerowlingclinic.org/research-trials/find-project/dash-digital-app-for-speech-health-monitoring

10.21437/Interspeech.2025-2000

Table 1: *Standardised speech recording tasks implemented in the speech recording app.*

Task	Type	Stimulus
Warm-up	Spontaneous	"What have you been up to this week?"
Reading 1	Scripted	Speech Accent Archive elicitation paragraph
Reading 2	Scripted	Rotating passages
Picture description task	Spontaneous	Rotating pictures, including the Cookie Theft Image
Phonetic fluency task	Spontaneous	F, A, S, T
Feedback (optional)	Spontaneous	"Please tell us about your experience using the app today!"

Integration within large scale interventional and observational studies in MND, MS, PD and Alzheimer's disease are planned. A snapshot of the corpus from 5th February 2025 was used for this paper.

2.2. Eligibility

In addition to eligibility criteria of individual studies in the ARNSC, participants also meet the following inclusion and exclusion criteria:

- **Inclusion:** Any one of the following: an individual with a diagnosis of cognitive disorder (all causes), PD, MND, and MS, a relative or carer of the above who does not report to have a neurological condition, or a healthy volunteer who does not report to have a neurological condition.
- **Exclusion:** Any one of the following: age under 16 years, significant and uncorrected visual or hearing impairment (precluding use of the App), or lacks capacity to consent to speech recordings due to cognitive impairment.

2.3. Participant groups

Participants are grouped according to documented diagnosis. Subgroups within each group allow for further stratification. Additional groups and subgroups may be added in the future.

- MND
  - Amyotrophic lateral sclerosis (ALS), primary lateral sclerosis (PLS), progressive bulbar palsy (PBP), progressive muscular atrophy (PMA)
  - Bulbar, limb and respiratory onset
- Dementia and cognitive disorders (all causes)
  - Alzheimer's disease (AD) and mild cognitive impairment (MCI)
- Multiple sclerosis (MS)
  - Relapsing remitting, primary progressive and secondary progressive
- Parkinson's disease (PD)
- Control (no known neurological diagnosis)

2.4. Assessment procedure

2.4.1. Speech recording app

All participants are assessed with a unified speech recording pipeline using a progressive web app[4] designed and co-produced with patient and public involvement and engagement (PPIE) input to have a user-friendly interface for speech acquisition that accommodates cognitive, visual, and motor deficits.As a web interface it is accessible via a web browser, and is compatible with all modern computer and mobile operating systems (including iOS, Android, Windows, macOS), minimising barriers to participation due to incompatibility.

[4]Example for the DASH Study: https://dash.ed.ac.uk

2.4.2. Speech recording tasks

Participants complete six speech recording tasks using the app, consisting of tasks that aim to elicit both scripted and spontaneous speech (Table 1). These tasks were selected due to their use in routine clinical practice (warm-up), other validated scales and tests (picture description, phonetic fluency tasks), or in other large speech corpus (fixed passage). Speech recording tasks are completed either under researcher supervision during clinic-based assessments, or unsupervised when completed remotely in the community. Additionally, participants are invited to provide an optional sixth recording where they provide open-ended feedback on participant experience of the tasks. All recording tasks were completed in British English.

2.4.3. Clinical annotation

Recordings are accompanied by deep clinical phenotyping undertaken by each study in the ARNSC. This includes:

- **Demographics** (age, sex, educational qualification, first language, accent)
- **Disease characteristics** (diagnosis group and subgroup, time from onset and diagnosis)
- **Disease severity** (ALS Functional Rating Scale Revised (ALSFRS-R), Clinical Dementia Rating (CDR), Expanded Disability Status Scale (EDSS), MDS-Unified Parkinson's Disease Rating Scale (MDS-UPDRS))
- **Cognitive tests** (Montreal Cognitive Assessment (MoCA), Addenbrooke's Cognitive Examination III (ACE-III), Edinburgh Cognitive and Behavioural ALS Screen (ECAS))
- **Blood biomarkers** (neurofilament light chains (Nfl), glial fibrillary acidic protein (GFAP), amyloid beta 40 and 42, phosphorylated tau at threonine 181 and 231)

Annotations are tailored to the participant group to provide more detailed ground-truth for development and evaluation of clinical models.

2.5. Ethics

All participating studies in the ARNSC have been reviewed and approved by NHS Research Ethics Committees and local NHS R&D departments. Participants are fully informed about the purpose, usage, and handling of their data. They are informed that due to the unique nature of speech, it is not possible to completely anonymise recordings collected. Speech, as a special category of personal data, is specifically consented for in terms of storage, use and future sharing.

2.6. Data Sharing

In alignment with open science and FAIR data principles [7], this study will endeavour to make data as accessible and usable as possible while ensuring data privacy and security are maintained. Standardised datasets may be made available via future speech processing challenges with specific tasks and evaluation criteria aimed at promoting innovation and benchmarking progress in the field of speech processing.

2.7. Benchmark analysis

Benchmark analysis was conducted in the MND, cognitive disorders and control groups to provide reference performance results for the data collected so far. Machine learning models were trained for the following tasks:

- Classification of disease groups (MND, cognitive disorders) and controls.
- Classification of disease subgroups (ALS, AD, MCI) and controls.
- Regression of ALSFRS-R speech sub scores

The data was subset to only the latest assessment from each participant, with audio samples from each assessment concatenated to produce one combined audio file per individual for feature extraction. For classification tasks, one-to-one coarsened exact matching[5] was conducted on age and sex to match and balance datasets, with a criteria for standardised mean difference (SMD) of <0.1.

2.7.1. Non-speech segment predictions

For each task, predictive performance of non-speech segments from the same audio samples are also presented. As demonstrated in previous studies [5], this provides a measure of the systematic bias present in the data, by presenting how much of a model's performance may be attributed to confounding characteristics present when speech is absent (such as recording quality, background noise). Non-speech segments were extracted using a pre-trained model.[6]

2.7.2. Features

Two types of acoustic features were extracted from samples to be used as inputs in machine learning models:

- **Conventional** - Geneva Minimalistic Acoustic Parameter Set 2 (eGeMAPSv02) in the openSMILE tooklit represents conventionally extracted acoustic features [8].
- **Deep learning** - Embeddings from publicly available pretrained deep learning models were extracted to represent emerging deep and transfer learning methods.

Three deep learning models identified in a separate benchmark by the authors were included in this analysis:

- **UniSpeech** - Fine-tuned version of *microsoft/unispeech-large-1500h-cv* on the *TIMIT_ASR-NA* dataset[7]
- **TitaNet** - TitaNet-L for speaker verification and diarisation[8]
- **CRDNN+CTC** - SpeechBrain automatic speech recognition model trained on LibriSpeech (EN)[9]

[5]https://cran.r-project.org/web/packages/MatchIt/
[6]https://huggingface.co/pyannote/speaker-diarization-3.1
[7]https://huggingface.co/patrickvonplaten/unispeech-large-1500h-cv-timit
[8]https://huggingface.co/nvidia/speakerverification_en_titanet_large
[9]https://huggingface.co/speechbrain/asr-crdnn-rnnlm-librispeech

2.7.3. Machine learning models

Random forest classification and regression models from *scikit-learn*[10] were used to benchmark extracted features. With the exception of *n_estimators* which was set to 1000, default hyperparameters were used. Models were evaluated in 5-fold cross validation with grouping by speaker identity to prevent data leakage across train and test sets in each fold. For classification tasks performance was reported using accuracy, which is valid as datasets are balanced. For regression tasks R^2 (coefficient of determination) was used.

3. Results

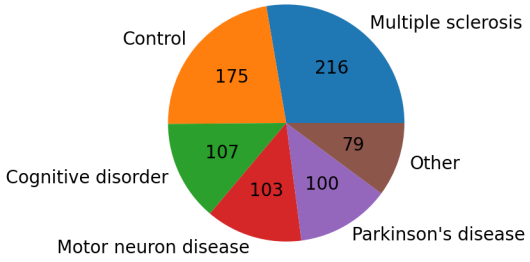

Figure 1: *Pie chart showing counts of participant groups in 780 unique participants.*

3.1. Dataset characteristics

5169 recordings in 1033 assessments from 780 unique participants were present in the ARNSC. Breakdown of participant groups in the corpus are shown in Figure 1. 134 (17%) participants have completed multiple assessments at different time points, including 48 with MND, 38 with cognitive disorder, 26 controls, 14 with MS and 8 with PD. Time interval between recordings range between 2 and 26 months.

3.1.1. Demographics

Overall mean age at the time of recording was 60 (range: 22-96, IQR 52-70) with male:female ratio (MFR) of 0.95. In matched datasets for benchmark analysis, demographics of MND (mean age 64, MFR 1.36) and cognitive disorder (mean age 66, MFR 1.25) were comparable to recent epidemiological studies, although for cognitive disorders participants were younger (66 vs 76) [9, 10].

3.1.2. Clinical annotation

Recordings were accompanied by 25,477 contemporaneous clinical annotations, which included:

- Disease severity: 187 ALSFRS-R and 24 CDR
- Cognitive tests: 142 ACE-III, 158 MoCA and 165 ECAS
- Blood biomarkers: 67 Nfl and GFAP

3.2. Predicting MND and cognitive disorder from controls

Figure 2 summarises predictive accuracies of classifiers trained on extracted conventional and deep learning features. For MND

[10]https://scikit-learn.org/stable/

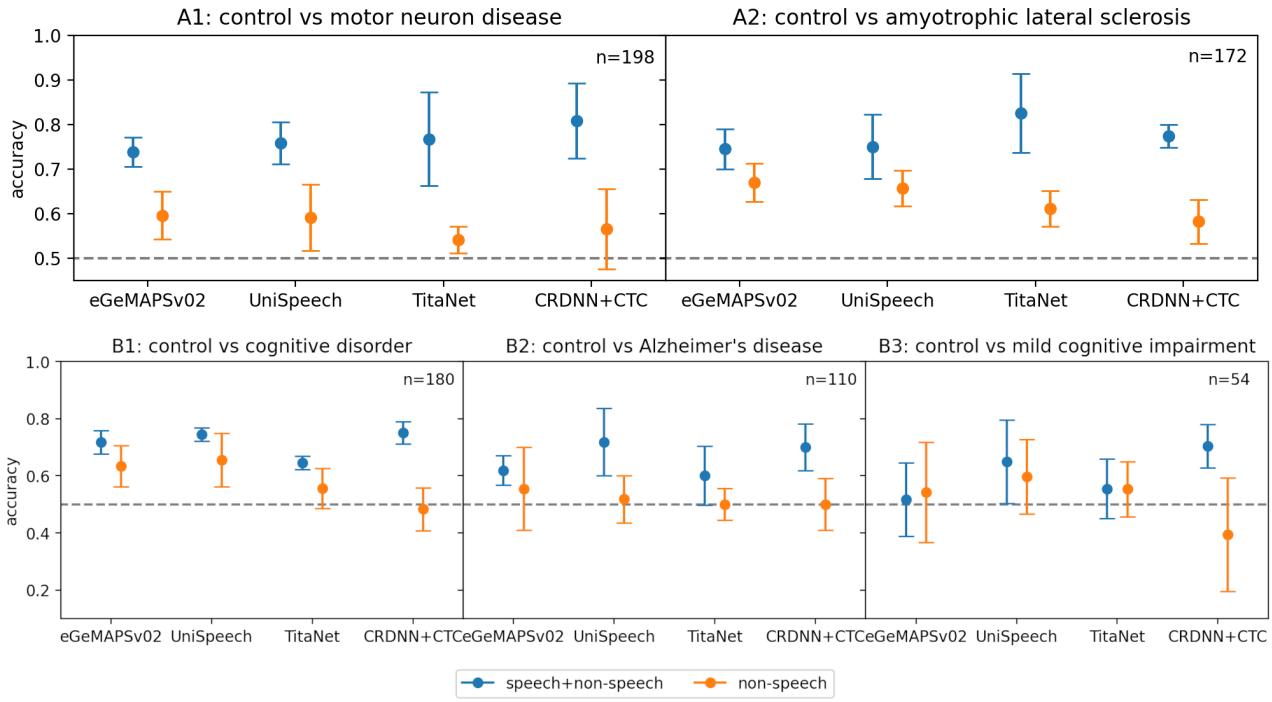

Figure 2: *Disease classifier accuracies from 5-fold cross validation for MND and cognitive disorder groups (A1, B1) and their aetiological subgroups (A2, B2, B3). ● represent mean accuracy, error bars represent standard deviation of accuracies from CV folds. Dotted line denotes randomness level (0.5).*

and cognitive disorder, promising accuracies of between 70-80% are reported across all features. A similar pattern is present when data is subset by aetiological subgroups (ALS, AD and MCI). Predictions from non-speech segments are either random or close to random, and have generally lower accuracies than predictions on both speech and non-speech segments. This suggests that only a small portion of the model's performance can be attributed to assumed confounding characteristics in non-speech segments, such as background noise and recording quality. Of note is the pattern that deep learning features seem have worse performance in non-speech segments compared to eGeMAPSv02, which suggest they may be more robust to these confounding characteristics in this dataset. Some parts of the non-speech signals may be true disease signature however, such as hesitancies, pauses and other non-speech sounds, which have been shown to be significant especially in cognitive disorders [11].

3.3. Predicting ALSFRS-R speech sub-scores

Figure 3 summarises regressors R^2 performance on ALSFRS-R speech sub-score predictions. With the exception of *titanet* all features achieve mean R^2 of between 0.40 and 0.48, with deep learning features being slightly more stable across folds than eGeMAPSv02. Performance of predictions on non-speech segments also display a similar pattern to classifier results in Figure 2, suggesting a similar robustness against assumed confounding characteristics in non-speech segments.

4. Conclusions

We present early findings from an ambitious and ongoing effort to create a large, longitudinal and deeply clinically annotated

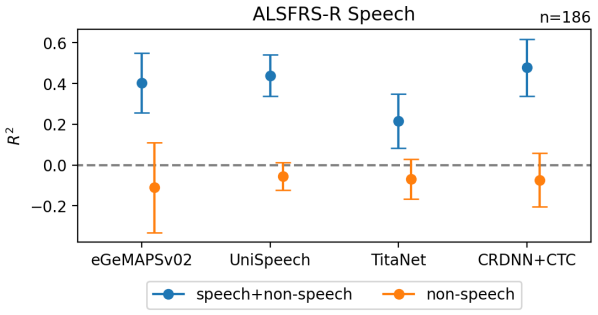

Figure 3: *ALSFRS-R speech sub-score regressors R^2 as evaluated with 5-fold cross validation (CV). ● represent mean R^2 and error bars represent standard deviation of R^2 from CV folds. Dotted line denotes randomness level (0).*

speech dataset for multiple neurodegenerative disorders. The size and depth of speech and contemporaneous clinical data in our corpus demonstrates the potential scalability and accessibility that speech biomarkers may add to clinical and research applications. Our benchmark analysis demonstrates significant informative signals in the corpus for both disease classification and regression to ground truth disease severity data. Furthermore we demonstrate limited bias due to confounding characteristics present in non-speech segments of audio, an issue that has plagued other publicly available datasets [5]. To our knowledge this is the largest multi-disorder neurological speech dataset, and progress is underway to increase scale and diversity of data collection and analysis in different patient populations, and develop performant and generalisable clinical speech systems for neurodegenerative disorders.

5. References

[1] A. J. Ferrari *et al.*, "Global incidence, prevalence, years lived with disability (ylds), disability-adjusted life-years (dalys), and healthy life expectancy (hale) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the global burden of disease study 2021," *The Lancet*, vol. 403, no. 10440, pp. 2133–2161, 2024. Link (data/code): https://www.sciencedirect.com/science/article/pii/S0140673624007578

[2] M. Bowden *et al.*, "A systematic review and narrative analysis of digital speech biomarkers in motor neuron disease," *npj Digital Medicine*, vol. 6, p. 228, 12 2023.

[3] S. de la Fuente Garcia, C. W. Ritchie, and S. Luz, "Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: A systematic review," *J. Alzheimers. Dis.*, vol. 78, pp. 1547–1574, 2020. Link (data/code): https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7836050/

[4] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 12 2012. Link (data/code): https://link.springer.com/article/10.1007/s10579-011-9145-0

[5] G. Schu, P. Janbakhshi, and I. Kodrasi, "On Using the UA-Speech and Torgo Databases to Validate Automatic Dysarthric Speech Classification Approaches," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2023-June, 2023.

[6] D. Leighton *et al.*, "Clinical audit research and evaluation of motor neuron disease (care-mnd): a national electronic platform for prospective, longitudinal monitoring of mnd in scotland," *Amyotrophic lateral sclerosis & frontotemporal degeneration*, vol. 20, pp. 242–250, 4 2019. Link (data/code): https://pubmed.ncbi.nlm.nih.gov/30889975/

[7] M. D. Wilkinson *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific Data 2016 3:1*, vol. 3, pp. 1–9, 3 2016. Link (data/code): https://www.nature.com/articles/sdata201618

[8] F. Eyben *et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing; The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," 2016. Link (data/code): http://www.speech.kth.se/wavesurfer/

[9] D. J. Leighton *et al.*, "Changing epidemiology of motor neurone disease in Scotland," *Journal of neurology*, vol. 266, no. 4, pp. 817–825, 4 2019. Link (data/code): https://pubmed.ncbi.nlm.nih.gov/30805795/

[10] J. J. Manly *et al.*, "Estimating the Prevalence of Dementia and Mild Cognitive Impairment in the US: The 2016 Health and Retirement Study Harmonized Cognitive Assessment Protocol Project," *JAMA Neurology*, vol. 79, no. 12, pp. 1242–1249, 12 2022. Link (data/code): https://jamanetwork.com/journals/jamaneurology/fullarticle/2797274

[11] E. L. Campbell, R. Y. Mesía, L. Docio-Fernandez, and C. García-Mateo, "Paralinguistic and linguistic fluency features for Alzheimer's disease detection," *Comput. Speech Lang.*, vol. 68, p. 101198, 7 2021.