Original Paper

Automatic Speech Recognition Error Patterns for Differentiating Clinically Diagnosed Dementia, Parkinson Disease, and Healthy Aging: Cross-Sectional Association Study

June-Woo Kim[1,2,3], PhD, Prof Dr; Sung-Hoon Yoon[4,5], MD, Prof Dr Med; Hyunju Lee[3,6*], PhD, Prof Dr; Do-Young Kang[7,8,9,10*], MD, PhD, Prof Dr Med

[1]Department of Electronic Engineering, Wonkwang University, Iksan, Jeollabuk-do, Republic of Korea
[2]Department of Digital Healthcare, Graduate School of JABA, Wonkwang University, Iksan, Jeollabuk-do, Republic of Korea
[3]GIST InnoCORE AI-Nano Convergence Institute for Early Detection of Neurodegenerative Diseases, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea
[4]Department of Psychiatry, School of Medicine, Wonkwang University, Iksan, Jeollabuk-do, Republic of Korea
[5]Department of Psychiatry, Wonkwang University Hospital, Iksan, Jeollabuk-do, Republic of Korea
[6]Department of AI, College of Artificial Intelligence, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea
[7]Innovation Center for Atomic Science, DAU G-LAMP Project Group, Dong-A University, Busan, Republic of Korea
[8]Department of Translational Biomedical Sciences, Graduate School of Dong-A University, Busan, Republic of Korea
[9]Department of Data Sciences Convergence, Graduate School, Dong-A University, Busan, Republic of Korea
[10]Department of Nuclear Medicine, Dong-A University Medical Center, Dong-A University College of Medicine, Busan, Republic of Korea
*these authors contributed equally

Corresponding Author:

Do-Young Kang, MD, PhD, Prof Dr Med
Innovation Center for Atomic Science, DAU G-LAMP Project Group
Dong-A University
32, Daesingongwon-ro
Busan 49236
Republic of Korea
Phone: +82-51-240-5630
Email: dykang@dau.ac.kr

Abstract

Background: Speech-based digital biomarkers have been studied for aging-related neurodegenerative disorders, but conventional acoustic features often show clearer abnormalities in people with Parkinson disease (PD) than in people with Alzheimer disease (AD).

Objective: This study aimed to characterize acoustic features and automatic speech recognition (ASR) error patterns among older adults with clinically diagnosed AD/dementia or PD and cognitively unimpaired comparison participants, and examine their associations with cognitive screening scores.

Methods: This cross-sectional study analyzed a fixed 40-item Korean reading task. Acoustic features and Whisper recognition performance were summarized at the participant level. Whisper-Base analyses examined character error rate (CER), insertion, deletion, and substitution ratios, repeated-insertion ratio (RIR), and excess repetition rate (ERR). Group-comparison sensitivity analyses were adjusted for age, sex, and education. Associations with the Mini-Mental State Examination (MMSE) and Cognitive Impairment Screening Test (CIST) scores were assessed using partial Pearson correlations adjusted for diagnostic group, age, sex, and education. Hybrid CIST retained observed CIST scores and used MMSE-derived estimates when CIST was unavailable, supporting sensitivity analyses. False discovery rate correction addressed multiple comparisons.

Results: The primary analysis included 2026 diagnostic-group-speaker records; cognitive screening scores were available for 1508 participants. Unadjusted acoustic differences were more extensive in comparisons involving participants with PD than between the dementia and comparison groups. In unadjusted analyses, longer speech duration correlated with lower observed CIST scores (r=–0.504; q<.001). In adjusted analyses, insertion ratio showed the largest absolute correlation among the six

Whisper-Base metrics with MMSE (partial r=–0.265), observed CIST (partial r=–0.241), and Hybrid CIST (partial r=–0.249; all q<.001). CER, RIR, and ERR were also inversely associated with all 3 cognitive measures.

Conclusions: Conventional acoustic features showed broader differences in comparisons involving the PD group, whereas insertion ratio, as well as RIR and ERR, was associated with lower cognitive screening scores. These findings support further investigation of ASR error composition alongside acoustic measures.

JMIR Aging 2026;9:e99664; doi: 10.2196/99664

Keywords: Alzheimer disease; dementia; Parkinson disease; automatic speech recognition; speech biomarkers; cognitive impairment

Introduction

As populations age, scalable and low-burden digital biomarkers for neurodegenerative disease are becoming increasingly important. Speech is particularly attractive in this context because it can be collected noninvasively, repeatedly, and at relatively low cost, while capturing changes across motor, acoustic, linguistic, and cognitive domains [1-6]. Recent work has shown that automated speech and language analysis can support the detection of cognitive decline and Alzheimer disease (AD), including in non-English cohorts, and that clinically interpretable timing and language features may generalize across languages better than many black-box feature sets [2,3,6-8]. However, despite this promise, speech-based biomarkers remain limited by heterogeneity in task design, feature definitions, diagnostic ascertainment, and disease specificity, which has slowed translation into clinically usable tools [1,3,4,6,9].

A central challenge is that not all neurodegenerative disorders affect speech in the same way. In people with AD, speech changes often involve lexical-semantic inefficiency, reduced information content, disorganization, and impaired fluency, whereas in Parkinson disease (PD), speech abnormalities are more strongly linked to motor-speech dysfunction such as hypophonia, monopitch, imprecise articulation, and altered timing [6,10-13]. Consistent with this distinction, recent AD studies have shown that speech and language models can detect cognitive decline across multiple feature domains [7,8,14], while interpretable semantic markers have been reported to distinguish people with AD from healthy controls (HCs) but not people with PD, suggesting that AD-related speech disruption may differ qualitatively from Parkinsonian speech impairment [15-17]. Meanwhile, a recent systematic review and meta-analysis concluded that acoustic speech markers in people with AD are promising but remain heterogeneous, with only a limited subset of acoustic parameters showing consistent evidence across studies [1,4,6]. By contrast, recent PD studies have demonstrated that automated speech measures can differentiate people with PD from HCs across languages and can even track clinically relevant motor state changes [10,11,18]. These findings suggest that conventional acoustic analysis may capture PD-related motor-speech abnormalities more readily than AD-related cognitive-linguistic disruption [17,18].

This distinction raises an important question for digital speech biomarker research: if low-level acoustic differences are relatively subtle in people with AD, where does the disease-related signal emerge? One underexplored possibility is that it appears in automatic speech recognition (ASR) failure patterns rather than in conventional acoustic features alone. ASR is typically used as a preprocessing step for transcript generation, but recent work suggests that ASR errors themselves may carry clinically meaningful information [19-23]. Li et al [24] showed that imperfect ASR output can still support downstream dementia classification and argued that ASR errors may be informative rather than merely nuisance noise [23,25]. However, most prior studies have treated ASR performance in aggregate terms, such as overall word error rate or character error rate (CER), or focused on spontaneous speech and downstream classification [7,8,24]. Much less attention has been paid to the internal composition of ASR errors, including insertion-heavy and repetition-related failure modes, even though such patterns may be relevant to disrupted speech planning, reduced sequential stability, or perseverative tendencies [26]. This gap is especially important in underrepresented languages such as Korean, where clinically interpretable and cross-linguistically validated speech biomarkers remain limited [3,4,7,10].

An additional practical issue is that large-scale screening-oriented speech datasets are often assembled on the basis of specialist clinical diagnosis rather than biomarker-confirmed research frameworks. Although biomarker-defined classification can improve pathophysiologic specificity, it is not always feasible in pragmatic speech collection intended to support scalable screening. It is therefore important to determine whether speech-derived measures remain informative in clinically clear, specialist-diagnosed cohorts that more closely reflect real-world neurology practice [9,18].

In this study, we characterized acoustic features and ASR-derived error profiles in older Korean speakers with clinically assigned AD/dementia, PD, and HC status. The fixed 40-item reading protocol reduced variation in the intended linguistic content across participants. We examined speaker-level acoustic group differences and associations with cognitive screening scores and evaluated Whisper [27] models from Tiny to Large to describe recognition performance across model sizes. Whisper-Base analyses included CER, insertion ratio, deletion ratio, substitution ratio, repeated-insertion ratio (RIR), and excess repetition rate (ERR). Cognitive analyses used observed Mini-Mental State Examination (MMSE) and the Cognitive Impairment Screening Test (CIST) scores, with Hybrid CIST treated as a supportive sensitivity measure. We hypothesized

that comparisons involving the PD group would show broader acoustic differences and explored whether insertion- and repetition-related ASR patterns were associated with diagnostic group and cognition. We further examined whether these ASR error mechanisms were associated with lower cognitive screening scores as a secondary analysis.

Methods

Participants

Older adults were recruited between July 1 and December 31, 2021, from hospitals and affiliated care facilities, specifically Dong-A University Hospital and local older-adult care centers. Diagnostic ascertainment was based on specialist clinical assessment rather than biomarker-defined research criteria. Participants assigned to the PD group met the UK Parkinson's Disease Society Brain Bank clinical diagnostic criteria [28]. Participants assigned to the AD/dementia group were enrolled from clinically diagnosed dementia cases evaluated in a dementia-specialty practice according to *DSM-IV* (*Diagnostic and Statistical Manual of Mental Disorders* [Fourth Edition])-based clinical criteria [29]. The AD/dementia label is used in this study to denote clinically diagnosed dementia cases evaluated in a dementia-specialty setting, rather than a biomarker-defined AD cohort. Clinical ascertainment was performed by 2 board-certified neurologists, including 1 dementia specialist and 1 PD specialist. Eligibility for the clinical groups also required participation in the relevant assessment items of the Seoul Neuropsychological Screening Battery. Participants with PD and comorbid dementia were not included in the PD group.

Participants with other neurologic disorders such as vascular disease were excluded. Cognitively unimpaired older adults recruited as controls were assigned to the HC group and were required to have normal neurologic examination findings and normal imaging or nonimaging evaluation results. When brain magnetic resonance imaging was available as part of control screening, individuals with abnormal striatal findings were excluded. Biomarker-defined stratification, such as ATN (amyloid, tau, neurodegeneration) classification [30], amyloid positron emission tomography, or dopamine positron emission tomography, was not part of the present enrollment protocol because the speech collection was designed as a pragmatic, screening-oriented assessment intended to support large-scale participant recruitment rather than to construct biomarker-defined cohorts. No enrolled participant had a visual impairment that precluded the completion of the reading task.

For the speaker-level acoustic and ASR analyses, participants were required to have all 40 analyzed recordings and the metadata necessary for diagnostic group assignment and downstream analysis. Participants were excluded from these analyses if any of the 40 recordings were missing or if required metadata were unavailable. For the clinically linked cognitive analyses, speech records were matched to cognitive annotation records, and participants were additionally required to have at least 1 observed MMSE or CIST score. Covariate-adjusted analyses including age, sex, and education years were restricted to participants with complete covariate data.

Recruitment and analysis were not stratified using a complete cohort-wide measure of disease stage. Clinical Dementia Rating [31], Hoehn and Yahr stage [32], Unified Parkinson's Disease Rating Scale scores [33], disease-specific severity measures, functional status, and direct literacy measures were available only for subsets of participants or were not consistently retained for the complete analytic cohort. These variables were therefore not imputed or included as cohort-wide covariates. Cognitive status was described using the available MMSE and CIST distributions.

Medication exposure was not systematically recorded at the individual level. The study protocol did not require temporary withdrawal of prescribed treatment, and patients were generally assessed under their usual treatment conditions. Specific medication identity, dose, time since the last dose, and PD ON/OFF state were not consistently available and therefore could not be included in the analyses.

Speech Recording Task

The analyzed speech task comprised 40 fixed Korean reading items drawn from the study protocol. The set consisted of 30 independent sentences and 10 sequential segments from a longer connected-text passage. All participants included in the analysis completed the identical 40-item set. Separate isolated-word and sustained-phonation tasks included in the broader data-collection protocol were not analyzed in the present study. Based on the normalized reference texts used for ASR scoring, item length ranged from 2 to 22 whitespace-delimited Korean orthographic units (eojeol) and from 4 to 64 characters excluding spaces. The items were not prospectively matched or stratified by lexical frequency, syntactic complexity, or psycholinguistic difficulty. Thus, the fixed-content design controlled between-participant variation in lexical content and task structure but did not eliminate item-level difficulty effects. The complete Korean reference texts and meaning-focused English translations are provided in Multimedia Appendix 1. The English translations were provided for interpretation by non-Korean readers and were not used for ASR scoring or statistical analysis.

Recording Protocol

Speech was recorded using Britz BE-STM700 microphones (Britz International Co, Ltd) and laptop-based acquisition systems. Data collection teams were deployed to several participating locations, including Dong-A University Hospital, welfare centers, local older-adult associations, day-care centers, senior centers, and nursing facilities. Recording spaces were selected to minimize environmental noise and interruptions that could materially affect the speech signal, and sound-absorbing partitions were used when needed. The collected speech data underwent independent external quality validation by the Telecommunications Technology Association in Korea. Before acoustic and ASR analysis, audio files were converted to mono and resampled to 16 kHz using a standardized preprocessing pipeline.

Cognitive Measures and Construction of Hybrid CIST

Because MMSE and CIST were not jointly available for many participants, we additionally defined a Hybrid CIST variable for sensitivity analyses aimed at maximizing the cognitively linked sample. To align MMSE with the CIST scale, we fit an ordinary least squares linear regression model using participants with both MMSE and CIST available. For participants with an observed CIST score, the original CIST value was retained. For participants with MMSE but without CIST, a mapped CIST-aligned value was generated using the fitted regression equation and clipped to the valid 0-30 range. These observed and mapped values were combined to create the Hybrid CIST variable, which was used only as a sensitivity cognitive measure to examine whether the overall association patterns were consistent in the largest available sample. Primary cognitive analyses were based on observed MMSE and observed CIST separately, whereas Hybrid CIST results were interpreted as supportive rather than primary evidence. The empirical relationship between MMSE and CIST, the fitted mapping equation, and the final score-availability counts are reported in the *Results* section.

Acoustic Feature Extraction

Utterance-level acoustic features were extracted from each raw audio sample using a standardized pipeline. The extracted features covered temporal, glottal, spectral, energy, pitch/prosodic, formant, and Mel-spectrogram domains, including speech duration; glottal opening, closing, and closed-phase summaries; spectral centroid, bandwidth, and rolloff; root-mean-square (RMS) energy; tempo; fundamental frequency (F0); pitch; prosody; spectral magnitude; zero-crossing rate (ZCR); formants F1-F4; and Mel-spectrogram. F0 was estimated using YIN [34] with a frequency range of 80 to 450 Hz. Spectral centroid, bandwidth, rolloff, and RMS were summarized in the log domain, and formants were estimated using Burg linear prediction coefficient-based [35] formant tracking. Acoustic features were extracted separately for each of the 40 recorded speech samples. For each participant, utterance-level values were then averaged across the complete 40-item set to obtain 1 speaker-level summary value per feature. Accordingly, all inferential analyses for acoustic group comparisons were conducted at the participant level, with each participant contributing a single aggregated value for each acoustic feature.

Automatic Speech Recognition Evaluation With Whisper

To evaluate ASR performance across model scales, we used Whisper [27] Tiny, Base, Small, Medium, and Large for Korean transcription. For Whisper decoding, each waveform was loaded, padded, or trimmed to 30 seconds according to the model's standard preprocessing routine and transcribed in Korean transcription mode. Reference and hypothesis texts were normalized using the same Korean text normalization procedure, including punctuation removal and whitespace normalization. CER was computed as the Levenshtein [36] edit distance between normalized reference and hypothesis character strings after removing spaces, divided by the number of reference characters. CER was first computed separately for each of the 40 reading items and then averaged within each participant to obtain one speaker-level mean CER value per model. Group-level CER trends across model sizes were summarized descriptively using these speaker-level mean CER values.

ASR Error Patterns Analysis

To characterize ASR errors beyond aggregate CER, we analyzed Whisper outputs using dynamic-programming alignment between whitespace-delimited tokens in the normalized Korean reference and hypothesis texts. These tokens correspond to Korean orthographic units (eojeol), not individual characters or morphologically segmented words. Alignment operations were classified as matches, substitutions, deletions, or insertions. Counts were calculated separately for each available utterance, converted to the measures defined below, and averaged within speakers. Whisper-Base was used for the primary error-pattern analysis. For each utterance, the numbers of substitution, deletion, and insertion operations were counted, and utterance-level measures were subsequently averaged across the 40 reading items to obtain speaker-level ASR error pattern measures.

Let I, D, and S denote the numbers of word-token insertion, deletion, and substitution operations, respectively. The insertion ratio was defined as $I/(I + D + S)$, deletion ratio as $D/(I + D + S)$, and substitution ratio as $S/(I + D + S)$. These measures describe the composition of edit operations rather than error counts divided by reference length. All 3 ratios were assigned a value of 0 when $I + D + S = 0$.

RIR was defined as R/I, where R is the number of word-token insertion operations for which the inserted token is identical to the immediately preceding or following token in the ASR hypothesis. RIR was assigned a value of 0 when I=0. This measure quantifies repetition-related insertions in the recognizer output.

ERR was defined as $\max(Hrep-Rrep, 0)/Nref$, where Hrep and Rrep are the numbers of adjacent identical word-token pairs in the hypothesis and reference, respectively, and Nref is the number of reference tokens. The implementation used $\max(1, Nref)$ as the denominator. RIR and ERR were calculated per utterance and averaged across available utterances within each speaker. Separately, substitution distance was computed for each aligned substituted token pair as the character-level Levenshtein distance divided by the longer token length. These distances were averaged across substituted pairs within an utterance and then across utterances with at least 1 substitution within a speaker.

Statistical Analysis

All inferential analyses were conducted at the participant level. Continuous demographic and clinical variables were compared across the AD/dementia, PD, and HC groups using Welch 1-way ANOVA, followed by Games-Howell tests for pairwise post hoc comparisons. Sex distributions were compared across the 3 groups using the Pearson chi-square

test, followed by pairwise Fisher exact tests with Holm correction.

For acoustic and ASR group comparisons, pairwise differences between AD/dementia and PD, AD/dementia and HC, and PD and HC were evaluated using 2-sided Mann-Whitney *U* tests [37]. The Cliff δ was used to summarize unadjusted acoustic effect sizes [38]. For sensitivity analyses in the cognitively linked cohort, each standardized speech measure was modeled as a function of diagnostic group, age, sex, and years of education. Pairwise adjusted differences were reported in outcome SD units with HC3 heteroskedasticity-robust 95% CIs and 2-sided Wald *P* values. Benjamini-Hochberg false discovery rate (BH-FDR) correction was applied across 19 acoustic features within each comparison [39]. For the ASR sensitivity analyses, correction was applied across 6 ASR error metrics in Multimedia Appendix 2 and across 7 metrics, including substitution distance, in the adjusted ASR error-profile analysis.

Cognitive association analyses included 19 acoustic features and 6 ASR metrics: CER, insertion ratio, deletion ratio, substitution ratio, RIR, and ERR. The original acoustic associations were evaluated using Pearson correlations with available MMSE, observed CIST, and Hybrid CIST scores. Covariate-adjusted acoustic associations were examined as sensitivity analyses using partial Pearson correlations. ASR-cognition associations were also evaluated using partial Pearson correlations adjusted for diagnostic group, age, sex, and years of education. For each analysis, both the speech measure and cognitive score were residualized against the same covariates, and Pearson correlations were calculated between the residuals. BH-FDR correction [39] was applied separately across the 19 acoustic features or 6 ASR metrics within each cognitive measure, with adjusted and unadjusted analyses corrected separately. Observed MMSE and CIST were the primary cognitive measures; Hybrid CIST was interpreted as supportive. Group comparisons are reported using unadjusted effect sizes or adjusted regression coefficients with 95% CIs, and cognitive associations are reported as Pearson or partial Pearson coefficients with *q* values. Cross-model sensitivity analyses repeated the 6-metric group comparisons and cognitive association analyses using Whisper-Base, Small, and Large outputs from the same participants and matched utterances. The same covariates were applied, with BH-FDR correction across 6 metrics separately within each model and group comparison or cognitive measure (Multimedia Appendices 3 and 4).

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Institutional Review Board of Dong-A University Hospital (DAUHURB-21-049, approved on March 17, 2021). Written informed consent was obtained from all participants or their legally authorized representatives, as applicable.

Results

Participant Flow and Characteristics

Of 2171 recruited older adults, 2026 participants with complete 40-item recordings were retained for the primary acoustic and ASR analyses (HC=937, AD=852, PD=237), as described in Figure 1. For clinically linked cognitive analyses, 1508 unique participants had at least 1 available cognitive screening score and were included in the final analytic cohort (HC=913, AD=369, PD=226). Within this cohort, MMSE scores were available for 484 participants, observed CIST scores for 1189 participants, MMSE only for 319 participants, CIST only for 1024 participants, and both MMSE and CIST for 165 participants. These participants had complete age, sex, and education data.

As shown in Table 1, the mean age was 76.37 (SD 8.01; range 53-98) years in the AD/dementia group, 67.66 (SD 8.57; range 43-90) years in the PD group, and 72.09 (SD 8.18; range 51-96) years in the HC group. Mean years of education were 8.79 (SD 4.43; range 0-21) in the AD/dementia group, 10.26 (SD 4.22; range 0.5-24; n=225) in the PD group, and 10.11 (SD 4.49; range 0-21) in the HC group. Age, sex distribution, and years of education differed across the 3 diagnostic groups, although the magnitude and pairwise pattern of these differences varied by characteristic.

Observed mean MMSE scores were 20.63 (SD 5.26; range 5-30; n=252) in the AD/dementia group, 27.18 (SD 2.67; range 15-30; n=221) in the PD group, and 26.45 (SD 2.70; range 21-30; n=11) in the HC group. Observed mean CIST scores were 12.65 (SD 6.71; range 2-30; n=199), 22.60 (SD 5.58; range 7-30; n=77), and 24.67 (SD 3.81; range 9-30; n=913), respectively. Mean Hybrid CIST scores were 13.50 (SD 6.27; n=369), 21.99 (SD 4.04; n=226), and 24.67 (SD 3.81; n=913), respectively. Cognitive-score availability differed substantially by diagnostic group, particularly for observed MMSE scores in the HC group (n=11).

Figure 1. Flow of participant selection and cognitive score availability in the study cohort. AD: Alzheimer disease; CIST: Cognitive Impairment Screening Test; HC: healthy control; MMSE: Mini-Mental State Examination; PD: Parkinson disease.

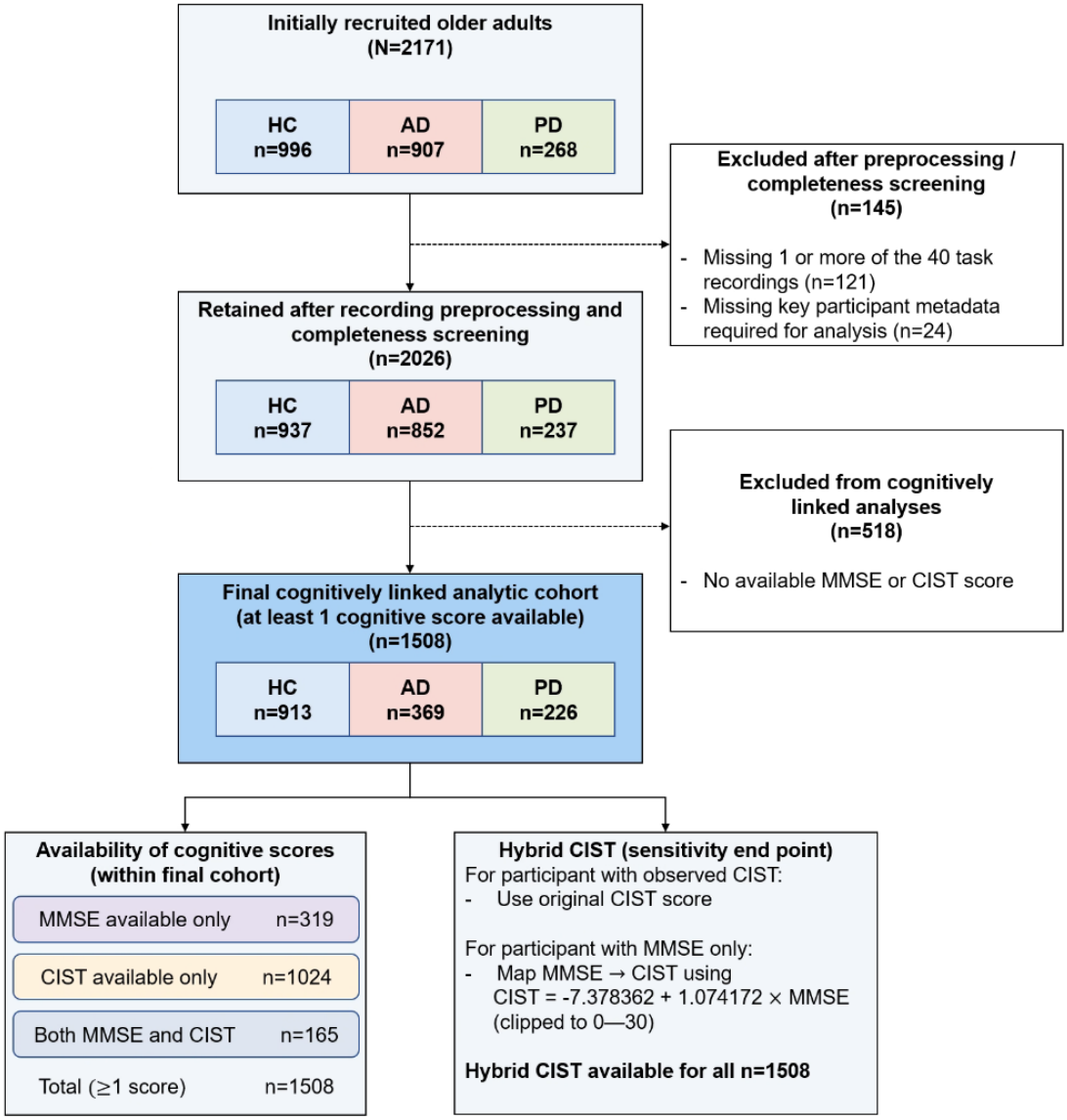


Table 1. Demographic and clinical characteristics of the cognitively linked analytic cohort[a].

Variables	AD[b] (n=369)[c]	PD[d] (n=226)	HC[e] (n=913)	Overall (*P* value)[f]	Post hoc (*P* value)[g]	Post hoc (*P* value)[h]	Post hoc (*P* value)[i]
Age[j] (y)	76.37 (8.01)	67.66 (8.57)	72.09 (8.18)	<.001[k]	<.001[k]	<.001[k]	<.001[k]
Sex[l], n (%)				.003[m]	.006[m]	.61	.006[m]
Male	141 (38.2)	116 (51.3)	363 (39.8)				
Female	228 (61.8)	110 (48.7)	550 (60.2)				
Education[n] (y)	8.79 (4.43)	10.26 (4.22)	10.11 (4.49)	<.001[k]	<.001[k]	<.001[k]	.885
MMSE score[o]	20.63 (5.26); n=252	27.18 (2.67); n=221	26.45 (2.70); n=11	<.001[k]	<.001[k]	<.001[k]	.668
CIST score[p]	12.65 (6.71); n=199	22.60 (5.58); n=77	24.67 (3.81); n=913	<.001[k]	<.001[k]	<.001[k]	.006[m]
Hybrid CIST score[q]	13.50 (6.27); n=369	21.99 (4.04); n=226	24.67 (3.81); n=913	<.001[k]	<.001[k]	<.001[k]	<.001[k]

[a]Data are presented as mean (SD) unless otherwise indicated. Available sample sizes are reported when data were missing.
[b]AD: Alzheimer disease.
[c]AD denotes the clinically diagnosed AD/dementia group.
[d]PD: Parkinson disease.
[e]HC: healthy control.
[f]Overall *P* value: Welch 1-way ANOVA for continuous variables and Pearson X^2 test for sex.
[g]Post hoc comparison between AD and PD.

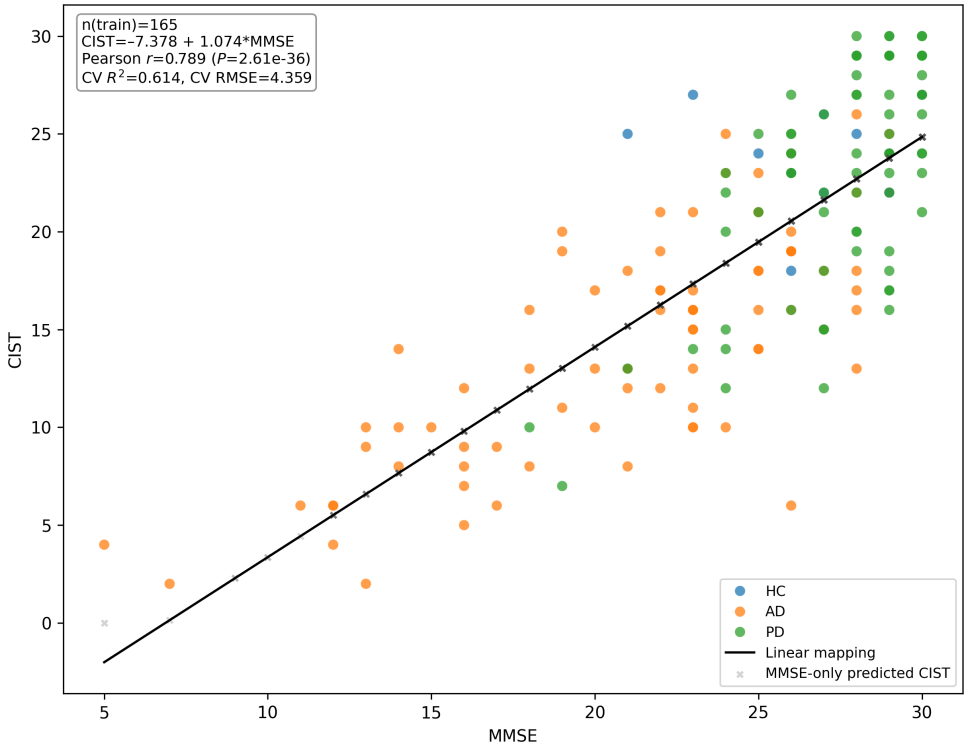
[h]Post hoc comparison between AD and HC.
[i]Post hoc comparison between PD and HC.
[j]Age: Welch 1-way ANOVA overall; Games-Howell post hoc tests.
[k]$P<.001$.
[l]Sex: Pearson chi-square test overall; pairwise 2-sided Fisher exact tests with Holm correction post hoc.
[m]$P<.01$.
[n]Education: Welch 1-way ANOVA overall; Games-Howell post hoc tests.
[o]MMSE: Mini-Mental State Examination. Welch 1-way ANOVA overall; Games-Howell post hoc tests.
[p]CIST: Cognitive Impairment Screening Test. Welch 1-way ANOVA overall; Games-Howell post hoc tests.
[q]Hybrid CIST retained observed CIST and used an MMSE-derived estimate when CIST was missing. Welch 1-way ANOVA overall; Games-Howell post hoc tests.

MMSE-to-CIST Mapping and Hybrid CIST Construction

Among the 165 participants with both MMSE and CIST, the 2 scores were strongly correlated (Pearson r=0.789; P<.001) as illustrated in Figure 2. This trend was already shown in previous studies [40]. The fitted linear mapping equation was as follows: CIST = –7.378362 + 1.074172 × MMSE.

Five-fold cross-validation yielded an R^2 of 0.614 and a root-mean-square error of 4.36, indicating moderate predictive alignment between the 2 screening instruments. This mapping was then used to generate CIST-aligned values for participants with MMSE but without observed CIST, yielding a Hybrid CIST variable available for all 1508 participants.

Figure 2. Mini-Mental State Examination (MMSE)-to-Cognitive Impairment Screening Test (CIST) mapping scatter. AD: Alzheimer disease; CV: cross-validation; HC: healthy control; PD: Parkinson disease; RMSE: root-mean-square error.



Acoustic Group Differences and Acoustic-Cognitive Associations

Speaker-level acoustic analyses showed fewer significant group differences between the AD/dementia and HC groups than in comparisons involving PD group. After BH-FDR correction across the 19 acoustic features, 5 features differed significantly between the AD/dementia and HC groups: speech duration, spectral centroid, tempo, ZCR, and formant F2. Compared with participants in the HC group, participants with AD/dementia had longer mean speech duration (6.04 vs 5.19 seconds) and lower ZCR (0.0629 vs 0.0653). Significant differences were observed in 14 features between the AD/dementia and PD groups and 13 features between PD and HC. Complete results are shown in Figure 3 and Multimedia Appendix 5, with selected comparisons presented in Table 2.

Figure 3. Unadjusted acoustic effect-size across diagnostic-group comparisons. Cells show Cliff δ for 19 acoustic features. Negative δ values indicate higher observations in the first-named group under the retained sign convention. Asterisks denote Benjamini-Hochberg false discovery rate (BH-FDR)-adjusted significance across 19 features within each comparison (*q<.05, **q<.01, ***q<.001). The corresponding adjusted coefficients are reported separately in Multimedia Appendices 6 and 7. AD: Alzheimer disease; GCL: glottal closed phase; GCP: glottal closing phase; GOP: glottal opening phase; HC: healthy control; PD: Parkinson disease; RMS: root-mean-square.

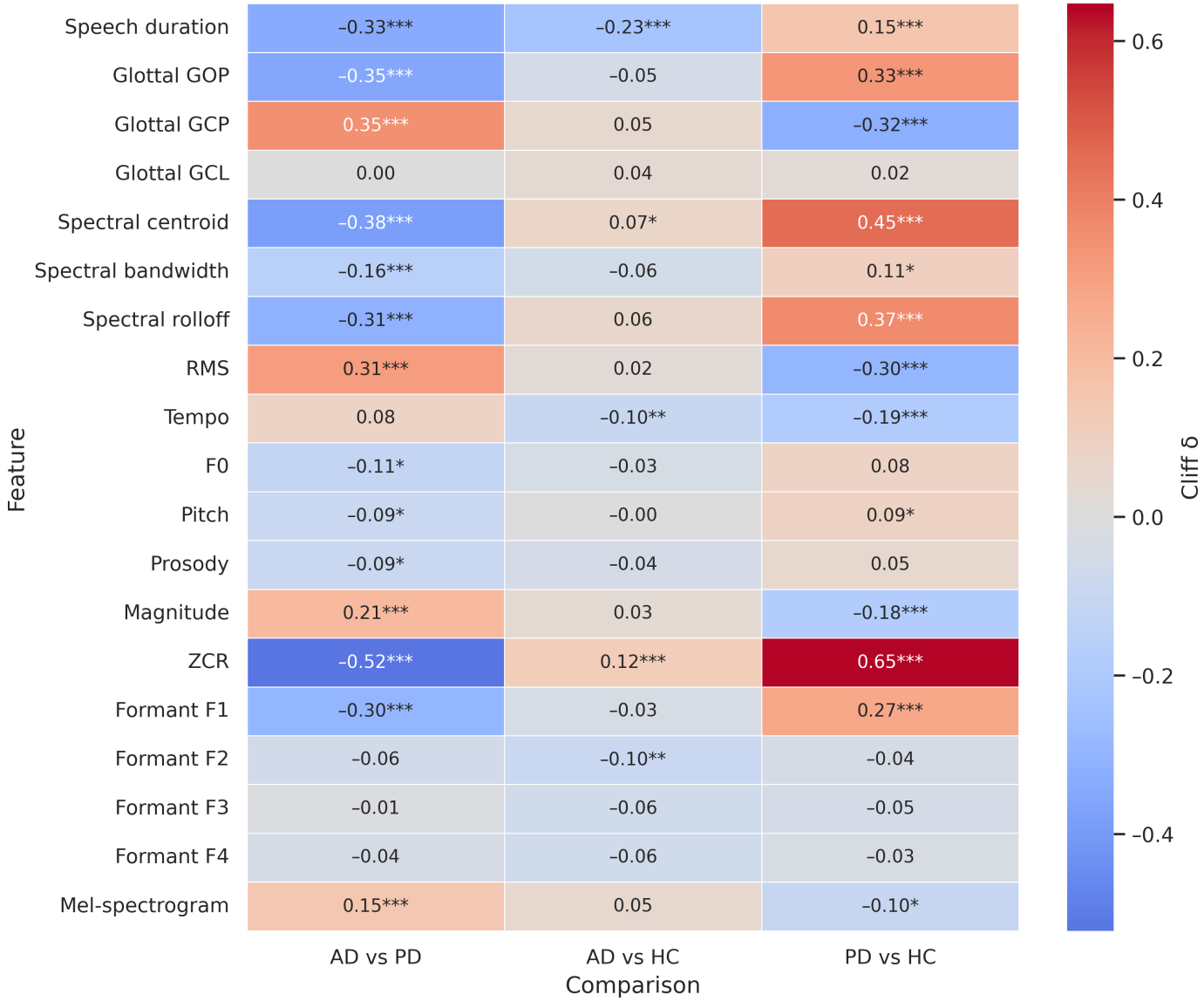


Table 2. Key acoustic group differences in the speaker-level analysis.

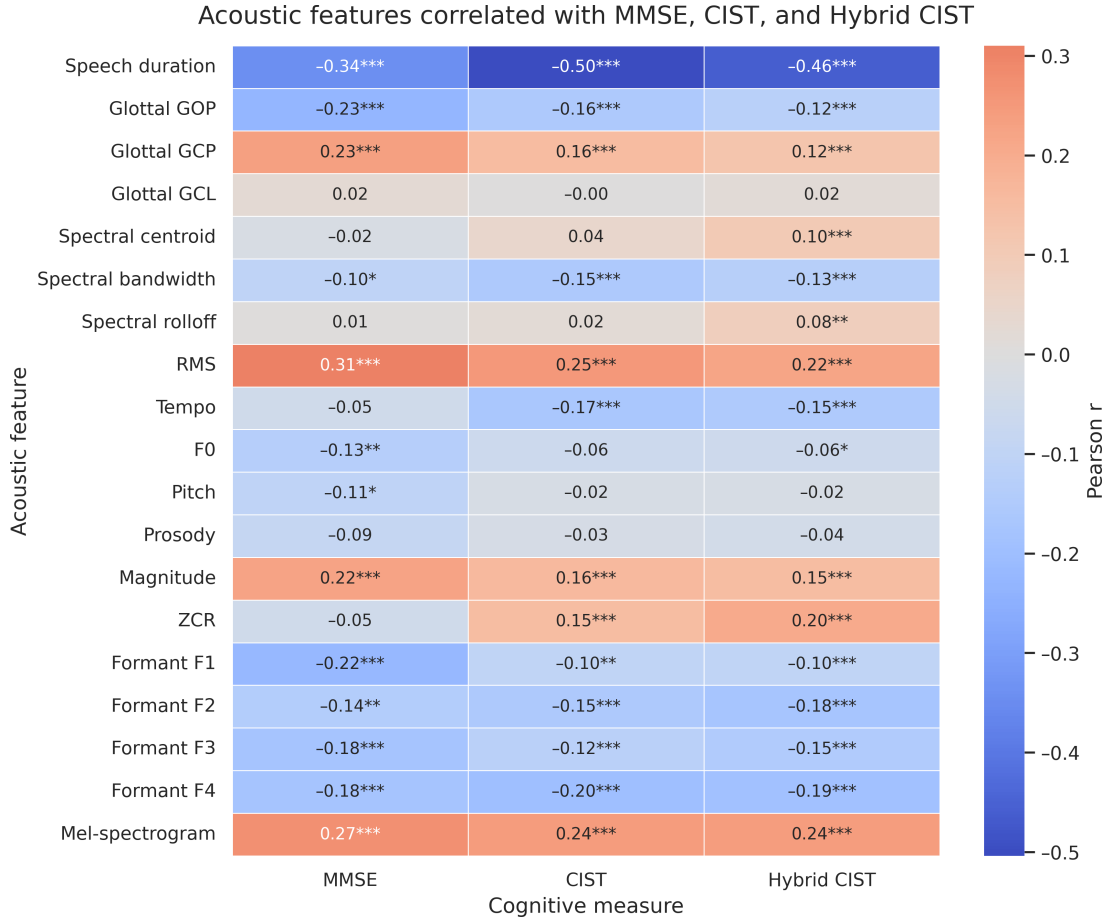
Comparison	Feature	Group 1, mean (SD)	Group 2, mean (SD)	Cliff, δ[a]	FDR[b] (q value)[c]
AD[d,e] vs HC[f]	Speech duration	6.0388 (2.7459)	5.1921 (1.5725)	–0.229	<.001
AD vs HC	ZCR[g]	0.0629 (0.0119)	0.0653 (0.0102)	0.119	<.001
AD vs PD[h]	ZCR	0.0629 (0.0119)	0.0517 (0.0115)	–0.522	<.001
AD vs PD	Spectral centroid	6.8796 (0.1308)	6.7869 (0.1482)	–0.382	<.001
AD vs PD	Speech duration	6.0388 (2.7459)	5.0937 (2.0036)	–0.335	<.001
PD vs HC	ZCR	0.0517 (0.0115)	0.0653 (0.0102)	0.647	<.001
PD vs HC	Spectral centroid	6.7869 (0.1482)	6.8977 (0.1207)	0.452	<.001
PD vs HC	Speech duration	5.0937 (2.0036)	5.1921 (1.5725)	0.152	<.001

[a]Cliff δ follows the retained convention P(group 1<group 2) – P(group 1>group 2); negative values indicate higher observations in the first-named group.
[b]FDR: false discovery rate.
[c]FDR q values were obtained by applying Benjamini-Hochberg false discovery rate correction across 19 acoustic features within each 2-sided Mann-Whitney U comparison.
[d]AD: Alzheimer disease.

[e]AD denotes the clinically diagnosed AD/dementia group.
[f]HC: healthy control.
[g]ZCR: zero-crossing rate.
[h]PD: Parkinson disease.

Longer speech duration showed the strongest absolute acoustic correlation with each cognitive measure in Figure 4: MMSE (r=–0.344), observed CIST (r=–0.504), and Hybrid CIST (r=–0.462; all q<.001). In the sensitivity analysis adjusted for diagnostic group, age, sex, and years of education, speech duration remained inversely associated with MMSE (partial r=–0.259), observed CIST (partial r=–0.226), and Hybrid CIST (partial r=–0.242; all q<.001; Multimedia Appendix 8). RMS and Mel-spectrogram values retained positive adjusted associations with all 3 cognitive measures. Hybrid CIST findings were treated as supportive because this measure partly consists of MMSE-derived estimates.

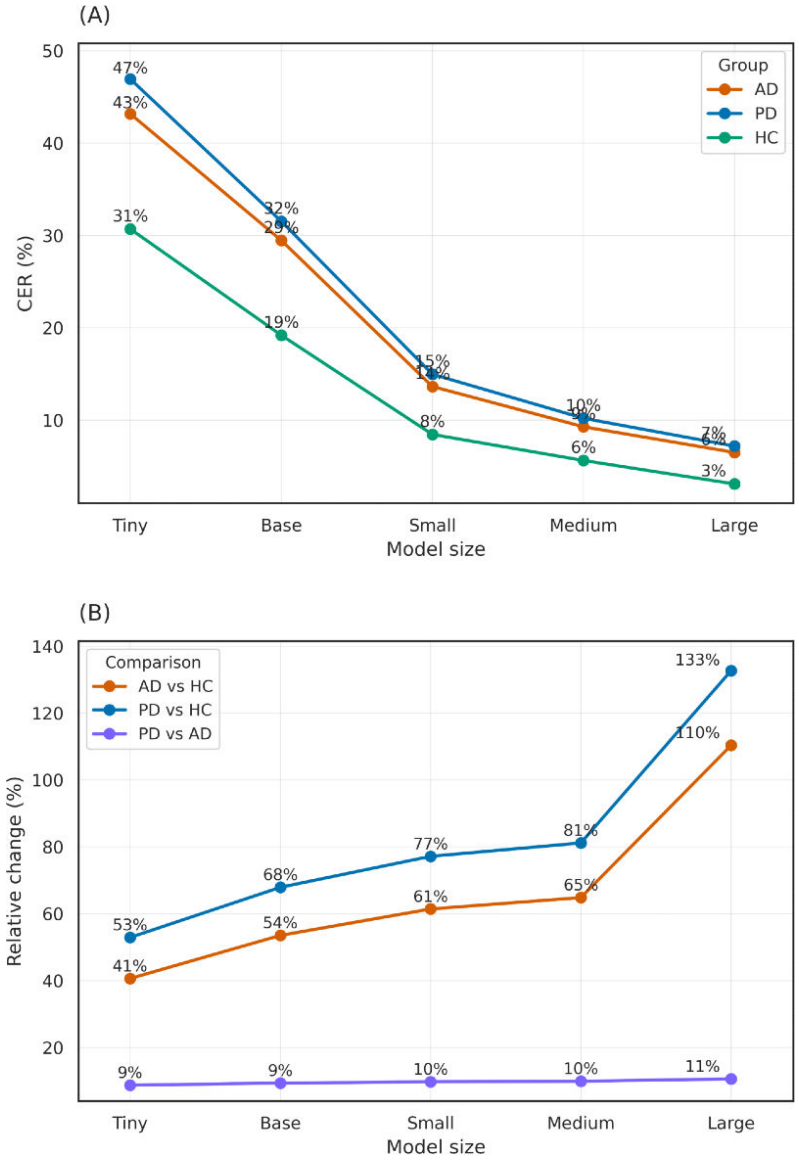
Figure 4. Unadjusted associations between speaker-level acoustic features and cognitive scores. Cells show Pearson correlation coefficients for Mini-Mental State Examination (MMSE; n=484), observed Cognitive Impairment Screening Test (CIST), and Hybrid CIST. Asterisks denote Benjamini-Hochberg false discovery rate (BH-FDR)–adjusted significance across the 19 acoustic features within each cognitive measure (*q<.05, **q<.01, ***q<.001). Covariate-adjusted sensitivity results are shown in Multimedia Appendices 8 and 9. GCL: glottal closed phase; GCP: glottal closing phase; GOP: glottal opening phase; RMS: root-mean-square; ZCR: zero-crossing.



ASR Performance Across Whisper Model Sizes

As illustrated in Figure 5, mean CER decreased from Whisper Tiny to Large in the AD/dementia group (mean 43.17%-6.50%, SD 33.44%-9.77%), the PD group (46.94%-7.19%, SD 39.09%-9.62%), and HC group (mean 30.70%-3.09%, SD 19.17%-5.16%). Relative disease-vs-HC CER gaps increased from 40.6% to 110.4% for the AD/dementia group and from 52.9% to 132.7% for the PD group. However, the absolute gaps decreased from 12.47 to 3.41 percentage points for the AD/dementia group and from 16.24 to 4.10 percentage points for the PD group. Thus, relative disparities persisted as overall error rates decreased. Item-level CER distributions are described separately in the original utterance-level supplement in Multimedia Appendix 9.

Figure 5. Automatic speech recognition (ASR) performance across Whisper model sizes. (A) Mean speaker-level CER in the healthy control (HC), Alzheimer disease (AD), and Parkinson disease (PD) groups across Whisper Tiny, Base, Small, Medium, and Large models. (B) Relative CER differences across model sizes for AD vs HC, PD vs HC, and PD vs AD. CER: character error rate.



Group Differences in ASR Error Patterns

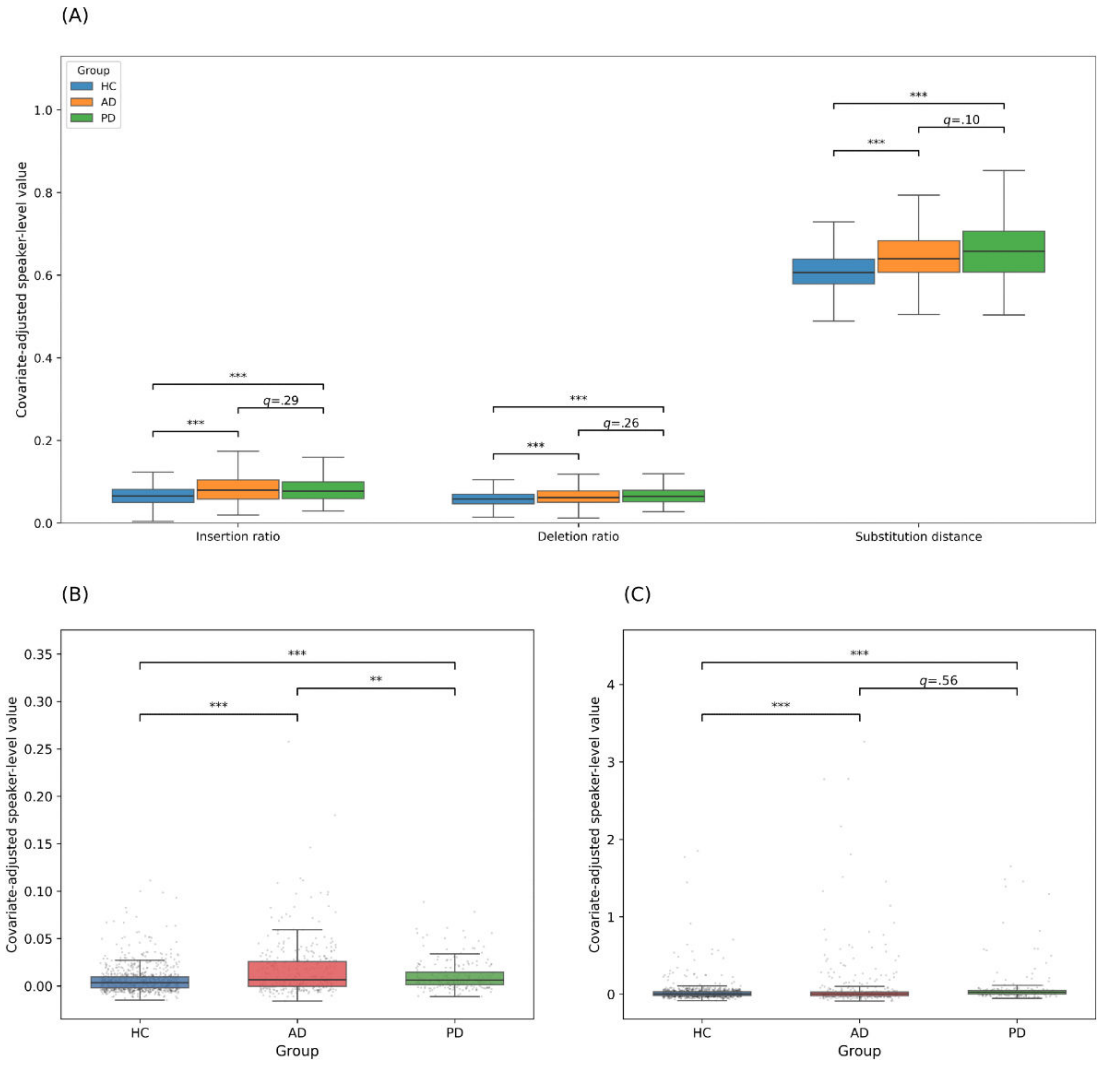
In the unadjusted analysis as described in Table 3, insertion ratio was higher in the AD/dementia group than the HC group (0.0858 vs 0.0664; q<.001) and in the PD group than in the HC group (0.0796 vs 0.0664; q<.001), whereas the AD/dementia-PD contrast was not significant after false discovery rate (FDR) correction (q=.10). Deletion ratio was higher in the PD group than in the HC group (0.0674 vs 0.0593; q<.001) and the AD/dementia group (0.0674 vs 0.0618; q=.04). Figure 6 instead presents age-, sex-, and education-adjusted sensitivity results. In that analysis, insertion ratio, deletion ratio, and substitution distance were higher in both clinical groups than in the HC group, but their AD/dementia-PD contrasts did not survive FDR correction (q=.29, .26, and .01, respectively). In full-cohort analysis, RIR and ERR were higher in both clinical groups than in the HC group, while AD/dementia-PD differences were not significant after FDR correction (Table 3). In the adjusted sensitivity analysis shown in the lower panels of Figure 6, RIR remained higher in both clinical groups than in the HC group and was also higher in AD/dementia than PD (q=.01). ERR remained higher in both clinical groups than in the HC group, but the AD/dementia-PD contrast was not significant (q=.56). The corresponding 6-metric adjusted regression results are provided in Multimedia Appendix 2, where the AD/dementia-PD RIR contrast has a q value of .009.

Table 3. Speaker-level automatic speech recognition (ASR) error pattern differences across diagnostic groups.

Metric	AD[a], mean (SD)[b]	PD[c], mean (SD)	HC[d], mean (SD)	AD vs HC (q value)[e]	PD vs HC (q value)	AD vs PD (q value)
CER[f]	0.2946 (0.2523)	0.3152 (0.2533)	0.1919 (0.1248)	<.001[g]	<.001[g]	.48
Insertion ratio	0.0858 (0.0468)	0.0796 (0.0419)	0.0664 (0.0308)	<.001[g]	<.001[g]	.10
Deletion ratio	0.0618 (0.0219)	0.0674 (0.0247)	0.0593 (0.0188)	.04[h]	<.001[g]	.04[h]
Substitution ratio	0.6467 (0.0965)	0.6471 (0.1090)	0.5990 (0.1024)	<.001[g]	<.001[g]	.75
RIR[i]	0.0137 (0.0256)	0.0093 (0.0163)	0.0060 (0.0147)	<.001[g]	<.001[g]	.12
ERR[j]	0.0742 (0.2897)	0.0655 (0.2386)	0.0231 (0.1314)	<.001[g]	<.001[g]	.53

[a]AD: Alzheimer disease.
[b]AD denotes the clinically diagnosed AD/dementia group.
[c]PD: Parkinson disease.
[d]HC: healthy control.
[e]q values are Benjamini-Hochberg false discovery rate-adjusted P values from 2-sided Mann-Whitney U tests.
[f]CER: character error rate.
[g]q<.001.
[h]q<.05.
[i]RIR: repeated-insertion ratio.
[j]ERR: excess repetition rate.

Figure 6. Covariate-adjusted automatic speech recognition (ASR) error profiles by diagnostic group in the cognitively linked group. (A) Insertion ratio, deletion ratio, and substitution distance by group. (B) Repeated-insertion ratio (RIR) by group. (C) Excess repetition rate (ERR) by group. Values were adjusted for age, sex, and years of education while retaining diagnostic-group differences. Boxes represent interquartile ranges, center lines indicate medians, and whiskers extend to the most extreme observations within 1.5 IQR. Brackets represent pairwise regression contrasts with HC3 robust inference. Benjamini-Hochberg false discovery rate (BH-FDR) correction was applied across seven ASR measures within each comparison (*q<.05, **q<.01, ***q<.001); exact q values are shown for nonsignificant comparisons.



Associations Between ASR Error Patterns and Cognition

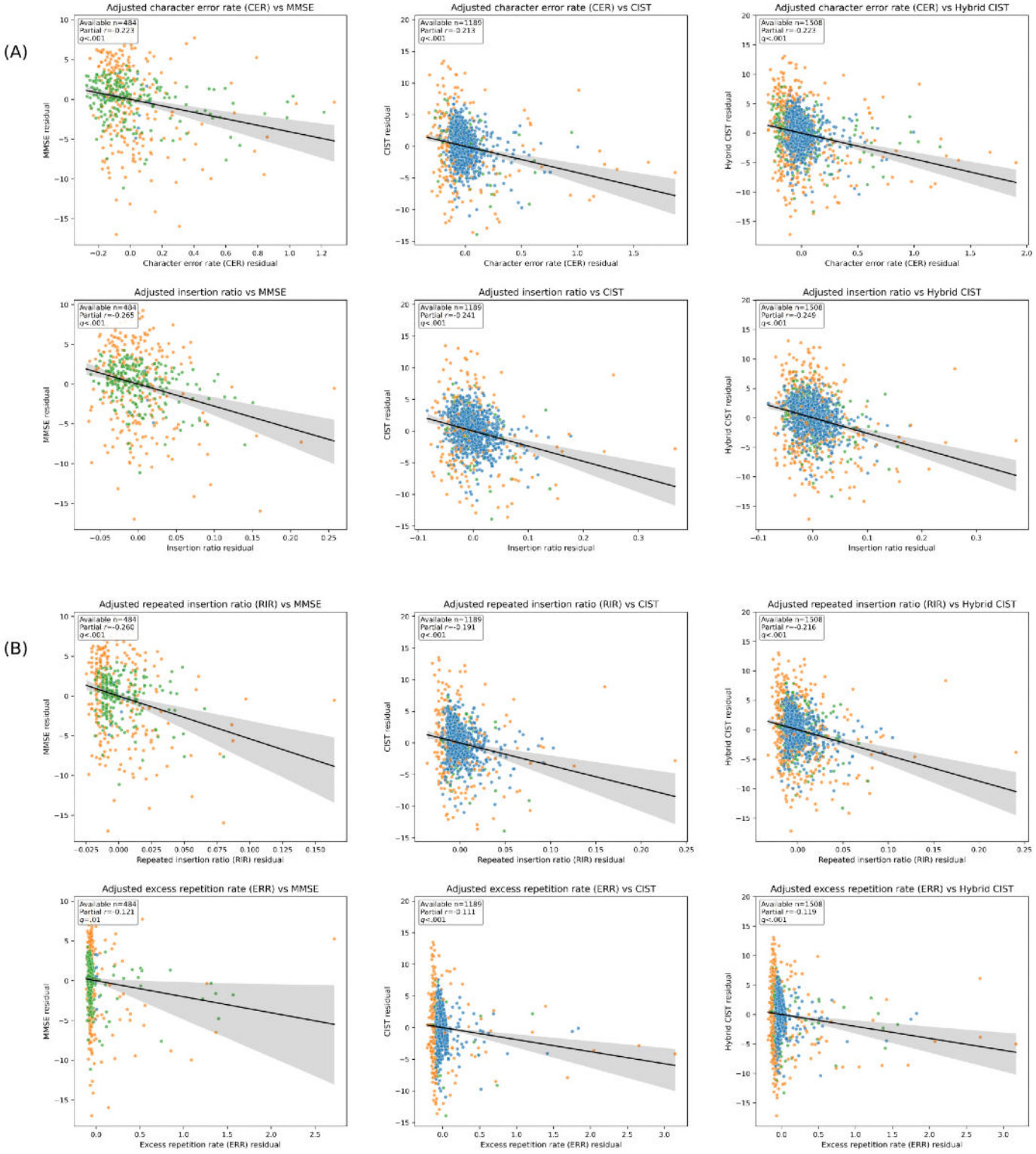
ASR-cognition associations were adjusted for diagnostic group, age, sex, and years of education using partial Pearson correlations (Table 4and Figure 7). For MMSE, partial correlations were –0.223 for CER, –0.265 for insertion ratio, –0.260 for RIR (all q<.001), and –0.121 for ERR (q=.01). Deletion ratio was also inversely associated with MMSE (partial r=–0.100; q=.04), whereas substitution ratio was not (partial r=–0.075; q=.10). For observed CIST, partial correlations were –0.213 for CER, –0.241 for insertion ratio, –0.191 for RIR, and –0.111 for ERR (all q<.001); deletion and substitution ratios were not significant after FDR correction. In the supportive Hybrid CIST analysis, CER, insertion ratio, RIR, and ERR remained inversely associated with scores (partial r=–0.223, –0.249, –0.216, and –0.119, respectively; all q<.001). Substitution ratio was also associated with Hybrid CIST (partial r=–0.080; q=.002), whereas deletion ratio was not (partial r=–0.045; q=.08). In Whisper-Base, insertion ratio showed the largest absolute adjusted correlation with each cognitive measure among the 6 ASR metrics. These findings indicate that cognition-related ASR degradation was not captured by total error burden alone and was particularly enriched for insertion- and repetition-related failure modes. Across Whisper-Base, Small, and Large, insertion ratio and RIR remained higher in both clinical groups than in the HC group, and their inverse associations with observed MMSE and CIST persisted after covariate adjustment (Multimedia Appendices 3 and 4). However, ERR-cognition associations and the AD/dementia-PD difference in RIR were not consistently reproduced.

Table 4. Covariate-adjusted partial correlations between automatic speech recognition (ASR) error metrics and cognitive scores.

Metric	MMSE[a], partial r[b]	MMSE FDR[c] (q value)[d]	CIST[e], partial r	CIST FDR (q value)	Hybrid CIST, partial r	Hybrid CIST FDR (q value)
CER[f]	–0.223	<.001[g]	–0.213	<.001[g]	–0.223	<.001[g]
Insertion ratio	–0.265	<.001[g]	–0.241	<.001[g]	–0.249	<.001[g]
Deletion ratio	–0.100	.04[h]	–0.023	.43	–0.045	.08
Substitution ratio	–0.075	.10	–0.056	.07	–0.080	.002[i]
RIR[j]	–0.260	<.001[g]	–0.191	<.001[g]	–0.216	<.001[g]
ERR[k]	–0.121	.01[h]	–0.111	<.001[g]	–0.119	<.001[g]

[a]MMSE: Mini-Mental State Examination.
[b]Partial Pearson correlations were adjusted for diagnostic group, age, sex, and years of education.
[c]FDR: false discovery rate.
[d]q values are Benjamini-Hochberg false discovery rate-adjusted P values, calculated separately within each cognitive measure across the 6 ASR error metrics.
[e]CIST: Cognitive Impairment Screening Test.
[f]CER: character error rate.
[g]q<.001.
[h]q<.05.
[i]q<.01.
[j]RIR: repeated-insertion ratio.
[k]ERR: excess repetition rate.

Figure 7. Associations between automatic speech recognition (ASR) error metrics and cognitive measures adjusted for diagnostic group, age, sex, and years of education. (A) The upper 6 panels display character error rate (CER) and insertion ratio, (B) lower 6 panels display repeated-insertion ratio (RIR) and excess repetition rate (ERR) against Mini-Mental State Examination (MMSE), observed Cognitive Impairment Screening Test (CIST), and Hybrid CIST. Points represent participants (healthy control [HC], blue; Alzheimer disease [AD]/dementia, orange; Parkinson disease [PD], green); black lines show linear fits to the residuals, and gray bands show 95% bootstrap CIs for the fitted mean. Annotations show partial Pearson *r* and Benjamini-Hochberg false discovery rate (BH-FDR) *q* values corrected across 6 ASR metrics within each cognitive measure.

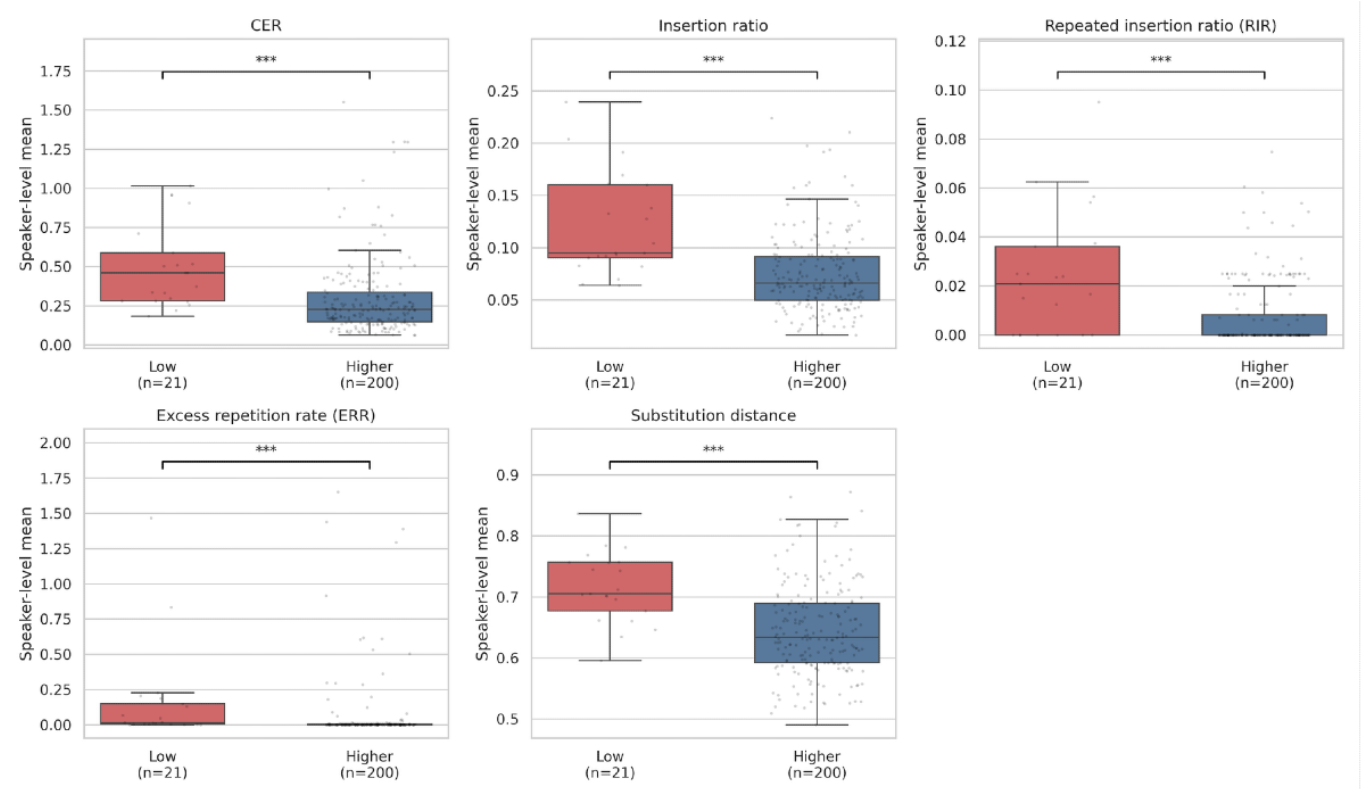


Exploratory Within-PD Analysis by MMSE Subgroup

As an exploratory within-PD analysis, we next compared participants with MMSE scores below 24 (n=21) and those with scores of 24 or higher (n=200). As shown in Figure 8, the lower-MMSE subgroup had higher CER (0.4968 vs 0.2950), insertion ratio (0.1214 vs 0.0748), RIR (0.0240 vs 0.0076), ERR (0.1623 vs 0.0600), and substitution distance (0.7156 vs 0.6448; all $q<.001$). Although this MMSE-based

split should not be interpreted as a formal PD dementia subgroup analysis, it suggests that insertion- and repetition-focused ASR abnormalities also track cognitive severity within the PD group.

Figure 8. Exploratory within-Parkinson disease (PD) automatic speech recognition (ASR) error differences by Mini-Mental State Examination (MMSE) subgroup. Participants with MMSE <24 (n=21) were compared with those with MMSE ≥24 (n=200). The 5 displayed metrics are character error rate (CER), insertion ratio, RIR (repeated-insertion ratio), ERR (excess repetition rate), and substitution distance. Brackets indicate 2-sided Mann-Whitney *U* comparisons with Benjamini-Hochberg false discovery rate (BH-FDR) correction across the 5 displayed metrics (*q<.05, **q<.01, ***q<.001). The MMSE threshold does not define a formal PD dementia diagnosis.



Discussion

Principal Results

This study characterized acoustic features and ASR error patterns across clinically diagnosed AD/dementia, PD, and HC groups. In the original unadjusted analysis, more acoustic features differed in comparisons involving the PD group than in the AD/dementia-HC comparison. However, the cohort-restricted adjusted analysis showed a broader AD/dementia-HC acoustic pattern. Both clinical groups showed elevated ASR error measures relative to the HC group, with insertion- and repetition-related measures associated with lower cognitive scores after adjustment for diagnostic group, age, sex, and years of education. The exploratory within-PD analysis also linked lower MMSE scores with higher ASR error measures. These results characterize group and cognitive associations.

The acoustic results are broadly consistent with prior literature showing that PD tends to produce more reproducible low-level acoustic abnormalities than AD [6,12,13,41-44]. Reviews and meta-analyses in PD have reported relatively consistent changes in prosodic and phonatory measures, including altered F0 variability, pause-related abnormalities, and broader voice perturbation patterns [12,13,45-48]. By contrast, recent work in AD suggests that acoustic markers are promising but more heterogeneous, with timing- and interruption-related measures among the more consistent findings [1,4-6,43,49]. In that context, our finding that the PD group showed broader multifeature acoustic separation whereas the AD/dementia group presented a narrower pattern dominated by speech duration is directionally consistent with existing evidence.

Our acoustic findings also fit this interpretation. Longer speech duration showed the strongest and most consistent association with lower MMSE, observed CIST, and Hybrid CIST scores. These associations remained after covariate adjustment, although their magnitudes were attenuated. This aligns with recent evidence that pause- and timing-related speech measures are among the most robust correlates of cognitive impairment in AD-spectrum conditions [50], and with cross-linguistic work showing that interpretable speech-timing features can predict MMSE better than many other feature classes [3,43,51]. Because our study used a fixed reading task rather than spontaneous narrative speech, the relatively modest AD/dementia-vs-HC acoustic separation is not unexpected [3,43,51]. Reading constrains lexical-semantic variability and likely attenuates some of the discourse-level abnormalities that are often more prominent in the AD/dementia group. Importantly, the present findings should

not be interpreted as directly indexing spontaneous speech or discourse-level language deficits in the AD/dementia group. Rather, our protocol captures recognizer-vulnerable speech patterns that emerge during a standardized reading task performed under neurocognitive burden. In this sense, the observed ASR error profiles likely reflect a composite manifestation of atypical speech production, reduced sequential stability, and task-constrained cognitive-linguistic inefficiency, rather than spontaneous language impairment.

The ASR analyses, however, revealed rationales that were not captured by conventional acoustics alone [24]. Across Whisper model sizes, absolute CER improved in all groups, but the relative disease-vs-HC gap widened rather than narrowed. This suggests that better-performing ASR models do not simply erase clinically relevant variation [24,49]. These results indicate persistent group-related differences in recognition performance within the Whisper model family but do not measure diagnostic sensitivity or establish robustness of the error-composition findings across recognizers.

The error-pattern analyses further suggested that examining ASR error composition may provide information not apparent from CER alone [24]. In the AD/dementia group, elevated insertion ratio and higher RIR and ERR were prominent findings relative to the HC group. In the unadjusted analysis, the PD group showed higher deletion ratios than both HC and AD/dementia groups, whereas the AD/dementia-PD difference in substitution distance was not significant after FDR correction. These findings are notable because CER did not significantly differ between the AD/dementia and PD groups, prompting closer examination of their ASR error profiles [23,43]. However, the cohort-restricted analysis adjusted for age, sex, and education showed a more qualified pattern: RIR was higher in the AD/dementia group than in the PD group, whereas insertion ratio, deletion ratio, substitution distance, and ERR did not significantly differ between the 2 groups after FDR correction. Prior work has suggested that ASR errors may carry useful dementia-related information rather than functioning only as nuisance noise [24]. Our findings extend this perspective by showing that insertion- and repetition-related ASR measures were associated with cognitive screening scores, with insertion ratio showing the largest absolute adjusted correlation among the 6 ASR metrics examined. RIR and ERR quantify repetition in ASR output, not independently verified repetition in the recorded speech, and may be influenced by recognizer hallucinations as well as speech and recording characteristics.

The exploratory within-PD comparison suggested that lower MMSE scores were associated with higher CER, insertion ratio, RIR, ERR, and substitution distance. This pattern suggests that cognition-related ASR vulnerability may also operate among participants with clinically diagnosed PD, not only across the AD-HC contrast [44]. These observations may help generate hypotheses about within-group heterogeneity, but the small lower-MMSE subgroup and the absence of formal cognitive diagnostic adjudication limit interpretation. They do not establish a PD dementia-specific ASR phenotype.

These findings also resonate with prior work suggesting that AD-related speech disruption may differ qualitatively from Parkinsonian speech impairment. For example, interpretable semantic markers have been reported to distinguish people with AD from HC participants but not people with PD, implying that AD may be better captured by markers of cognitive-linguistic instability than by motor-acoustic disturbance alone [15]. Our results extend this logic to ASR behavior. Even in a tightly controlled reading task that minimizes discourse variability, disease-related signal still emerged in how the recognizer failed, particularly through insertion- and repetition-related mechanisms. Overall, these results suggest that ASR error patterns may complement conventional acoustic biomarkers by capturing a different level of speech dysfunction, one that reflects the interaction between atypical speech production and recognizer failure modes.

In terms of future applications, these findings may inform the development of community-based cognitive screening approaches. Data collection included cognitive assessments at community sites, such as senior centers, and the predominance of CIST over MMSE reflects its use in this community-based assessment program. Future work will investigate whether ASR-derived measures from spontaneous speech can support early cognitive screening without requiring the fixed 40-item reading task. Such use would require prospective validation and would complement, rather than replace, clinical assessment. Prospective validation should assess false-positive rates and robustness across languages, recording conditions, and ASR model updates.

Limitations

Several limitations should be considered. First, this was a cross-sectional study, and therefore we could not determine whether ASR-derived error patterns track longitudinal cognitive decline, motor progression, or future diagnostic conversion. Second, the speech material consisted of a fixed 40-item Korean reading task. Although identical content improved between-participant comparability, the items were not prospectively matched for lexical frequency, syntactic complexity, or psycholinguistic difficulty. Item-level ASR performance may therefore partly reflect differences in sentence length and linguistic complexity. The fixed reading task also limits generalizability to spontaneous speech, other languages, and real-world recording environments. Recording-site differences between diagnostic groups and quantitative channel-quality effects were not evaluated in the present analysis. Third, diagnostic classification was based on specialist clinical judgment rather than biomarker-confirmed research criteria, such as ATN [30] profiling, amyloid positron emission tomography, or dopamine positron emission tomography. Accordingly, the AD/dementia label should be interpreted as clinically diagnosed dementia evaluated in a dementia-specialty setting rather than as a biomarker-defined AD cohort. This pragmatic design reflects the screening-oriented and large-scale nature of the speech collection, but it limits pathophysiologic specificity.

Fourth, diagnostic groups were unequal in size, reflecting the recruitment settings. Covariate-adjusted group analyses used a cognitively linked subset; differences from the full-cohort results may therefore reflect both sample selection and adjustment. Fifth, disease-specific severity measures, including Clinical Dementia Rating [31], Hoehn and Yahr stage [32], and Unified Parkinson's Disease Rating Scale scores [33], as well as functional status and literacy, were unavailable or incomplete. Residual confounding remains possible despite adjustment for age, sex, and education. Sixth, individual medication identity, dose, timing, and PD ON/OFF state were not systematically recorded. Although most patients were assessed while continuing their usual treatment, medication-related effects on speech could not be estimated [45]. Seventh, Hybrid CIST was a derived cognitive measure rather than a directly observed clinical score and should be interpreted as a sensitivity analysis cognitive measure rather than a substitute for direct cognitive assessment. Eighth, the within-PD MMSE subgroup analysis was exploratory and based on a screening-score threshold rather than formal cognitive diagnostic adjudication. Finally, the additional analyses support partial reproducibility within the Whisper model family, but validation using independent ASR systems remains necessary. Future prospective studies should validate these markers in biomarker-defined cohorts, spontaneous speech settings, and PD subgroups characterized by formal cognitive diagnosis, motor severity, and medication state.

Conclusions

In conclusion, conventional acoustic features captured PD-related speech abnormality more clearly than AD/dementia-related abnormality in this standardized reading dataset, whereas ASR error patterns revealed complementary disease-related structure that conventional acoustics did not fully capture. In the primary analysis, insertion ratio, RIR, and ERR were inversely associated with cognitive screening scores. These findings support further investigation of ASR error composition as a candidate marker of diagnostic-group differences and cognitive variation. Future studies incorporating PD-specific motor severity measures may further clarify motor-related contributions to ASR vulnerability in people with PD.

Acknowledgments

DYK and HL are co-corresponding authors of this work, and the latter can be reached by email at hyunjulee@gist.ac.kr.

Generative AI Use

ChatGPT (OpenAI) was used to assist with grammatical corrections and language editing during manuscript preparation.

Funding

This research was supported by Global - Learning & Academic research institution for Master's PhD students, and Postdocs (LAMP) Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (RS-2025-25440216), by InnoCORE program of the Ministry of Science and ICT (GIST InnoCORE KH0860), and by the Regional Innovation System & Education (RISE) program through the Jeonbuk RISE Center, funded by the Ministry of Education (MOE) and the Jeonbuk State, Republic of Korea (2026-RISE-13-WKU).

Data Availability

The dataset analyzed in this study is available through AI Hub [52] upon request and subject to the platform's data access procedures.

Authors' Contributions

JWK, HL, and DYK contributed to the conception and design of the study. DYK recruited study participants and administered the study assessments. DYK organized the dataset. JWK, SHY, HL, and DYK contributed to the analysis. DYK contributed to data interpretation. JWK drafted the manuscript. HL and DYK jointly supervised the study, contributed equally to the supervision of the work, and reviewed and edited the manuscript (co-corresponding authors). All authors had full access to the study design information and all data, approved the final version to be published, and agreed to be accountable for all aspects of the work to ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Korean speech materials used for controlled reading tasks and their English translations.
[DOCX File (Microsoft Word File), 30 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Age-, sex-, and education-adjusted automatic speech recognition error group comparisons.
[DOCX File (Microsoft Word File), 32 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Covariate-adjusted group differences in automatic speech recognition error metrics across Whisper-Base, Whisper-Small, and Whisper-Large.
[DOCX File (Microsoft Word File), 101 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Automatic speech recognition-cognition associations across Whisper-Base, Small, and Large.
[DOCX File (Microsoft Word File), 99 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Unadjusted comparisons of all 19 acoustic features.
[DOCX File (Microsoft Word File), 35 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Age-, sex-, and education-adjusted acoustic group comparisons in the cognitively linked cohort.
[DOCX File (Microsoft Word File), 36 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Covariate-adjusted group differences in acoustic features.
[DOCX File (Microsoft Word File), 171 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Covariate-adjusted associations between acoustic features and cognitive measures.
[DOCX File (Microsoft Word File), 170 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Utterance-level mean CERcharacter error rate across the 40 elicited utterances by diagnostic group.
[DOCX File (Microsoft Word File), 379 KB-Multimedia Appendix 9]

References

1. Saeedi S, Hetjens S, Grimm MOW, Barsties V Latoszek B. Acoustic speech analysis in Alzheimer's disease: a systematic review and meta-analysis. J Prev Alzheimers Dis. 2024;11(6):1789-1797. [doi: 10.14283/jpad.2024.132] [Medline: 39559890]
2. Huang L, Yang H, Che Y, Yang J. Automatic speech analysis for detecting cognitive decline of older adults. Front Public Health. 2024;12:1417966. [doi: 10.3389/fpubh.2024.1417966] [Medline: 39175901]
3. Pérez-Toro PA, Ferrante FJ, Pérez G, et al. Automated speech markers of Alzheimer dementia: test of cross-linguistic generalizability. J Med Internet Res. Oct 15, 2025;27:e74200. [doi: 10.2196/74200] [Medline: 41091545]
4. Ding H, Lister A, Karjadi C, et al. Detection of mild cognitive impairment from non-semantic, acoustic voice features: the Framingham Heart Study. JMIR Aging. Aug 22, 2024;7:e55126. [doi: 10.2196/55126] [Medline: 39173144]
5. Ding H, Mandapati A, Karjadi C, et al. Association between acoustic features and neuropsychological test performance in the Framingham Heart Study: observational study. J Med Internet Res. Dec 22, 2022;24(12):e42886. [doi: 10.2196/42886] [Medline: 36548029]
6. Pulido MLB, Hernández JBA, Ballester MÁF, González CMT, Mekyska J, Smékal Z. Alzheimer's disease and automatic speech analysis: a review. Expert Syst Appl. Jul 2020;150. [doi: 10.1016/j.eswa.2020.113213]
7. 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. Mar 2023;19(3):946-955. [doi: 10.1002/alz.12721] [Medline: 35796399]
8. Gosztolya G, Vincze V, Tóth L, Pákáski M, Kálmán J, Hoffmann I. Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. Comput Speech Lang. Jan 2019;53:181-197. [doi: 10.1016/j.csl.2018.07.007]
9. Sadeghian R, Schaffer JD, Zahorian SA. Towards an automatic speech-based diagnostic test for Alzheimer's disease. Front Comput Sci. 2021;3. [doi: 10.3389/fcomp.2021.624594]

10. Lim WS, Chiu SI, Peng PL, et al. A cross-language speech model for detection of Parkinson's disease. J Neural Transm (Vienna). Apr 2025;132(4):579-590. [doi: 10.1007/s00702-024-02874-z] [Medline: 39739129]
11. Thies T, Mallick E, Tröger J, Baykara E, Mücke D, Barbe MT. Automatic speech analysis combined with machine learning reliably predicts the motor state in people with Parkinson's disease. NPJ Parkinsons Dis. May 2, 2025;11(1):105. [doi: 10.1038/s41531-025-00959-4] [Medline: 40316531]
12. Cao F, Vogel AP, Gharahkhani P, Renteria ME. Speech and language biomarkers for Parkinson's disease prediction, early diagnosis and progression. NPJ Parkinsons Dis. Mar 24, 2025;11(1):57. [doi: 10.1038/s41531-025-00913-4] [Medline: 40128529]
13. Rusz J, Hlavnicka J, Tykalova T, et al. Smartphone allows capture of speech abnormalities associated with high risk of developing Parkinson's disease. IEEE Trans Neural Syst Rehabil Eng. Aug 2018;26(8):1495-1507. [doi: 10.1109/TNSRE.2018.2851787] [Medline: 29994713]
14. He R, Chapin K, Al-Tamimi J, et al. Automated classification of cognitive decline and probable Alzheimer's dementia across multiple speech and language domains. Am J Speech Lang Pathol. Sep 11, 2023;32(5):2075-2086. [doi: 10.1044/2023_AJSLP-22-00403] [Medline: 37486774]
15. Sanz C, Carrillo F, Slachevsky A, et al. Automated text-level semantic markers of Alzheimer's disease. Alzheimers Dement (Amst). 2022;14(1):e12276. [doi: 10.1002/dad2.12276] [Medline: 35059492]
16. Haulcy R, Glass J. Classifying Alzheimer's disease using audio and text-based representations of speech. Front Psychol. 2020;11:624137. [doi: 10.3389/fpsyg.2020.624137] [Medline: 33519651]
17. Kovac D, Novakova L, Mekyska J, et al. Digital speech biomarkers for assessing cognitive decline across neurodegenerative conditions. Comput Biol Med. Nov 2025;198(Pt B):111251. [doi: 10.1016/j.compbiomed.2025.111251] [Medline: 41175714]
18. Rusz J, Krack P, Tripoliti E. From prodromal stages to clinical trials: the promise of digital speech biomarkers in Parkinson's disease. Neurosci Biobehav Rev. Dec 2024;167:105922. [doi: 10.1016/j.neubiorev.2024.105922] [Medline: 39424108]
19. Miner AS, Haque A, Fries JA, et al. Assessing the accuracy of automatic speech recognition for psychotherapy. NPJ Digit Med. 2020;3:82. [doi: 10.1038/s41746-020-0285-8] [Medline: 32550644]
20. Just SA, Elvevåg B, Pandey S, et al. Moving beyond word error rate to evaluate automatic speech recognition in clinical samples: lessons from research into schizophrenia-spectrum disorders. Psychiatry Res. Oct 2025;352:116690. [doi: 10.1016/j.psychres.2025.116690] [Medline: 40849998]
21. Orozco-Arroyave JR, Hönig F, Arias-Londoño JD, et al. Automatic detection of Parkinson's disease in running speech spoken in three different languages. J Acoust Soc Am. Jan 2016;139(1):481-500. [doi: 10.1121/1.4939739] [Medline: 26827042]
22. Ahn T, Hong Y, Im Y, et al. Automatic speech recognition (ASR) for the diagnosis of pronunciation of speech sound disorders in Korean children. Clin Linguist Phon. Oct 2025;39(10):913-926. [doi: 10.1080/02699206.2024.2387609] [Medline: 39162064]
23. Liu YL, Li Y, Feng R, et al. Beyond manual transcripts: exploring the potential of automatic speech recognition errors in improving Alzheimer's disease detection. J Biomed Inform. Jan 2026;173:104968. [doi: 10.1016/j.jbi.2025.104968] [Medline: 41421688]
24. Li C, Xu W, Cohen T, Pakhomov S. Useful blunders: can automated speech recognition errors improve downstream dementia classification? J Biomed Inform. Feb 2024;150:104598. [doi: 10.1016/j.jbi.2024.104598] [Medline: 38253228]
25. Tang L, Zhang Z, Feng F, Yang LZ, Li H. Explainable Alzheimer's disease detection using linguistic features from automatic speech recognition. Dement Geriatr Cogn Disord. 2023;52(4):240-248. [doi: 10.1159/000531818] [Medline: 37433284]
26. Balagopalan A, Shkaruta K, Novikova J. Impact of ASR on Alzheimer's disease detection: all errors are equal, but deletions are more equal than others. Proc 6th Workshop Noisy User-generated Text (W-NUT 2020). 2020:159-164. [doi: 10.18653/v1/2020.wnut-1.21]
27. Radford A, Kim JW, Xu T, Brockman G, McLeavey C. Robust speech recognition via large-scale weak supervision. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. Proceedings of the 40th International Conference on Machine Learning. PMLR (Proceedings of Machine Learning Research); 2023:28492-28518. URL: https://proceedings.mlr.press/v202/radford23a/radford23a.pdf [Accessed 2026-11-12]
28. Clarke C, Patel S, Ives N, et al. UK parkinson's disease society brain bank diagnostic criteria. NIHR Journals Library; 2016. URL: https://www.ncbi.nlm.nih.gov/books/NBK379754/ [Accessed 2026-11-16]
29. Diagnostic and Statistical Manual of Mental Disorders. 5th ed. American Psychiatric Association; 2013. URL: https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596 [Accessed 2026-11-16]
30. Jack CR, Bennett DA, Blennow K, et al. NIA-AA research framework: toward a biological definition of Alzheimer's disease. Alzheimers Dement. Apr 2018;14(4):535-562. [doi: 10.1016/j.jalz.2018.02.018] [Medline: 29653606]

31. Hughes CP, Berg L, Danziger WL, Coben LA, Martin RL. A new clinical scale for the staging of dementia. Br J Psychiatry. Jun 1982;140(6):566-572. [doi: 10.1192/bjp.140.6.566] [Medline: 7104545]
32. Hoehn MM, Yahr MD. Parkinsonism: onset, progression and mortality. Neurology. May 1967;17(5):427-442. [doi: 10.1212/wnl.17.5.427] [Medline: 6067254]
33. Goetz CG, Tilley BC, Shaftman SR, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord. Nov 15, 2008;23(15):2129-2170. [doi: 10.1002/mds.22340] [Medline: 19025984]
34. de Cheveigné A, Kawahara H. YIN, a fundamental frequency estimator for speech and music. J Acoust Soc Am. Apr 2002;111(4):1917-1930. [doi: 10.1121/1.1458024] [Medline: 12002874]
35. Boersma P. Praat, a system for doing phonetics by computer. Glot Int. 2001;5(9/10):341-345. URL: https://www.fon.hum.uva.nl/paul/papers/speakUnspeakPraat_glot2001.pdf [Accessed 2026-11-12]
36. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Sov Phys Dokl. 1965;10(8):707-710. URL: https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=565 [Accessed 2026-11-12]
37. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist. Mar 1947;18(1):50-60. [doi: 10.1214/aoms/1177730491]
38. Cliff N. Dominance statistics: ordinal analyses to answer ordinal questions. Psychol Bull. 1993;114(3):494-509. [doi: 10.1037/0033-2909.114.3.494]
39. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. Jan 1, 1995;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
40. Kang KW, Park G, Kim H, et al. Improved dementia screening for elderly with low education in South Korea using the Cognitive Impairment Screening Test (CIST). Front Neurosci. 2025;19:1599019. [doi: 10.3389/fnins.2025.1599019] [Medline: 40438629]
41. Moya-Galé G, Walsh SJ, Goudarzi A. Automatic assessment of intelligibility in noise in Parkinson disease: validation study. J Med Internet Res. Oct 20, 2022;24(10):e40567. [doi: 10.2196/40567] [Medline: 36264608]
42. Lee J, Kim N, Ha JW, et al. Exploring voice acoustic features associated with cognitive status in Korean speakers: a preliminary machine learning study. Diagnostics (Basel). Dec 17, 2024;14(24):2837. [doi: 10.3390/diagnostics14242837] [Medline: 39767198]
43. García AM. Speech and language in healthy ageing and Alzheimer's dementia. Nat Rev Psychol. 2026;5(5):312-326. [doi: 10.1038/s44159-026-00553-2]
44. Scimeca S, Amato F, Olmo G, et al. Robust and language-independent acoustic features in Parkinson's disease. Front Neurol. 2023;14:1198058. [doi: 10.3389/fneur.2023.1198058] [Medline: 37384279]
45. Brabenec L, Mekyska J, Galaz Z, Rektorova I. Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation. J Neural Transm (Vienna). Mar 2017;124(3):303-334. [doi: 10.1007/s00702-017-1676-0] [Medline: 28101650]
46. Chiaramonte R, Bonfiglio M. Acoustic analysis of voice in Parkinson's disease: a systematic review of voice disability and meta-analysis of studies. Rev Neurol. Jun 1, 2020;70(11):393-405. [doi: 10.33588/rn.7011.2019414] [Medline: 32436206]
47. Fumel J, Bahuaud D, Weed E, Fusaroli R, Basirat A. A systematic review and Bayesian meta-analysis of acoustic measures of prosody in Parkinson's disease. J Speech Lang Hear Res. Aug 5, 2024;67(8):2548-2564. [doi: 10.1044/2024_JSLHR-23-00588] [Medline: 39018262]
48. Kong Q, Cao Y, Iqbal T, Wang Y, Wang W, Plumbley MD. PANNs: large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans Audio Speech Lang Process. 2020;28:2880-2894. [doi: 10.1109/TASLP.2020.3030497]
49. Shankar R, Goh Z, Devi F, Xu Q. A systematic review of explainable artificial intelligence methods for speech-based cognitive decline detection. NPJ Digit Med. Nov 26, 2025;8(1):724. [doi: 10.1038/s41746-025-02105-z] [Medline: 41298840]
50. Cohen AS, Divers R, Calamia M, et al. Speech pause and speech rate for evaluating Alzheimer's and mild cognitive impairment: a meta-analysis. J Int Neuropsychol Soc. Jan 2026;32(1):24-31. [doi: 10.1017/S1355617725101677] [Medline: 41378466]
51. De Looze C, Dehsarvi A, Crosby L, et al. Cognitive and structural correlates of conversational speech timing in mild cognitive impairment and mild-to-moderate Alzheimer's disease: relevance for early detection approaches. Front Aging Neurosci. 2021;13:637404. [doi: 10.3389/fnagi.2021.637404] [Medline: 33986656]
52. Voice information of older adults and patients with major neurodegenerative diseases in the Busan region [Article in Korean]. AI-Hub. URL: https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=565 [Accessed 2026-11-12]

Abbreviations

AD: Alzheimer disease
ASR: automatic speech recognition
ATN: amyloid, tau, neurodegeneration
BH-FDR: Benjamini-Hochberg false discovery rate
CER: character error rate
CIST: Cognitive Impairment Screening Test
***DSM-IV*:** *Diagnostic and Statistical Manual of Mental Disorders* (Fourth Edition)
ERR: excess repetition rate
FDR: false discovery rate
HC: healthy control
MMSE: Mini-Mental State Examination
PD: Parkinson disease
RIR: repeated-insertion ratio
RMS: root-mean-square
ZCR: zero-crossing

Edited by Megan O'Connell; peer-reviewed by Israel Martinez-Nicolas, Sudarsana Reddy Kadiri; submitted 28.Apr.2026; initial comments to author 19.Aug.2026; final revised version received 13.Oct.2026; accepted 20.Oct.2026; published 26.Nov.2026

Please cite as:
Kim JW, Yoon SH, Lee H, Kang DY
Automatic Speech Recognition Error Patterns for Differentiating Clinically Diagnosed Dementia, Parkinson Disease, and Healthy Aging: Cross-Sectional Association Study
JMIR Aging 2026;9:e99664
URL: https://aging.jmir.org/2026/1/e99664
doi: 10.2196/99664

© June-Woo Kim, Sung-Hoon Yoon, Hyunju Lee, Do-Young Kang. Originally published in JMIR Aging (https://aging.jmir.org), 26.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Aging, is properly cited. The complete bibliographic information, a link to the original publication on https://aging.jmir.org, as well as this copyright and license information must be included.