


RESEARCH ARTICLE

Automatic classification of AD pathology in FTD phenotypes using natural speech

Sunghye Cho¹  | Christopher A. Olm² | Sharon Ash² | Sanjana Shellikeri² | Galit Agmon² | Katheryn A. Q. Cousins² | David J. Irwin² | Murray Grossman² | Mark Liberman¹ | Naomi Nevler²

¹Linguistic Data Consortium, Department of Linguistics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

²Penn Frontotemporal Degeneration Center, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Correspondence

Sunghye Cho, Linguistic Data Consortium, University of Pennsylvania, 3600 Market Street, Philadelphia, PA 19104, USA.
Email: csunghye@ldc.upenn.edu

Funding information

National Institutes of Health, Grant/Award Numbers: P01-AG066597, K99-AG073510-01, P30-AG072979; Department of Defense, Grant/Award Numbers: W81XWH-20-1-0531, AL220035; Alzheimer's Association, Grant/Award Number: AARF-21-851126

Abstract

INTRODUCTION: Screening for Alzheimer's disease neuropathologic change (ADNC) in individuals with atypical presentations is challenging but essential for clinical management. We trained automatic speech-based classifiers to distinguish frontotemporal dementia (FTD) patients with ADNC from those with frontotemporal lobar degeneration (FTLD).

METHODS: We trained automatic classifiers with 99 speech features from 1 minute speech samples of 179 participants (ADNC = 36, FTLD = 60, healthy controls [HC] = 89). Patients' pathology was assigned based on autopsy or cerebrospinal fluid analytes. Structural network-based magnetic resonance imaging analyses identified anatomical correlates of distinct speech features.

RESULTS: Our classifier showed 0.88 ± 0.03 area under the curve (AUC) for ADNC versus FTLD and 0.93 ± 0.04 AUC for patients versus HC. Noun frequency and pause rate correlated with gray matter volume loss in the limbic and salience networks, respectively.

DISCUSSION: Brief naturalistic speech samples can be used for screening FTD patients for underlying ADNC in vivo. This work supports the future development of digital assessment tools for FTD.

KEYWORDS

Alzheimer's disease, automated speech analysis, frontotemporal lobar degeneration, machine learning classification, natural speech, pathology

Highlights

- We trained machine learning classifiers for frontotemporal dementia patients using natural speech.
- We grouped participants by neuropathological diagnosis (autopsy) or cerebrospinal fluid biomarkers.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *Alzheimer's & Dementia* published by Wiley Periodicals LLC on behalf of Alzheimer's Association.

- Classifiers well distinguished underlying pathology (Alzheimer's disease vs. frontotemporal lobar degeneration) in patients.
- We identified important features through an explainable artificial intelligence approach.
- This work lays the groundwork for a speech-based neuropathology screening tool.

1 | BACKGROUND

Alzheimer's disease (AD) affects > 50 million individuals globally,¹ and much emphasis has been given to cognitive profiling of AD, including in the language domain. The pathology of AD consists of abnormal accumulation of extracellular amyloid beta ($A\beta$) plaques and intracellular neurofibrillary tangles. Previous studies have shown that individuals with amnesic AD presentation produce an abundance of generic words, circumlocutions, and few content units.²⁻⁸ One of the non-amnesic AD presentations, logopenic variant primary progressive aphasia (lvPPA), is characterized by slow speaking rate with frequent partial words or false starts and frequent, long pauses due to word-finding difficulties.^{9,10} A recent study identified several linguistic similarities between people with the amnesic AD presentation and those with a non-amnesic lvPPA presentation with AD neuropathologic change (ADNC), suggesting that speech features may be useful in identifying individuals with ADNC, regardless of their clinical presentation.¹¹

Early in vivo identification of underlying ADNC can be crucial to the success of future disease-modifying treatments. However, ADNC is observed not only in individuals with the typical, amnesic clinical presentation but also in people with atypical, non-amnesic presentations.^{12,13} Atypical presentations of ADNC with predominant linguistic and behavioral changes can present a challenging differential diagnosis among subtypes of frontotemporal dementia (FTD) due to frontotemporal lobar degeneration (FTLD), where common presentations include behavioral variant FTD (bvFTD) and primary progressive aphasia (PPA). Because novel therapeutics target underlying neuropathologies, it is essential to differentiate FTD patients with underlying ADNC from those with underlying FTLD. Developing in vivo assessments that identify the underlying pathology will help better guide patient management.

Indeed, speech has been studied as an objective tool for automatically differentiating people with amnesic AD presentation from healthy controls (HC). For example, a large number of lexical, syntactic, and acoustic features was used to classify patients with the amnesic AD presentation versus HC in the DementiaBank corpus with 82% accuracy.¹⁴ Jarrold et al. also implemented a similar approach and showed 88% accuracy with 18 participants (9 AD and 9 HC).¹⁵ A more recent study used a large dataset and a natural language processing (NLP) approach to differentiate individuals with dementia from HC with 87.1% accuracy and to distinguish individuals with mild cognitive impairment (MCI) from HC with 71.2% accuracy.¹⁶ Other reports

based on language and speech features ranged from 70% to 89.6% accuracy.¹⁷⁻²⁵

While previous studies have shown promising results with speech features classifying the amnesic AD presentation versus HC, a clinically meaningful challenge that has not been addressed is distinguishing underlying ADNC from FTLD pathology in individuals presenting with non-amnesic phenotypes. In this study, we used digital speech features that were extracted with automatic lexical and acoustic processing pipelines from digitized 1 minute picture descriptions, and we trained machine learning classifiers to automatically classify participants with clinical FTD and biological evidence of the underlying pathology (based on available autopsy or cerebrospinal fluid [CSF] biomarkers) into underlying ADNC or FTLD groups. We asked the following research questions: (1) Can speech and language features distinguish pathology in individuals with non-amnesic FTD phenotypes? (2) What features distinguish individuals with ADNC from those with FTLD? (3) How does a speech-based classifier perform compared to classifiers trained with traditional cognitive tests and basic demographics? (4) What specific features are related to patterns of regional brain atrophy in participants? We trained machine learning classifiers with different feature sets and applied an interpretable artificial intelligence (AI) approach to answer these questions. We also performed neuroimaging analyses relating speech features to regional brain volumes to provide anatomical validation for this biologically defined sample. The main goal of our study was to test machine learning classifiers trained with speech features extracted from brief naturalistic speech samples to identify the underlying ADNC versus FTLD pathology of participants presenting with a non-amnesic phenotype.

2 | METHODS

2.1 | Participants

We conducted a retrospective study and used brief oral picture descriptions that were produced by 179 participants (Table 1). All patients ($n = 96$) were clinically evaluated by expert neurologists (M.G., D.J.I.) at the time of recording at the frontotemporal degeneration center (FTDC) in the neurology department of the Hospital of the University of Pennsylvania, and they were classified at a weekly consensus meeting at the FTDC using established clinical criteria.^{9,26,27} Patients presented with various clinical phenotypes, including lvPPA, non-fluent/agrammatic variant PPA (naPPA), and semantic variant PPA

(svPPA) as detailed in Table 1. We note that one svPPA patient and another lvPPA patient in the AD group later developed (>1 year) clinical symptoms of posterior cortical atrophy and dementia with Lewy bodies, respectively. Patients were included if they had either a *post mortem* neuropathologic diagnosis or an *in vivo* CSF analyte profile suggestive of the underlying neuropathology based on cut-offs previously validated in autopsy series.^{28,29} We excluded patients with primary psychiatric conditions or other non-neurodegenerative conditions that could affect speech or cognition. Autopsy analyses classified the level of ADNC from “none” to “high” using an established scoring system when available.^{30,31} We grouped patients by their underlying pathology (Table 1). Twenty-one participants had primary ADNC based on CSF (phosphorylated-tau [p-tau]/A β 42 \geq 0.1²⁸ and total-tau [t-tau]/A β 42 \geq 0.34²⁹; $n = 21$); 15 patients had a primary neuropathological diagnosis of ADNC (high likelihood) per autopsy.³⁰ Sixty patients with FTD clinical symptoms had FTLN pathology by either CSF (p-tau/A β 42 < 0.1²⁸ and t-tau/A β 42 < 0.34²⁹; $n = 36$) or a neuropathological diagnosis per autopsy (e.g., corticobasal degeneration, Pick's disease, FTLN with TDP-43 immunoreactivity pathology, FTD with parkinsonism linked to chromosome 17, progressive supranuclear palsy, $n = 24$), with negligible AD co-pathology (none or low ADNC). One FTLN patient with intermediate ADNC was excluded from the dataset due to the clinically relevant ADNC comorbidity. Exclusion criteria for the autopsy series included clinical presentations other than FTD, primary pathology other than AD or FTLN, and co-occurring neurological or psychiatric conditions.

Among the 83 healthy participants, 45 were tested at the FTDC and had Mini-Mental State Examination (MMSE) scores. The other 38 healthy participants were volunteers who remotely provided their picture descriptions via a webpage that we designed to collect speech samples from HC (speechbiomarkers.org). The remote participants did not have MMSE scores available. All participant groups did not differ in demographic characteristics, including age, sex ratio, and years of education. The ADNC and FTLN groups had similar demographic characteristics, including age and disease duration. The ADNC group had lower MMSE scores than the FTLN group on average ($P = 0.028$; Table 1), but they were still within the range of intermediate disease severity.

2.2 | Data collection

We digitally recorded the participants' oral descriptions of the Cookie Theft picture from the Boston Diagnostic Aphasia Examination.³² The mean duration of the recordings was 69.8 ± 25.2 seconds. Recordings were transcribed by trained annotators at the Linguistic Data Consortium (LDC) of the University of Pennsylvania using a standard semi-automatic annotation protocol. Only the earliest recording from each participant was included in the study. The study was approved by the institutional review board of the Hospital of the University of Pennsylvania, and we certify that the study was performed in accordance with the ethical standards as in the 1964 Declaration of Helsinki and its later amendments.

RESEARCH IN CONTEXT

- 1. Systematic review:** The authors conducted a literature review using traditional sources, such as PubMed, as well as meeting abstracts and presentations. While speech-based automatic classification tasks of clinical phenotypes (e.g., distinguishing participants with the amnesic Alzheimer's disease (AD) presentation from healthy speakers) have been frequently attempted, automatic classification of underlying pathology has not yet been studied.
- 2. Interpretation:** Our findings demonstrate that speech-based automatic classifiers effectively distinguished underlying pathology (AD vs. frontotemporal lobar degeneration [FTLD]) among individuals with frontotemporal dementia (FTD) phenotypes. We also identified distinct speech patterns in participants with AD neuropathologic change compared to those with FTLN.
- 3. Future directions:** We plan to validate our findings in a larger dataset. This line of work will lay the groundwork for developing a convenient screening tool for clinical trials targeting the neuropathology of patients.

2.3 | Features

2.3.1 | Lexical features

Lexical features were extracted using our automated lexical pipeline. The pipeline automatically tagged the part-of-speech (POS) category of all tokenized words, using spaCy,³³ which is a natural language processing library (NLP) in Python. The pipeline used one of the large language models of spaCy (“en_core_web_lg”) for English. It automatically tallied the count of each POS category and calculated the POS counts per 100 words, controlling for the total number of words in each description.

The pipeline also automatically rated words for concreteness,³⁴ semantic ambiguity,³⁵ word frequency,³⁶ age of acquisition (AoA),³⁷ and word familiarity,³⁸ based on published norms. The concreteness measure indicates how concrete or abstract a word's meaning is on a scale from 1 (most abstract) to 5 (most concrete). Semantic ambiguity measures the potential number of a word's meaning in a given context, where a high score means high ambiguity. Word frequency was calculated from the SUBTLEX-US corpus³⁶ and word frequency per million words was transformed with a \log_{10} scale for simplicity. AoA indicates the average age in years when children acquire a given word, and word familiarity is a z scored scale of the number of people who answered that they knew a given word. The pipeline also measured word length as the number of phonemes using the CMU pronouncing dictionary³⁹ in the NLTK package.⁴⁰ After defining these scores, the pipeline calcu-

TABLE 1 Mean (SD) demographic and clinical characteristics of participants.

	AD (N = 36)	FTLD (N = 60)	HC (N = 83)	P value
Age (years)	64.0 (7.9)	64.7 (7.2)	66.0 (8.3)	0.391
Sex, N (%)				0.889
Female	19 (52.8%)	29 (48.3%)	43 (51.8%)	
Male	17 (47.2%)	31 (51.7%)	40 (48.2%)	
Education (years)	16.0 (2.9)	15.2 (3.1)	16.3 (2.2)	0.051
Disease duration (years)	4.0 (1.8)	3.6 (2.0)	NA	
Clinical phenotype				<0.001
ALS-FTD	0 (0.0%)	2 (3.3%)	0 (0.0%)	
bvFTD	0 (0.0%)	16 (26.7%)	0 (0.0%)	
CBS	0 (0.0%)	3 (5.0%)	0 (0.0%)	
lvPPA	28 (77.8%)	3 (5.0%)	0 (0.0%)	
naPPA	3 (8.3%)	13 (21.7%)	0 (0.0%)	
HC	0 (0.0%)	0 (0.0%)	83 (100.0%)	
PSP	0 (0.0%)	3 (5.0%)	0 (0.0%)	
svPPA	5 (13.9%)	20 (33.3%)	0 (0.0%)	
ADNC				<0.001
High	12 (33.3%)	0 (0.0%)	0 (0.0%)	
Low	0 (0.0%)	10 (16.7%)	0 (0.0%)	
None	0 (0.0%)	12 (20.0%)	45 (54.2%)	
N/A	24 (66.7%)	38 (63.3%)	38 (45.8%)	
MMSE (0-30)	20.6 (6.1)	23.4 (5.6)	29.2 (1.0)	<0.001

Abbreviations: AD, Alzheimer's disease; ADNC, Alzheimer's disease neuropathologic change; ALS-FTD, amyotrophic lateral sclerosis frontotemporal dementia; bvFTD, behavioral variant frontotemporal dementia; CBS, corticobasal syndrome; FTLD, frontotemporal lobar degeneration; HC, healthy control; lvPPA, logopenic variant primary progressive aphasia; MMSE, Mini-Mental State Examination; naPPA, nonfluent/agrammatic variant primary progressive aphasia; PSP, progressive supranuclear palsy; SD, standard deviation; svPPA, semantic variant primary progressive aphasia.

lated the mean scores of these measures across all words, all content words, and all nouns per participant.

Last, the pipeline automatically calculated lexical diversity, that is, how diverse one's word usage was in the picture description, using the moving-average type-token ratio (MATTR).⁴¹ This method calculates a type-token ratio (TTR) for a fixed length of the window, moving the window one word at a time from the beginning to the end of a description, and then it averages all calculated TTR scores. MATTR has been described as one of the most reliable measures for calculating lexical diversity.⁴²

The calculated average lexical scores, POS counts per 100 words, and lexical diversity scores were used for model training. A detailed description of this automatic lexical pipeline has been published previously.^{43,44}

2.3.2 | Acoustic features

Our automatic acoustic pipeline used an in-house Gaussian mixture models-hidden Markov models-based speech activity detector (SAD) developed at LDC to segment audio files into speech segments and silent pauses. The minimum duration of speech and pause segments

was set at 250 and 150 ms, respectively. After segmenting the audio files, we visually reviewed the segments to validate the SAD outputs. Pause segments at the beginning and end of each recording were excluded, as well as pauses after an interviewer's prompt. We also excluded the interviewer's speech segments.

Using the time information from the SAD outputs, the acoustic pipeline automatically calculated durational measurements, including mean speech segment duration, mean silent pause duration, total speech time, total pause time, total time (= total speech + total pause), pause count per minute of speech (= number of pauses/total speech time), speech segment count, pause rate per minute, and percent of speech (= total speech/total time).

Additionally, the pipeline pitch-tracked all speech segments with Praat⁴⁵ and calculated the 10th to 90th percentile pitch estimates in fundamental frequency (f0) for each speech segment with a uniform pitch setting of 75–300 Hz. To normalize physiological differences in voice, we converted pitch values from Hz to semitones (st) using the 10th percentile of each participant as a baseline: $st = 12 \cdot \log_2(f0/\text{baseline } f0)$. We used the converted 90th percentile as a measure of the pitch range of a speaker. A detailed description of this automatic acoustic pipeline has been published previously.¹⁰

2.3.3 | Feature sets

To investigate how well speech features predict patients' underlying pathology, we experimented with five feature sets when training models. The first set included MMSE and demographics (age, sex, education level) only and the performance of this set served as a baseline model. The second feature set included speech features only (90 lexical + 9 acoustic features). The third (99 speech features + 3 demographics) and fourth feature sets (99 speech features + MMSE) added demographics and MMSE, respectively, to speech features to assess whether demographic characteristics or MMSE had additional predictive power in predicting patients' underlying pathology. The last feature set (99 speech features + 3 demographics + MMSE) included speech features, demographics, and MMSE together.

2.3.4 | Feature selection

Because we had a large number of features (maximum $N = 103$) with speech features compared to the number of samples (maximum $N = 179$ participants), we used the elastic net regression as a feature selection method to reduce the number of features. The elastic net regression combines the penalty functions of the LASSO and ridge regression models, L1 and L2 regularizations, respectively, to overcome the limitations of those models. When using the elastic net for feature selection, we varied the L1 regularization ratio from 0.1 to 1 with an 0.1 increment and the value of alpha, which is a constant term that multiplies the penalty terms in the L2 regularization: [0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1]. We did not decrease the feature dimensions to look at individual features that were selected, and we interpret the results based on feature importance values, which is explained in more detail below.

2.4 | Model training

Using the five feature sets (Section 2.3.3), we trained machine learning classifiers for two binary classification tasks: one for participants with ADNC versus those with FTLD ($n = 96$) and the other for HC versus patients with either pathology ($n = 176$). For each binary task ($n = 2$) and feature set combination ($n = 5$), we trained three classifiers: support vector machine (SVM), random forest (RF), and multi-layer perceptron (MLP). For all models, missing values, including MMSE scores of HC participants from speechbiomarkers.org, were imputed using the SimpleImputation function, and all feature values were standardized with the StandardScaler function in scikit-learn⁴⁶ in Python 3.7. Using 10 train–test split folds, all models were trained on the nine folds (the train set) and tested on the 10th fold (the test set). During this process, features were selected using the elastic net regression for each train set. This process was repeated 10 times to test the model performance in all the samples we had, and we report the best averaged performance of the models over 10 folds after hyperparameter tuning and feature selection.

2.5 | Feature interpretation

From the best performing model that was trained with only speech features for each binary task, we counted the frequency of selected features in 10 folds (i.e., how many times a feature was selected). Features that are selected at least 9 times out of 10 folds were reported in the results. We also computed Shapley additive explanations (SHAP) values⁴⁷ of the most frequently selected features to explain the importance of the selected features in the best performing speech models. SHAP values are frequently used in the literature to provide an explanation of a model's prediction by computing the relative impact of each feature on the prediction. We further explored group differences in the most frequently selected features. First, we decided if the data were normally distributed and the variances of the two groups were homogeneous. Depending on the results, we performed *t* test or Wilcoxon signed rank tests to compare the groups and reported significant results.

2.6 | Imaging analysis

High-resolution T1-weighted magnetic resonance imaging (MRI) was available for a subset of participants within 10 months of the speech sample (median 1 month, mean 2.04 months, range 0–10 months). All images were acquired on a 3T Siemens scanner. Most images (ADNC = 20, FTLD = 28, HC = 16) were acquired using an axially acquired magnetization prepared rapid acquisition gradient echo (MP-RAGE) protocol with repetition time = 1620 ms; echo time = 3.87 ms; flip angle = 15°; matrix = 192 × 256, 160 slices, and resolution = 0.9766 mm × 0.9766 mm × 1.0 mm. Additional participants (ADNC = 1, FTLD = 9) were imaged using a sagittal MP-RAGE sequence with repetition time = 2300 ms, echo time = 2.91 ms, flip angle = 9°, matrix = 240 × 256, 176 slices, and resolution = 1.055 mm × 1.055 mm × 1.2 mm. A sagittal multiplanar reconstruction multi-echo (4) virtual navigator sequence with repetition time = 2400 ms, echo time = 1.96 ms, flip angle = 8°, matrix = 320 × 320, 224 slices, and resolution = 0.8 mm × 0.8 mm × 0.8 mm was used to acquire the remaining images (ADNC = 2, FTLD = 10, HC = 5). Images were processed using the antsCorticalThickness pipeline.⁴⁸ Cortical gray matter (GM) volumes from the Yeo 17 networks⁴⁹ were summarized per hemisphere to create 34 network regions which were each normalized for age and sex to account for associated population differences. Linear models were used to compare GM volumes between participants with ADNC and HC and between those with FTLD and HC to establish atrophy patterns for each patient group. ADNC and FTLD were also contrasted to identify differences in atrophy localization between ADNC and FTLD directly. Within each patient group, linear models were then used to determine relationships between speech features and network volumes in atrophied regions for the respective patient group to HC comparisons. The speech measures included in this regression analysis were those most frequently selected in the best performing speech model where the patient group means were further than one standard deviation from the control means. All linear models

TABLE 2 Classification results of ADNC versus FTLD.

Value	Model	Accuracy	AUC	Sensitivity	Specificity
Demo + MMSE	MLP	0.66 (0.13)	0.65 (0.07)	0.38 (0.26)	0.81 (0.18)
	RF	0.61 (0.18)	0.60 (0.12)	0.24 (0.23)	0.82 (0.19)
	SVM	0.59 (0.17)	0.42 (0.20)	0.10 (0.14)	0.89 (0.15)
Speech	MLP	0.83 (0.10)	0.88 (0.03)	0.81 (0.25)	0.87 (0.14)
	RF	0.82 (0.07)	0.84 (0.05)	0.64 (0.16)	0.96 (0.07)
	SVM	0.77 (0.13)	0.85 (0.08)	0.54 (0.27)	0.94 (0.08)
Speech + demo	MLP	0.81 (0.13)	0.85 (0.05)	0.76 (0.19)	0.86 (0.13)
	RF	0.80 (0.15)	0.86 (0.05)	0.64 (0.27)	0.94 (0.11)
	SVM	0.78 (0.12)	0.83 (0.05)	0.68 (0.27)	0.87 (0.13)
Speech + MMSE	MLP	0.83 (0.11)	0.83 (0.04)	0.74 (0.28)	0.91 (0.11)
	RF	0.84 (0.09)	0.85 (0.04)	0.68 (0.19)	0.96 (0.09)
	SVM	0.81 (0.11)	0.82 (0.03)	0.71 (0.18)	0.89 (0.13)
Speech + demo + MMSE	MLP	0.81 (0.13)	0.84 (0.05)	0.64 (0.20)	0.93 (0.12)
	RF	0.82 (0.09)	0.83 (0.03)	0.62 (0.19)	0.95 (0.07)
	SVM	0.80 (0.10)	0.82 (0.04)	0.67 (0.28)	0.89 (0.13)

Notes: The values represent the mean (SD) over 10 folds. Sensitivity in all models refers to the ability to distinguish the AD group, whereas specificity refers to the ability to distinguish the FTLD group. Results of the best performing models based on the accuracy and AUC are highlighted in bold.

Abbreviations: ADNC, Alzheimer's disease neuropathologic change; AUC, area under the curve; FTLD, frontotemporal lobar degeneration; MLP, multi-layer perceptron classifier; MMSE, Mini-Mental State Examination; RF, random forest classifier; SD, standard deviation; SVM, support vector machine classifier.

included acquisition pulse sequence as a covariate of no interest, and results were considered significant at $P < 0.05$ after false discovery rate (FDR) correction for multiple comparisons.

3 | RESULTS

3.1 | ADNC versus FTLD

3.1.1 | Classification results

Among the models trained with speech features, the MLP model showed the highest accuracy (83%) with an area under the curve (AUC) of 0.88 in distinguishing patients with ADNC from those with FTLD (Table 2). The model correctly identified 52 FTLD patients out of 60 and 28 patients with ADNC out of 36. The eight participants with FTLD that the model missed included five with svPPA, one with bvFTD, one with amyotrophic lateral sclerosis FTLD, and one with naPPA. The eight participants with AD that the classifier incorrectly predicted included seven patients with lvPPA and one with svPPA. This model's performance was comparable to the RF model that was trained with speech features and MMSE scores (accuracy = 84%, AUC = 0.85). All models generally showed good specificity scores (the ability to distinguish FTLD pathology in this task > 0.85) except the baseline models, which were trained without speech features. In particular, RF models trained with speech features only or speech + MMSE showed the highest specificity (0.96). Adding MMSE scores increased the accuracy of the RF and SVM models (RF = 2%, SVM = 4%), whereas adding demographic features decreased the accuracy of the MLP and RF models

(−2% for both models). Adding MMSE and demographics together to the models showed mixed effects on the models (MLP = −2%, RF = 0%, SVM = 3%).

3.1.2 | Selected features from the best performing speech model

The average number of features selected in each fold in the best performing speech model was 37.9 (± 2.13), with a range from 34 to 41. Features selected at least 9 times in the 10 folds included verbs in base form per 100 words ($n = 10$), verbs (in 3rd sg.) per 100 words ($n = 10$), word length of nouns ($n = 10$), percent of speech ($n = 9$), ambiguity of content words ($n = 10$), lexical diversity ($n = 10$), pause rate per minute ($n = 10$), frequency of nouns ($n = 10$), total number of words ($n = 10$), pitch range ($n = 10$), familiarity of content words ($n = 9$), total number of syllables ($n = 10$), concreteness of nouns ($n = 9$), and comparative adjective counts per 100 words ($n = 9$). The feature importance of these features in absolute SHAP values is displayed in Figure 1A.

Among the 14 features that were most frequently selected, seven features showed significant group differences: base verb counts per 100 words, third singular verbs per 100 words, comparative adjectives per 100 words, averaged ambiguity of content words, lexical diversity, total number of words, and total number of syllables (Figure 1B). Patients with ADNC produced more verbs in base form and comparative adjectives per 100 words compared to FTLD patients (base verbs: $t = 2.79$, $P = 0.007$, comparative adjectives: $W = 1214.5$, $P = 0.016$), whereas they produced fewer verbs in the third person singular form ($W = 636.5$, $P = 0.001$). Content words that patients with ADNC pro-

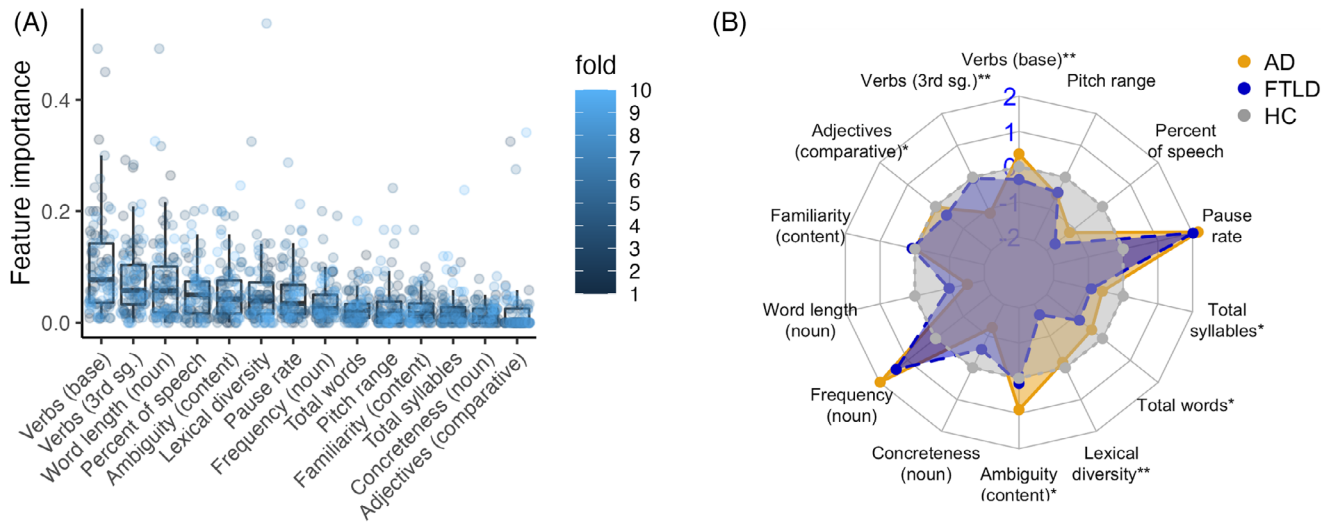


FIGURE 1 Selected features in the best performing speech-based classifier for the distinction between ADNC and FTLD. A, Feature importance values (Shapley additive explanations) of the 14 most frequently selected speech features per fold in the distinction of ADNC versus FTLD. Each data point represents the feature importance value for one patient in the fold. B, Group comparisons of the 14 most frequently selected speech features. The scales of the features were z scored based on HC's mean and standard deviation for better visualization. HC's values are plotted in gray for reference. Seven features showed significant group comparisons between ADNC and FTLD: * $P < 0.05$; ** $P < 0.01$. ADNC, Alzheimer's disease neuropathologic change; FTLD, frontotemporal lobar degeneration; HC, healthy controls

duced were more ambiguous than those produced by patients with FTLD ($t = 2.48$, $P = 0.015$). Patients with FTLD used more limited vocabularies (lower lexical diversity) compared to those with ADNC ($t = 3.6$, $P = 0.001$), and they produced fewer words and syllables when describing the same picture compared to those with ADNC (total words: $t = 2.56$, $P = 0.013$, total syllables: $t = 2.09$, $P = 0.04$). The other features, such as word frequency or length, seemed to help the classification of ADNC versus FTLD, but did not significantly differ by group.

3.2 | Patients versus HC

3.2.1 | Classification results

The MLP model trained only with speech features showed the highest accuracy (93%) with 0.93 AUC, correctly identifying 89 patients out of 96 (= 36 AD + 60 FTLD) and 77 HCs out of 83 (Table 3). The seven patients that the model incorrectly predicted included three participants with lvPPA, two with bvFTD, and two with progressive supranuclear palsy. The models generally showed high performance (accuracy > 89%) when speech features were included, compared to the baseline models that were trained with demographics and MMSE scores only. The sensitivity (the ability to distinguish patients in this task) improved when MMSE scores were added to speech features (speech only models: 0.89–0.93, speech + MMSE: 0.9–0.95). The specificity (the ability to identify HC in this task) of the MLP speech model (0.9) was the second highest, preceded by that of the SVM speech + MMSE model (0.91). The RF model trained with speech features and MMSE scores showed a slightly higher AUC value (0.94) with

lower accuracy (0.92) compared to the MLP model trained with speech features. Adding demographics or MMSE scores did not dramatically change the performance of the models in terms of accuracy.

3.2.2 | Selected features of the best performing speech model

The average number of features selected in each fold for the classification of HC versus patients was 31.7 ± 3.23 , ranging from 25 to 35. Features that were selected at least 9 times out of 10 folds included the number of conjunctions ($n = 10$), determiners ($n = 10$), and possessives per 100 words ($n = 9$); averaged word frequency of all words ($n = 10$) and of nouns ($n = 10$); averaged word length of all words ($n = 10$) and of content words ($n = 10$); lexical diversity ($n = 10$); total number of words ($n = 9$); total number of unique adjectives ($n = 10$); unique content words ($n = 10$); unique words ($n = 10$); pause rate per minute ($n = 9$); and total speech duration ($n = 10$). The feature importance values of these features are presented in Figure 2A.

Patients produced fewer unique adjectives ($t = 5.41$, $P < 0.001$) and unique content words ($W = 2757$, $P = 0.008$) compared to HC, while they produced more conjunctions than HC ($W = 1656.5$, $P = 0.026$; Figure 2B). Words that patients produced were shorter (all: $W = 3558.5$, $P < 0.001$; content words: $W = 3482$, $P < 0.001$) and more frequent (all: $t = -5.61$, $P < 0.001$; nouns: $W = 1164$, $P < 0.001$) than those produced by HC. Patients spent less time describing the same picture ($t = 3.73$, $P < 0.001$), with more frequent pauses ($W = 633$, $P < 0.001$) compared to HC. They also used less diverse vocabulary ($W = 2843$, $P = 0.001$) with fewer words in total ($t = 4.24$, $P < 0.001$) than HC.

TABLE 3 Classification results of patients versus HC.

Value	Model	Accuracy	AUC	Sensitivity	Specificity
Demo + MMSE	MLP	0.72 (0.23)	0.86 (0.05)	0.82 (0.14)	0.62 (0.25)
	RF	0.84 (0.16)	0.92 (0.03)	0.82 (0.13)	0.79 (0.29)
	SVC	0.54 (0.27)	0.75 (0.09)	0.57 (0.21)	0.66 (0.32)
Speech	MLP	0.93 (0.06)	0.93 (0.04)	0.93 (0.08)	0.90 (0.12)
	RF	0.91 (0.09)	0.89 (0.05)	0.92 (0.08)	0.86 (0.21)
	SVC	0.89 (0.09)	0.90 (0.06)	0.89 (0.07)	0.84 (0.17)
Speech + demo	MLP	0.89 (0.11)	0.92 (0.03)	0.90 (0.08)	0.85 (0.22)
	RF	0.91 (0.09)	0.90 (0.05)	0.83 (0.32)	0.87 (0.22)
	SVC	0.91 (0.11)	0.83 (0.09)	0.84 (0.32)	0.84 (0.26)
Speech + MMSE	MLP	0.92 (0.08)	0.94 (0.03)	0.93 (0.07)	0.89 (0.15)
	RF	0.91 (0.09)	0.88 (0.07)	0.95 (0.06)	0.83 (0.21)
	SVC	0.91 (0.06)	0.89 (0.06)	0.90 (0.07)	0.91 (0.10)
Speech + demo + MMSE	MLP	0.92 (0.08)	0.91 (0.05)	0.94 (0.08)	0.87 (0.16)
	RF	0.92 (0.10)	0.91 (0.05)	0.74 (0.40)	0.88 (0.18)
	SVC	0.91 (0.08)	0.86 (0.07)	0.94 (0.06)	0.86 (0.21)

Notes: The values are mean (SD) over 10 folds. Sensitivity in all models refers to the ability to distinguish the patients. Results of the best performing models based on the accuracy and AUC are highlighted in bold.

Abbreviations: AUC, area under the curve; HC, healthy control; MLP, multi-layer perceptron classifier; MMSE, Mini-Mental State Examination; RF, random forest classifier; SD, standard deviation; SVM, support vector classifier.

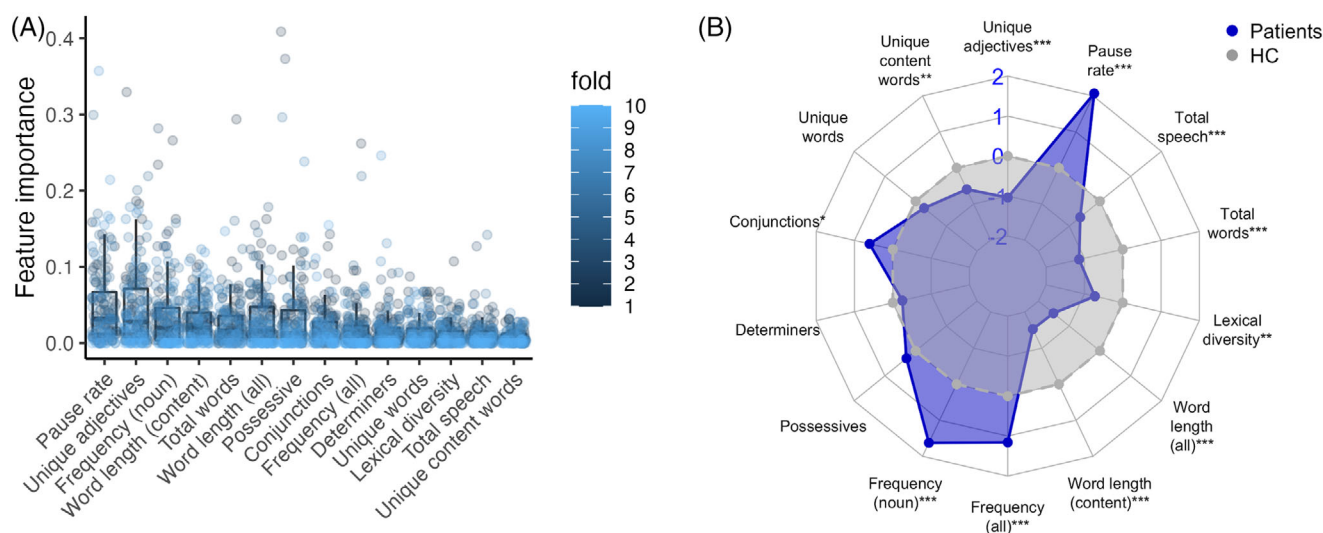


FIGURE 2 Selected features in the best-performing speech-based classifiers for the distinction between patients and HCs. A, Feature importance values (|Shapley additive explanations|) of the most frequently selected features per fold in the distinction of patients versus HC. Each data point represents the feature importance value for one patient in the fold. B, Group comparisons of the most frequently selected features. The scales of the features were z scored based on the HC's mean and standard deviation for better visualization. HC's values are plotted in gray, and patients are in blue. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$; HC, healthy controls

3.3 | Imaging

3.3.1 | GM atrophy

As shown in Figure 3 and Table 4, participants with ADNC demonstrated lower GM volumes than HC (Figure 3A) in the left hemisphere

in the default mode network (DMN) parts A and B, as well as the left temporal parietal, visual A, dorsal attention A and B, limbic A, salience A, and control A and B networks, which cover much of the frontal, temporal, and parietal cortex, as well as some occipital lobe. Participants with FTLN exhibited lower GM volume relative to HC in bilateral limbic A and B, temporal parietal, and salience A networks, in addition to left

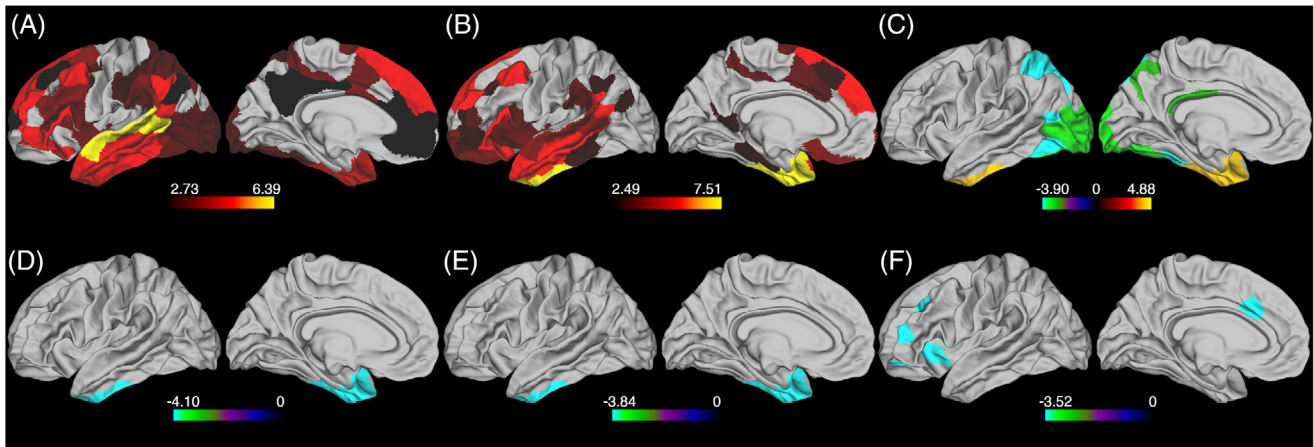


FIGURE 3 Comparisons of cortical gray matter (GM) volumes between (A) participants with Alzheimer's disease neuropathologic change (ADNC) and healthy controls (HC). B, Patients with frontotemporal lobar degeneration (FTLD) versus HC. C, Patients with ADNC compared to those with FTLD. D, Significant association of GM volume to noun frequency in ADNC. E, Significant association of GM volume to noun frequency in FTLD. F, Significant association of GM volume to pause rate in FTLD. All scale bars represent *t* statistics

networks salience B, control B and DMN B and C, and right DMN B network regions, which cover most of the frontal and temporal cortices, as well as some parietal cortex (Figure 3B). Participants with FTLD exhibited lower GM volumes relative to those with ADNC in bilateral limbic A, as well as right limbic B and salience A, which consist of bilateral anterior inferior temporal lobe, along with right orbitofrontal, insular, medial frontal, and precentral cortex, and supramarginal gyrus (Figure 3C). Participants with ADNC demonstrated atrophy relative to FTLD in left visual A, dorsal attention A, and control C networks, which consist of much of the lateral occipital lobe, as well as some posterior parietal lobule and posterior cingulate cortex (Figure 3C).

3.3.2 | Imaging regressions

As shown in Figure 3 and Table 4, for participants with ADNC (Figure 3D) and FTLD (Figure 3E), increased frequency of nouns was significantly associated with atrophy in left limbic regions, which largely consist of anterior inferior temporal cortex. Furthermore, in participants with FTLD, a higher pause rate was associated with atrophy in the left salience network, which includes inferior frontal, orbitofrontal, and middle frontal cortex, along with cingulate cortex and insula (Figure 3F).

4 | DISCUSSION

We used automated lexical and acoustic pipelines to extract speech features from naturalistic speech samples and trained machine learning classifiers to distinguish participants with FTD phenotypes by their likely underlying pathology (ADNC vs. FTLD). Our classification systems trained with speech features performed well, with 83% accuracy (AUC = 0.88) for identifying patients with likely underlying ADNC versus FTLD, and 93% accuracy (AUC = 0.93) for distinguishing all

participants with FTD phenotypes from healthy individuals. These results suggest that speech features automatically extracted from 1 minute picture descriptions can be used to help identify participants who present with an FTD phenotype but are likely to have underlying ADNC. We applied an interpretable AI approach to show important features and tested group differences in those features. For additional neuroanatomical validity we associated some of the most distinct speech features with GM atrophy in functional neural networks. We discuss these findings below.

Although all patients in our study presented with an FTD phenotype, our models indicated linguistic patterns that were distinct for the likely underlying pathology, differentiating underlying ADNC from FTLD. Participants with likely ADNC produced more verbs in base form (e.g., "take") but fewer verbs in the third singular present tense (e.g., "takes") per 100 words. We and others have observed a similar finding in previous studies,^{11,50,51} in which patients with amnesic or non-amnesic AD presentations produced more inflectional errors (e.g., "She run") instead of "She runs") and fewer tense-inflected verbs than HC. We have previously shown that individuals with bvFTD or svPPA had normal counts of tense-inflected verbs, and they were the majority of our current FTLD group (>50%).⁴⁴ Other reports⁵¹ have also found that participants with the amnesic AD presentation produced more frequent inflectional errors than HC, whereas individuals with svPPA did not differ from HC. It seems that the number of tense inflections is a useful measure in distinguishing ADNC from FTLD pathology. The exact nature of this phenomenon should be investigated further, preferably with detailed neuroanatomical and neuropathological correlates.

We found high ambiguity scores in participants with ADNC but not in those with FTLD. Words with high ambiguity scores (e.g., "something") are easier to access compared to less ambiguous, more concrete words (e.g., "stool"). In a previous study, we found that high ambiguity scores were a distinct feature of patients with a semantic impairment rather than grammatical or behavioral impairment.⁴⁴ svPPA patients may compensate for impaired grammatical capacity by selecting more

TABLE 4 Imaging results from comparing GM volumes from each of the Yeo 17 networks, split by hemisphere, between patients with ADNC and HC, patients with FTLN and HC, and between patients with ADNC and FTLN.

Imaging analysis	Region	t value	P value
ADNC atrophy	Left default B	4.78	<0.001
	Left default A	2.73	0.032
	Left temporal parietal	6.39	<0.001
	Left visual central (visual A)	3.23	0.011
	Left dorsal attention B	3.10	0.013
	Left dorsal attention A	3.29	0.01
	Left limbic A	4.29	<0.001
	Left salience/ventral attention A	3.49	0.007
	Left control A	3.71	0.004
	Left control B	4.84	<0.001
FTLN atrophy	Left limbic A	7.51	<0.001
	Left limbic B	3.77	0.003
	Right limbic A	6.06	<0.001
	Right limbic B	2.49	0.04
	Left temporal parietal	3.99	0.002
	Right temporal parietal	2.56	0.037
	Left salience/ventral attention A	3.40	0.007
	Left salience/ventral attention B	3.14	0.011
	Right salience/ventral attention A	3.13	0.011
	Left control B	2.80	0.023
	Left default B	5.55	<0.001
	Left default C	2.76	0.023
Right default B	3.04	0.013	
ADNC to FTLN atrophy	Left limbic A	4.33	<0.001
	Right limbic A	4.88	<0.001
	Right limbic B	2.91	0.033
	Right salience/ventral attention A	2.68	0.045
	Left visual central (visual A)	-3.22	0.017
	Left dorsal attention A	-3.90	<0.001
	Left control C	-2.80	0.038
ADNC frequency (noun)	Left limbic A	-4.10	0.042
FTLN frequency (noun)	Left limbic A	-3.84	0.036
FTLN pause rate	Left salience/ventral attention B	-3.52	0.047

Notes: Regressions were also run, relating speech features to the GM volumes from each of the atrophied networks for each patient group, respectively. All results considered significant at $P < 0.05$ after FDR correction for multiple comparisons.

Abbreviations: ADNC, Alzheimer's disease neuropathologic change; FDR, false discovery rate; FTLN, frontotemporal lobar degeneration; GM, gray matter; HC, healthy control.

precise words with a lower ambiguity score.^{52,53} Many people with amnesic AD presentation also have difficulty accessing words, which may be related to impaired working memory or to the spread of disease from medial to more lateral parts of the left temporal lobe. Indeed, in another study, we found impaired naming in people presenting with synucleinopathies who had concomitant ADNC and this was directly associated with a high p-tau level in their CSF.⁵⁴ Thus, ambiguity scores may not be sensitive enough to detect individuals with likely underly-

ing FTLN versus ADNC when various FTD phenotypes are included. In our highly phenotypically diverse patient cohort, word ambiguity was sensitive to likely underlying ADNC. Future studies of digital speech assessments in neurodegenerative conditions should consider different clinical scenarios when planning their analyses.

Noun frequency was high in our AD and FTLN groups compared to healthy speakers, and we found partial associations between this feature and decreased GM volumes in the left limbic A network. Both

ADNC and FTLD groups demonstrated atrophy in this network, which encompasses much of the left anterior temporal lobe (ATL), and previous studies have shown that atrophy in this region is associated with increased word frequency in patients with FTD.^{35,44,55,56} Atrophy in these areas in both groups seems to be generally associated with increased word frequency. Words with higher frequency (e.g., “chair” vs. “stepstool”) are typically more accessible as they are learned earlier and accessed more frequently in the language acquisition process.³⁶ This is a frequently seen hallmark in aphasia patients with injury to temporal regions involved in language processing including anterior parts of the left temporal lobe.

Pause rate was increased in both groups, yet it was significantly associated with GM volume loss in parts of the left salience network only in patients with FTLD. Increased pausing while speaking is often related to word finding difficulty, but this could result from injury to different cognitive processes: in people with ADNC this may result from impairment of the audio-verbal loop,⁵⁷ while in FTLD this could arise from impaired social-executive interaction in the process of lexical selection or agrammatism. The fronto-insular cortex, part of the salience network, has been structurally and functionally linked to FTD syndromes, particularly bvFTD,⁵⁸ and this is suggested by our MRI analysis. In another study, we found high pause rates in all PPA variants, most profoundly in naPPA,⁵⁹ and in another classification study this feature differentiated subtypes of PPA.⁶⁰ Our MRI analysis supports the association of frequent pausing in FTLD to the salience network. However, we were unable to localize this speech feature in our ADNC group, and this could be related to our chosen methodology and network scheme, designed to capture structural involvement by network. Future studies can consider alternative approaches using structural or functional neuroimaging to localize increased pausing in ADNC.

It has been observed that adjective counts and lexical diversity of patients with cognitive impairment decreased compared to HC.^{11,44,61} Additionally, patients with various types of dementia tended to produce shorter and more frequent words than HC.^{11,44,62,63} These features were also selected in our analysis. These observations along with the high performance of our speech model provide accumulating evidence that naturalistic speech is informative in screening individuals with different presentations of dementia and diverse underlying pathologies. Our study emphasized the power of speech analysis in *in vivo* screening for the underlying pathology in a phenotypically diverse group of patients.

Our participants did not differ in any major demographic characteristics, and this could explain the insignificant effect of adding demographics to our models. However, this finding may not be generalizable to other cohorts, when considering individuals with the typical amnesic AD presentation, as they tend to be older than FTLD patients. MMSE scores, however, had a limited effect on our model's performance, dissociating the underlying pathology of patients only when combined with speech features. Our participants with ADNC had significantly lower MMSE scores compared to the FTLD group. The lower MMSE scores in ADNC may have helped distinguish patients' underlying pathology in our particular cohort, but this finding might not be generalizable to all patient populations.

While this study demonstrated high performance in automatic classification of likely underlying pathology in people with FTD phenotypes, it has limitations. First, the models included several similar features. Elastic net regression is robust to collinearity, but other feature dimensionality reduction techniques, such as principal component analysis, may also be considered. We did not try feature reduction here, because reduced features are sometimes difficult to interpret. Second, the sample size of the AD group was smaller than that of the FTLD group, and our study was not validated in an independent sample of patients. Future studies will need to verify our findings with more patients with biological evidence of the pathological grouping. Additionally, we could not examine the effect of mild depression and any depression medication in our cohort. Mild depression is a commonly co-occurring psychiatric symptom with neurodegeneration, but we only had limited information about the medication usage in our participants. Future studies will need to consider the effect of any psychiatric medication when studying these populations. For our structural imaging analysis, we used regions of interest based upon control resting state functional MRI, which may be suboptimal for investigating associations with language performance, though we found significant associations between some of these regions and the language features in our study. Last, we extracted lexical features from transcripts generated by human annotators. Current automatic speech recognition systems have a high word error rate when processing patients' speech, but we expect future work will enable complete automation of transcription pipelines for scalable analyses of patient speech with higher accuracy.

Interesting directions for future studies include the comparison of acoustic and lexical features, as well as the investigation of other clinical assessments in the baseline models. Considering that acoustic features were less frequently selected by the models than lexical features, it might be the case that classifiers trained with lexical features outperform those trained with acoustic features. Also, we only included MMSE scores in the baseline models due to limited availability of other clinical assessments, but standard language tests may improve the performance of baseline models. Future studies will need to determine whether speech-based systems consistently outperform baseline models trained with various clinical assessment scores.

To conclude, natural speech of people with FTD syndromes briefly describing a picture can be used to automatically classify them by their underlying neuropathology as AD or FTLD. Such *in vivo* automatic classifiers derived from low-cost, non-invasive digital speech assessments could be extremely useful in future multi-center trials.

ACKNOWLEDGMENTS

In memory of Murray Grossman. One of the co-authors, Murray Grossman, MD, FAAN, passed away on April 4, 2023. This study was funded by grants from the National Institutes of Health (P01-AG066597, K99-AG073510-01, P30-AG072979), Alzheimer's Association (AARF-21-851126), and the Department of Defense (W81XWH-20-1-0531, AL220035).

CONFLICT OF INTEREST STATEMENT

Author disclosures are available in the [supporting information](#).

CONSENT STATEMENT

All human subjects provided informed consent.

ORCID

Sunghye Cho  <https://orcid.org/0000-0003-1569-7608>

REFERENCES

- Alzheimer's Association. 2020 Alzheimer's disease facts and figures. 2020. doi:10.1002/alz.12068
- Nicholas M, Obler L, Albert M, Goodglass H. Lexical retrieval in healthy aging. *Cortex*. 1985;21(4):595-606. doi:10.1016/S0010-9452(58)80007-6
- Kavé G, Dassa A. Severity of Alzheimer's disease and language features in picture descriptions. *Aphasiology*. 2018;32(1):27-40. doi:10.1080/02687038.2017.1303441
- Ahmed S, Haigh AMF, de Jager CA, Garrard P. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*. 2013;136:3727-3737. doi:10.1093/brain/awt269
- Boschi V, Catricalà E, Consonni M, Chesi C, Moro A, Cappa SF. Connected speech in neurodegenerative language disorders: a review. *Front Psychol*. 2017;8:269. doi:10.3389/fpsyg.2017.00269
- Snowdon DA. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. Findings from the Nun Study. *JAMA*. 1996;275(7):528-532. doi:10.1001/jama.275.7.528
- Kemper S, Thompson M, Marquis J. Longitudinal change in language production: effects of aging and dementia on grammatical complexity and propositional content. *Psychol Aging*. 2001;16(4):600-614. doi:10.1037/0882-7974.16.4.600
- Riley KP, Snowdon DA, Desrosiers MF, Markesbery WR. Early life linguistic ability, late life cognitive function, and neuropathology: findings from the Nun Study. *Neurobiol Aging*. 2005;26(3):341-347. doi:10.1016/j.neurobiolaging.2004.06.019
- Gorno-Tempini ML, Hillis AE, Weintraub S, et al. Classification of primary progressive aphasia and its variants. *Neurology*. 2011;76(11):1006-1014. doi:10.1212/WNL.Ob013e31821103e6
- Nevler N, Ash S, Jester C, Irwin DJ, Liberman M, Grossman M. Automatic measurement of prosody in behavioral variant FTD. *Neurology*. 2017;89:1-8.
- Cho S, Cousins KAQ, Shellikeri S, et al. Lexical and acoustic speech features relating to Alzheimer disease pathology. *Neurology*. 2022;99(4):E313-E322. doi:10.1212/WNL.000000000000200581
- Giannini LAA, Irwin DJ, Mcmillan CT, et al. Clinical marker for Alzheimer disease pathology in logopenic primary progressive aphasia. *Neurology*. 2017;88(24):2276-2284. doi:10.1212/WNL.0000000000004034
- Bergeron D, Gorno-Tempini ML, Rabinovici GD, et al. Prevalence of amyloid- β pathology in distinct variants of primary progressive aphasia. *Ann Neurol*. 2018;84(5):729-740. doi:10.1002/ana.25333
- Fraser K, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. *J Alzheimer's Dis*. 2016;49(2):407-422. doi:10.3233/JAD-150520
- Jarrold W, Peintner B, Wilkins D, et al. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 2014:27-37. doi:10.3115/v1/w14-3204
- Amini S, Hao B, Zhang L, et al. Automated detection of mild cognitive impairment and dementia from voice recordings: a natural language processing approach. *Alzheimers Dement*. 2022;19(3):946-955. doi:10.1002/alz.12721
- 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. doi:10.1080/026870300401603
- Rentoumi V, Paliouras G, Danasi E, et al. Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: a computational linguistics analysis. In: 8th IEEE International Conference on Cognitive Infocommunications, CogInfoCom. 2017;33-38. doi:10.1109/CogInfoCom.2017.8268212
- Guinn C, Habash A. Language analysis of speakers with dementia of the Alzheimer's type. AAAI Fall Symposium: Artificial Intelligence for Gerontechnology. 2012; FS-12-01:8-13.
- Meilán JGG, Martínez-Sánchez F, Carro J, López DE, Millian-Morell L, Arana JM. Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia? *Dement Geriatr Cogn Disord*. 2014;37(5-6):327-334. doi:10.1159/000356726
- Orimaye SO, Wong JS-M, Golden KJ. Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. Workshop on computational linguistics and clinical psychology: from linguistic signal to. Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. 2014:78-87. <http://www.alz.org/dementia/>
- 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. *Alzheimers Dement*. 2018;10:260-268. doi:10.1016/j.jad.2018.02.004
- Eyigöz E, Courson M, Sedeño L, et al. From discourse to pathology: automatic identification of Parkinson's disease patients via morphological measures across three languages. *Cortex*. 2020;132:191-205. doi:10.1016/j.cortex.2020.08.020
- Yuan J, Bian Y, Cai X, Huang J, Ye Z, Church K. Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 2020:2162-2166. doi:10.21437/Interspeech.2020-2516
- Yuan J, Cai X, Church K. Pause-encoded language models for recognition of Alzheimer's disease and emotion. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE; 2021:7293-7297. doi:10.1109/ICASSP39728.2021.9413548
- McKhann GM, Knopman DS, Chertkow H, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*. 2011;7(3):263-269. doi:10.1016/j.jalz.2011.03.005
- Rascovsky K, Hodges JR, Knopman D, et al. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain*. 2011;134(9):2456-2477. doi:10.1093/brain/awr179
- Irwin D, McMillan CT, Toledo JB, et al. Comparison of cerebrospinal fluid levels of tau and A β 1-42 in Alzheimer disease and frontotemporal degeneration using 2 analytical platforms. *Arch Neurol*. 2012;69(8):1018-1025. doi:10.1001/archneurol.2012.26
- Lleó A, Irwin DJ, Illán-Gala I, et al. A 2-step cerebrospinal algorithm for the selection of frontotemporal lobar degeneration subtypes. *JAMA Neurol*. 2018;75(6):738-745. doi:10.1001/jamaneurol.2018.0118
- Montine TJ, Phelps CH, Beach TG, et al. National institute on aging-Alzheimer's association guidelines for the neuropathologic assessment of Alzheimer's disease: a practical approach. *Acta Neuropathol*. 2012;123(1):1-11. doi:10.1007/s00401-011-0910-3
- Toledo JB, Van Deerlin VM, Lee EB, et al. A platform for discovery: the University of Pennsylvania integrated neurodegenerative disease biobank. *Alzheimers Dement*. 2014;10(4):477-484. e1. doi:10.1016/j.jalz.2013.06.003
- Goodglass H, Kaplan E, Weintraub S. *Boston Diagnostic Aphasia Examination*. Lea & Febiger; 1983.
- Honnibal M, Johnson M. An improved non-monotonic transition system for dependency parsing. In: *EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*. 2015:1373-1378. doi:10.18653/v1/d15-1162

34. Brysbaert M, Warriner AB, Kuperman V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behav Res Methods*. 2014;46(3):904-911. doi:10.3758/s13428-013-0403-5
35. Hoffman P, Lambon Ralph MA, Rogers TT. Semantic diversity: a measure of semantic ambiguity based on variability in the contextual usage of words. *Behav Res Methods*. 2013;45(3):718-730. doi:10.3758/s13428-012-0278-x
36. Brysbaert M, New B. Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav Res Methods*. 2009;41(4):977-990. doi:10.3758/BRM.41.4.977
37. Kuperman V, Stadthagen-Gonzalez H, Brysbaert M. Age-of-acquisition ratings for 30,000 English words. *Behav Res Methods*. 2012;44(4):978-990. doi:10.3758/s13428-012-0210-4
38. Brysbaert M, Mandera P, Keuleers E. Word prevalence norms for 62,000 English lemmas. *Behav Res Methods*. 2018;51:467-479.
39. Carnegie Mellon Speech Group. The Carnegie Mellon University Pronouncing Dictionary. 2014. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>
40. Loper E, Bird S. NLTK: the Natural Language Toolkit. In: *ETMTNLP '02: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. 2002:63-70. doi:10.3115/1118108.1118117
41. Covington MA, McFall JD. Cutting the Gordian knot: the moving average type-token ratio (MATTR). *J Quant Linguist*. 2010;17(2):94-100.
42. Fergadiotis G, Wright HH, Green SB. Psychometric evaluation of lexical diversity indices: assessing length effects. *J Speech Lang Hear Res*. 2015;58:840-852. doi:10.1044/2015
43. Cho S, Nevler N, Shellikeri S, et al. Lexical and acoustic characteristics of young and older healthy adults. *J Speech Lang Hear Res*. 2021;64(2):302-314. doi:10.1044/2020_JSLHR-19-00384
44. Cho S, Nevler N, Ash S, et al. Automated analysis of lexical features in frontotemporal degeneration. *Cortex*. 2021;137:215-231. doi:10.1016/j.cortex.2021.01.012
45. Boersma P, Weenink D. Praat: doing phonetics by computer. 2019. <http://www.praat.org>
46. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830. doi:arXiv/1201.0490
47. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. 2017. <http://arxiv.org/abs/1705.07874>
48. Tustison NJ, Cook PA, Klein A, et al. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *Neuroimage*. 2014;99:166-179. doi:10.1016/j.neuroimage.2014.05.044
49. Yeo BTT, Krienen FM, Sepulcre J, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J Neurophysiol*. 2011;106(3):1125-1165. doi:10.1152/jn.00338.2011
50. Altmann LJP, Kempler D, Andersen ES. Speech errors in Alzheimer's disease: reevaluating morphosyntactic preservation. *J Speech Lang Hear Res*. 2001;44(5):1069-1082. doi:10.1044/1092-4388(2001/085)
51. Sajjadi SA, Patterson K, Tomek M, Nestor PJ. Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. *Aphasiology*. 2012;26(6):847-866. doi:10.1080/02687038.2012.654933
52. Whitwell JL, Jones DT, Duffy JR, et al. Working memory and language network dysfunctions in logopenic aphasia: a task-free fMRI comparison with Alzheimer's dementia. *Neurobiol Aging*. 2015;36(3):1245-1252. doi:10.1016/j.neurobiolaging.2014.12.013
53. Rezaei N, Mahowald K, Ryskin R, Dickerson B, Gibson E. A syntax-lexicon trade-off in language production. *Proc Natl Acad Sci*. 2022;119(25):1-11. doi:10.1073/pnas.2120203119
54. Howard E, Irwin DJ, Rascovsky K, et al. Cognitive profile and markers of Alzheimer disease-type pathology in patients with lewy body dementias. *Neurology*. 2021;96(14):E1855-E1864. doi:10.1212/WNL.00000000000011699
55. Cousins K, Ash S, Olm CA, Grossman M. Longitudinal changes in semantic concreteness in Semantic Variant Primary Progressive Aphasia (svPPA). *eNeuro*. 2018;5(6):1-10. doi:10.1523/ENEURO.0197-18.2018
56. Cousins K, York C, Bauer L, Grossman M. Cognitive and anatomic double dissociation in the representation of concrete and abstract words in semantic variant and behavioral variant frontotemporal degeneration. *Neuropsychologia*. 2016;84:244-251. doi:10.1016/j.neuropsychologia.2016.02.025
57. Cousins K, Bove J, Giannini LAA, et al. Longitudinal naming and repetition relates to AD pathology and burden in autopsy-confirmed primary progressive aphasia. *Alzheimers Dement (N Y)*. 2021;7(1):1-10. doi:10.1002/trc2.12188
58. Massimo L, Powers JP, Evans LK, et al. Apathy in frontotemporal degeneration: neuroanatomical evidence of impaired goal-directed behavior. *Front Hum Neurosci*. 2015;9(November):1-10. doi:10.3389/fnhum.2015.00611
59. Nevler N, Ash S, Irwin DJ, Liberman M, Grossman M. Validated automatic speech biomarkers in primary progressive aphasia. *Ann Clin Transl Neurol*. 2019;6(1):4-14. doi:10.1002/acn3.653
60. Cho S, Nevler N, Shellikeri S, Ash S, Liberman M. Automatic classification of primary progressive aphasia patients using lexical and acoustic features. In: *Proceedings of Language Resources and Evaluation Conference 2020 Workshop on Resources and Processing of Linguistic, Para-Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric/Developmental Impairments*. 2020:60-65.
61. Shellikeri S, Cho S, Cousins K, et al. Parkinsonism and related disorders natural speech markers of Alzheimer's disease co-pathology in Lewy body dementias. *Parkinsonism Relat Disord*. 2022;102(August):94-100. doi:10.1016/j.parkreldis.2022.07.023
62. Lancashire I, Hirst G. Vocabulary changes in Agatha Christie's mysteries as an indication of dementia: a case study. Presented at the 19th Annual Rotman Research Institute Conference, *Cognitive Aging: Research and Practice*, 8-10 March 2009, Toronto. 2009;(March):1-5.
63. Kavé G, Goral M. Word retrieval in picture descriptions produced by individuals with Alzheimer's disease. *J Clin Exp Neuropsychol*. 2016;38(9):958-966. doi:10.1080/13803395.2016.1179266

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Cho S, Olm CA, Ash S, et al. Automatic classification of AD pathology in FTD phenotypes using natural speech. *Alzheimer's Dement*. 2024;1-13.
<https://doi.org/10.1002/alz.13748>