

Review

Voice, Speech, and Large Language Models in Neurology: From Acoustic Biomarkers to Conversational AI

Shahar Shelly ^{1,2} ¹ Department of Neurology, Rambam Health Care Campus, Haifa 3655306, Israel; shahar.shell@technion.ac.il² Ruth and Bruce Rappaport Faculty of Medicine, Technion—Israel Institute of Technology, Haifa 3655306, Israel

Abstract

Background: Speech models (wav2vec 2.0, HuBERT, Whisper), large language models (GPT, LLaMA), and conversational AI have expanded computational speech analysis from handcrafted acoustic features to dialogue-based neurological assessment. How well these approaches address clinical practice has not been evaluated. **Methods:** We conducted a narrative review searching PubMed, Google Scholar, and IEEE Xplore, supplemented by Interspeech and ICASSP proceedings. Findings are organized along three layers: acoustic-motor (voice quality, prosody, articulation), language-transcript (lexical, syntactic, semantic, and discourse analysis), and integrated multimodal-conversational (interactive dialogue systems). Traditional acoustic biomarkers provide background; the primary focus is on foundation models, LLMs, and conversational AI. **Findings:** Speech foundation models outperform handcrafted features on several classification tasks but degrade on severely impaired speech due to domain mismatch with healthy training data. LLMs classify transcripts and score cognitive tests, but operate on text alone and cannot access acoustic-motor information. Conversational AI can administer cognitive screening through naturalistic dialogue, but validation is limited to small single-centre feasibility studies. Prospective clinical validation remains limited. Cross-linguistic generalizability is untested for most methods. **Interpretation:** The field is moving toward integrated speech-language assessment, but the gap between technical capability and clinical utility remains wide. Closing it requires diverse multilingual datasets, standardized benchmarks, prospective validation, and ethical governance.

Keywords: speech biomarkers; voice analysis; large language models; foundation models; Parkinson's disease; Alzheimer's disease; dysarthria; automatic speech recognition; conversational AI; natural language processing; neurological assessment; digital biomarkers



Academic Editor: Xueyi Li

Received: 16 June 2026

Revised: 13 July 2026

Accepted: 14 July 2026

Published: 16 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Speech requires coordinated engagement of respiratory, laryngeal, and articulatory musculature under precise neural control, real-time lexical and syntactic processing, and continuous monitoring of acoustic output [1,2]. Because speech production depends on distributed circuits spanning the motor cortex, basal ganglia, cerebellum, perisylvian language network, and prefrontal systems, disruptions produce characteristic alterations in voice quality, articulation, prosody, fluency, and discourse coherence [3,4]. The hypophonia of Parkinson's disease, the scanning speech of cerebellar ataxia, the effortful output of Broca's aphasia, and the empty fluent speech of Alzheimer's disease are established clinical signs [5,6].

Over two decades, researchers have extracted acoustic features from speech recordings and shown they can discriminate patients from controls in controlled settings [7,8]. NLP applied to transcripts extended this to lexical, syntactic, and discourse-level features reflecting cognitive function [9,10]. These streams established that speech contains extractable diagnostic information, but the translational gap remains substantial: most studies rely on small cohorts, unstandardized features, and retrospective designs [11,12]. Previous reviews examined acoustic biomarkers for specific conditions [12,13] or digital speech methodology [11], but none have traced the trajectory from traditional features through foundation models to conversational AI.

Foundation models have begun to change this. Self-supervised speech models (wav2vec 2.0 [14], HuBERT [15], WavLM [16]) learn acoustic representations from unlabeled speech, replacing handcrafted features with learned representations that capture spectro-temporal patterns inaccessible to traditional measures. Whisper [17] enables reliable transcription across languages. Large language models (GPT [18], LLaMA [19], biomedical adaptations [20]) analyze transcripts at the levels of semantic coherence and clinical reasoning, capabilities beyond earlier NLP.

We organize the reviewed work along three conceptual layers: (1) the acoustic-motor layer, capturing voice quality, prosody, and articulation from audio; (2) the language-transcript layer, analyzing lexical, syntactic, semantic, and discourse features from transcripts; and (3) the integrated multimodal-conversational layer, combining acoustic and linguistic analysis within interactive dialogue systems. The convergence of these technologies has opened a new direction: conversational AI systems that conduct naturalistic clinical dialogues [21,22], though the gap between capability and clinical reality remains wide. Each approach carries characteristic limitations: domain mismatch for speech models [23], text-only access for LLMs [24], and absent clinical validation for conversational AI [25].

This review traces four stages: traditional acoustic biomarkers (as background), speech foundation models and ASR, LLM-based transcript analysis, and conversational AI for neurological assessment. For each, we summarize current evidence, identify limitations, and assess clinical readiness, with particular attention to multimodal integration, the gap between conversational fluency and clinical reasoning, cross-linguistic generalizability, and ethical considerations. Neurological conditions affect an estimated 3.4 billion people worldwide [26], and the global shortage of neurologists means scalable automated assessment tools could address a real unmet need. That estimate includes headache disorders; the burden of neurodegenerative and speech-affecting conditions is smaller but still substantial.

2. Methods

This is a narrative review examining voice and speech analysis, foundation models, and LLMs in neurological assessment. We adopted the narrative format because the topic spans multiple research traditions with fundamentally different methodologies and publication conventions [27,28].

We searched PubMed, Google Scholar, and IEEE Xplore for publications from January 2000 through April 2026, combining three concept domains: (1) neurological conditions; (2) speech and voice analysis; and (3) computational methods (deep learning, foundation models, LLMs, transformers, conversational AI). We supplemented with reference screening, forward citation tracking of seminal papers (wav2vec 2.0, HuBERT, Whisper, GPT-4), and hand-searching of Interspeech (2015 to 2025) and ICASSP (2018 to 2025) proceedings. Traditional acoustic biomarker studies (pre-2020) were included selectively as background; the primary focus was on foundation models, LLMs, and conversational AI from 2020 onward. The complete search strings, database-specific modifications, and applied filters

for each database are provided in Supplementary Table S5. A simplified PRISMA-style flow diagram summarizing this process is provided in Figure S1 in the Supplementary Material.

Studies were included if they met at least one of the following criteria: (a) applied speech foundation models or LLMs to neurological assessment; (b) presented benchmark datasets or evaluation frameworks for speech-based neurological analysis; (c) tested conversational AI for clinical cognitive or motor-speech assessment; or (d) addressed translational challenges (regulatory, ethical, clinical workflow) specific to speech-based clinical AI. Pre-2020 acoustic biomarker studies were included selectively as representative exemplars based on citation frequency, methodological influence on subsequent work, and coverage of the major neurological conditions addressed in the review; this section provides historical context rather than exhaustive coverage.

Studies were excluded if they focused on speech therapy outcomes without a diagnostic or monitoring component, addressed psychiatric conditions without neurological framing, reported purely technical speech-processing advances without clinical motivation, or were not available in English.

We organized findings along a conceptual spectrum (Figure 1): (1) traditional acoustic biomarkers (background); (2) speech foundation models and ASR; (3) LLMs and NLP for clinical speech; and (4) conversational AI and dialogue-based assessment.

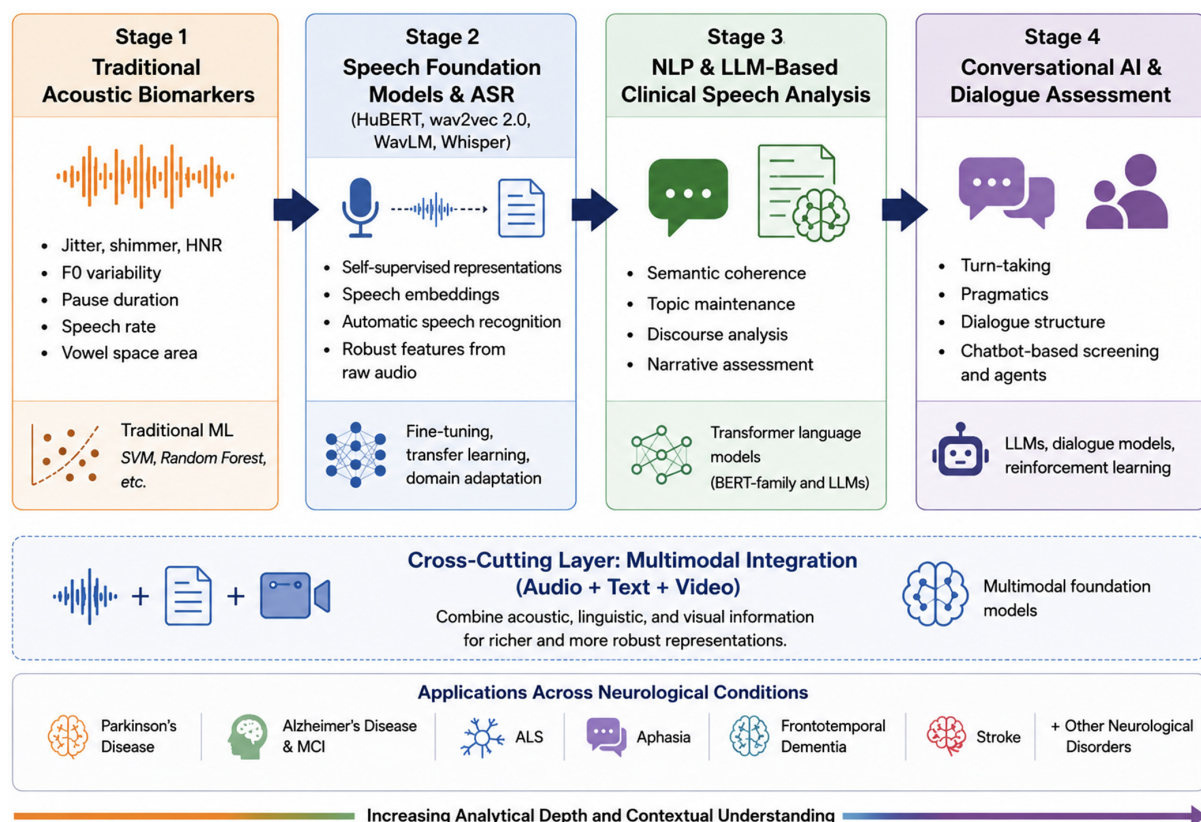


Figure 1. Conceptual framework for the four stages of computational speech analysis in neurology, corresponding to Sections 3–6 of this review. The stages progress from traditional acoustic biomarker extraction (Section 3, left) through speech foundation models and automatic speech recognition (Section 4), NLP- and LLM-based clinical speech analysis (Section 5), and conversational AI and dialogue-based assessment (Section 6, right), with increasing analytical depth and contextual understanding. A cross-cutting multimodal integration layer (Section 7) combines acoustic, linguistic, and visual information across stages to support integrated neurological assessment. This figure was generated using the AI tool Google's Gemini Pro 3 based on author-defined prompts; all values were verified against the original data tables.

The initial database search across PubMed, Google Scholar, and IEEE Xplore yielded approximately 1600 records. After removing duplicates and screening titles and abstracts, approximately 340 articles underwent full-text review. Of these, 128 studies met our inclusion criteria and are cited in the final manuscript; 15 of the 128 were identified through forward citation tracking and hand-searching of conference proceedings rather than the primary database search. A simplified PRISMA-style flow diagram summarizing this process is provided in Figure S1.

We did not apply a formal risk-of-bias tool (e.g., QUADAS-2). Instead, methodological rigour (sample size, validation strategy, dataset limitations) is evaluated for individual studies throughout the text. Study selection was performed by the author with iterative refinement during manuscript preparation. This approach, while standard for narrative reviews, does not carry the reproducibility guarantees of a systematic review, and this limitation should be considered when interpreting the scope and completeness of the evidence presented.

AI Tools Used

Figures 1 and 2 were generated using Google's Gemini image generation model based on author-defined prompts. The underlying numerical content was prepared in Excel tables, and all values shown in the figures were verified against the original tables to ensure accuracy.

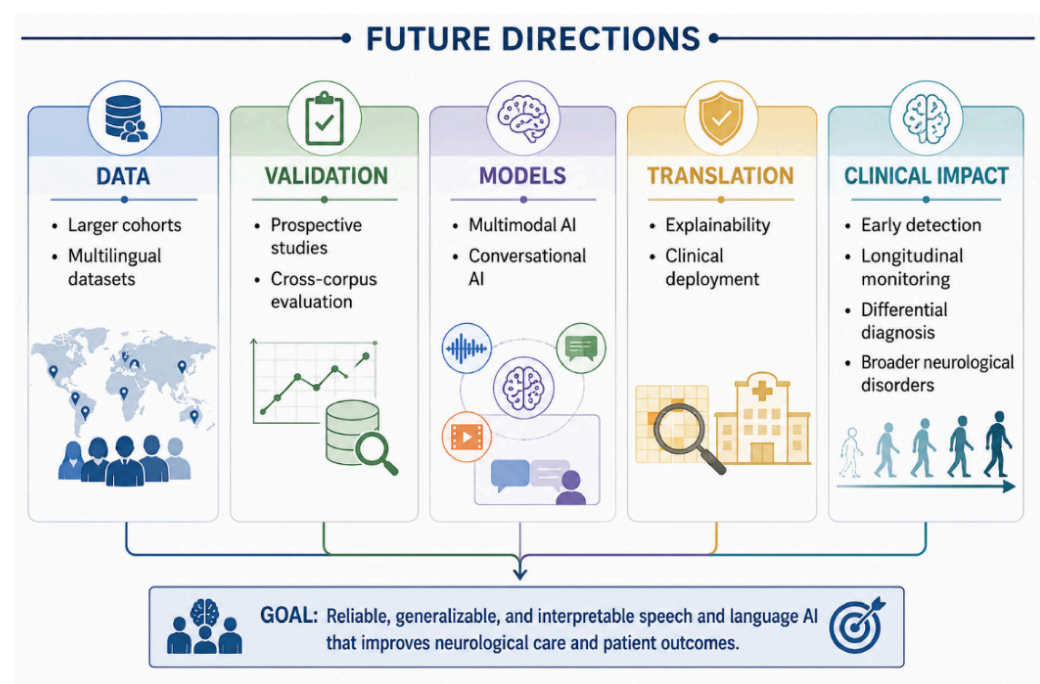


Figure 2. Key future directions for speech and language AI in neurology, derived from the limitations identified in Section 8. The figure summarizes major challenges and opportunities across five domains: data acquisition (Sections 8.1 and 8.2), validation (Section 8.1), model development (Sections 4, 5 and 7), clinical translation (Sections 8.5 and 8.6), and clinical impact (Sections 8.3 and 8.6). Future efforts should focus on larger and multilingual datasets, prospective and cross-corpus validation, multimodal and conversational AI, explainable and deployable systems, and clinically meaningful applications, including early detection, longitudinal monitoring, differential diagnosis, and broader neurological disease coverage. This figure was generated using the AI tool Google's Gemini Pro 3 based on author-defined prompts; all values were verified against the original data tables.

3. Traditional Acoustic Biomarkers: Background

The computational analysis of speech in neurological disorders has a two-decade history rooted in the extraction of handcrafted acoustic features from controlled recordings. This body of work established that the speech signal contains diagnostically relevant information, but systemic limitations have constrained clinical translation. We summarize the key findings and limitations here as context for the foundation model and LLM-based approaches that follow.

Parkinson's disease (PD) has received the most attention, reflecting the high prevalence of hypokinetic dysarthria [5,29,30]. Early studies extracted jitter, shimmer, and harmonic-to-noise ratio from sustained phonation, reporting high classification accuracies [7,31], though some used sample-level rather than subject-level cross-validation. Later work on connected speech showed articulatory rate, pause patterns, and prosodic variability were more sensitive to early disease [3,32,33], with cross-linguistic analyses showing substantial influence of language and recording conditions [34].

In Alzheimer's disease (AD), the changes are cognitive-linguistic: word-finding pauses, reduced lexical diversity, simplified syntax, and loss of discourse coherence [9,35,36]. Linguistic transcript analysis proved more discriminating than acoustic features alone, with Fraser and colleagues achieving over 81% accuracy on the DementiaBank corpus [9]. The ADReSS/ADReSSo challenges established standardized benchmarks [37,38]. Studies of ALS showed that speaking rate is sensitive to early bulbar involvement and can track progression over months [39,40]. Acoustic studies also extended to aphasia [10], autism spectrum disorder [41,42], multiple sclerosis [43,44], Huntington's disease [45], and traumatic brain injury [46].

Despite these achievements, several systemic limitations have prevented clinical translation. First, handcrafted feature sets are not standardized across laboratories, impeding cross-study comparison [11,13]. Second, the near-universal reliance on small, single-centre datasets (typically fewer than 250 participants) recorded under controlled conditions limits generalizability to real clinical environments [12]. Third, the predominant framework (binary classification of patients versus controls) does not address clinically relevant tasks of disease staging, progression monitoring, or treatment response. Fourth, acoustic and linguistic dimensions have typically been analyzed in separate pipelines despite neurological disorders often affecting both simultaneously. No prospective clinical validation study has been published for any traditional approach. These limitations motivated the foundation model and LLM-based approaches described in the following sections.

4. Speech Foundation Models and Automatic Speech Recognition

4.1. From Handcrafted Features to Learned Representations

Self-supervised learning (SSL) models, trained on large corpora of unlabeled speech, learn general-purpose representations capturing complex spectro-temporal patterns without prior specification of which features to extract [13,47,48]. Neurological speech disorders produce subtle, multidimensional alterations that may be partially captured by individual features but are more completely represented in high-dimensional learned embeddings. The key question is whether representations learned from healthy speech transfer effectively to pathological populations.

4.2. wav2vec 2.0, HuBERT, and WavLM

wav2vec 2.0 [14] learns representations through contrastive self-supervised learning on masked raw audio, producing contextualized representations encoding both local spectral and longer-range temporal dependencies. HuBERT [15] employs offline clustering for masked prediction with more stable layer-wise representations. WavLM [16] incorporates

denoising objectives, producing representations more resistant to environmental noise and speaker variation, a property relevant for clinical recordings outside controlled settings.

Several studies have applied SSL models to neurological speech. Bayerl and colleagues showed that SSL representations outperformed acoustic baselines for dysfluency detection [49]. Javanmardi and colleagues reported improvements for dysarthria severity classification [50]. Balagopalan and colleagues found that combining speech representations with linguistic features improved AD detection on ADReSS [51]. Pappagari and colleagues showed competitive performance even with limited labelled data [52]. Favaro and colleagues found that SSL embeddings outperformed handcrafted features across six languages for PD detection [53].

A critical limitation is the domain mismatch between pretraining data (predominantly healthy adult speech from podcasts, audiobooks, and web media) and the target population (patients with dysarthric, aphasic, or cognitively impaired speech). Violeta and colleagues investigated SSL pretraining frameworks for pathological speech recognition and found significant performance degradation on severely disordered speech compared with mild or moderate impairment, suggesting that standard SSL models may require domain adaptation or clinical pretraining to handle the full spectrum of neurological speech disorders [23]. This echoes a broader concern in clinical AI: general-purpose foundation models may not reliably generalize to the distributional shifts encountered in medical populations without explicit adaptation.

4.3. *Whisper and Clinical ASR*

OpenAI's Whisper [17], trained on 680,000 h of multilingual weakly supervised data, achieves reliable transcription across languages, accents, and acoustic environments for typical speech. For clinical speech research, Whisper is primarily an enabling technology: automatic transcription removes the bottleneck that has limited the scale of NLP-based speech analysis studies.

However, Whisper's accuracy on neurological speech is not assured. Rowe and colleagues reported substantially higher word error rates for severe dysarthria compared with mild impairment or healthy speech [54], consistent with the broader ASR literature showing degraded performance on atypical speech. The same distributional mismatch affecting SSL models also affects ASR: models trained predominantly on healthy speech underperform on the populations most in need of automated analysis. Efforts to adapt ASR for pathological speech include fine-tuning on clinical corpora and data augmentation strategies that simulate dysarthric characteristics [55]. The TORGO database [56] and UA-Speech corpus [57] have served as benchmarks, though both are limited in size and linguistic diversity.

4.4. *Transfer Learning*

The practical value of speech foundation models depends on transfer learning effectiveness. Fine-tuning SSL models on small clinical datasets ($n = 50$ to 200) can yield performance that is competitive with or superior to traditional pipelines, particularly for subtle distinctions (MCI versus healthy ageing, early PD versus controls) that are poorly captured by handcrafted features [49,51,58].

Transfer learning introduces its own challenges. The high dimensionality of SSL representations (typically 768 or 1024 dimensions per time step) relative to small clinical sample sizes creates substantial overfitting risk, particularly when fine-tuning the full model rather than extracting fixed features from pretrained layers. The optimal SSL layer for feature extraction is itself a consequential hyperparameter: earlier layers encode acoustic-phonetic information while later layers capture more abstract linguistic representations,

and the optimal layer depends on whether the condition primarily affects motor speech or language processing [59]. These considerations point to the need for rigorous validation (held-out test sets, cross-corpus evaluation, prospective validation) that the field has not yet consistently adopted.

5. Large Language Models and NLP for Clinical Speech Analysis

5.1. Transformer-Based NLP for Speech Transcripts

The application of NLP to speech transcripts in neurological populations predates the current era of large language models. Early work relied on rule-based analysis, bag-of-words models, and hand-engineered features such as type-token ratio and mean length of utterance [9,60]. The introduction of transformer-based language models, BERT [61] and biomedical variants (BioBERT [62], ClinicalBERT [63]), enabled contextualized semantic representations capturing meaning at a granularity inaccessible to earlier methods. Balagopalan and colleagues showed that BERT representations combined with acoustic features achieved state-of-the-art AD detection on the ADReSS benchmark [51,64]. Yuan and colleagues found BERT-derived pausing patterns predicted cognitive status independent of lexical content [65]. Agbavor and Liang showed that transformers could predict MMSE scores from transcripts with accuracy approaching clinical utility [66].

The transition from feature-extraction NLP to generative language models represents a qualitative shift. BERT-family models excel at classification and regression but cannot generate interpretive clinical text, engage in dialogue, or reason about the relationship between language patterns and neurological mechanisms. These capabilities became possible with large generative models.

5.2. GPT and LLM-Based Analysis

The GPT family [18] introduced open-ended generation and in-context learning, enabling qualitatively new applications. Rather than classifying a transcript as “impaired” or “normal,” an LLM can describe specific linguistic features, generate differential considerations, and produce structured reports.

Agbavor and Liang investigated whether GPT-3 could detect cognitive impairment from speech transcripts, finding that GPT-3 embeddings, used as input to a downstream classifier, distinguished AD from controls with approximately 80% accuracy without fine-tuning the language model itself [66]. B T B and colleagues evaluated ChatGPT-3.5, ChatGPT-4, and Google Bard in zero-shot classification of AD versus cognitively normal individuals, finding all three surpassed chance-level detection but did not reach clinical-application thresholds [67]. Taherinezhad and colleagues systematically compared LLM adaptation strategies, including in-context learning, reasoning-augmented prompting, and fine-tuning, with fine-tuned LLaMA 3B achieving an F1 score of 0.83 and AUC of 0.91 on ADReSSo [68]. Chlasta and colleagues combined GPT-4o-derived linguistic features with HuBERT speech representations for dementia classification, achieving RMSE of 2.78 for cognitive score regression [69]. In the aphasia domain, Kurland and colleagues compared three LLMs to human raters for main concept scoring of story retelling discourse from 96 persons with aphasia, finding strong correlations supporting clinical automation [70]. Fei and colleagues compared GPT-3.5 and GPT-4 to physician-led cognitive evaluations in 60 healthy individuals and 30 stroke survivors, finding that GPT-4 scores aligned closely with clinician ratings [71].

These studies show the flexibility of LLM-based analysis but also reveal fundamental limitations. Five of six relied on transcripts alone as input; only Chlasta and colleagues incorporated audio-derived SSL embeddings. LLMs operate on text and therefore cannot access the acoustic-motor information (voice quality, prosody, speech rate, pause durations)

that constitutes a critical dimension of neurological speech assessment. The LLM does not hear the patient; it reads a transcript. The distinction matters. Moreover, ASR errors on dysarthric or aphasic speech propagate through the LLM analysis and may produce spurious findings.

Clinical reliability of LLM-based speech analysis depends on factors beyond classification accuracy. Model architecture, pretraining corpus, instruction tuning, and alignment strategies all influence how an LLM interprets clinical transcripts. Models pretrained on biomedical text may handle clinical terminology more accurately than general-purpose models, but instruction-tuned models may generate more structured outputs. Alignment methods (reinforcement learning from human feedback, constitutional AI) shape whether a model hedges appropriately or produces overconfident clinical inferences. These architectural differences mean that results obtained with one LLM family may not generalize to another, and clinical evaluation must be model-specific.

Retrieval-augmented generation (RAG) offers a potential approach to grounding LLM outputs in external knowledge sources such as clinical guidelines, validated scoring rubrics, or domain-specific literature. By retrieving relevant documents before generating a response, RAG architectures could reduce hallucination and improve the factual accuracy of automated clinical assessments. In neurological speech analysis, a RAG-enabled system could cross-reference a generated interpretation against published diagnostic criteria or normative data for a given speech task. RAG has shown promise in other healthcare applications but remains largely unexplored in speech-based neurological assessment.

5.3. Discourse Coherence Analysis

One area where NLP and LLM methods are well-suited is the analysis of discourse coherence, the logical and thematic connectedness of extended speech, which deteriorates diagnostically in AD, frontotemporal dementia, and thought-disordered states [72,73]. Computational approaches have applied semantic similarity metrics, topic modelling, and graph-based coherence analysis. Embedding-based coherence measures applied to picture descriptions showed that semantic coherence declined in a gradient from healthy controls through MCI to AD [74]. Mota and colleagues showed graph-based speech connectivity could distinguish psychotic from nonpsychotic states [75]. Ilias and Askounis proposed multimodal architectures combining BERT with vision transformers for AD detection [76].

LLMs extend this by enabling assessment of pragmatic appropriateness, inferential coherence, and referential tracking [77], dimensions inaccessible to counting-based metrics, though this remains untested at scale in clinical populations.

5.4. Automated Test Scoring

A practical application of NLP and LLM technologies is automated scoring of standardized cognitive assessments that rely on verbal responses, including the Boston Naming Test, verbal fluency tasks, and story recall [78,79]. Automating this process could reduce administration time, eliminate inter-rater variability, and enable cognitive screening at scale. Toth and colleagues showed automated scoring of verbal fluency tests approaching human rater accuracy [80]. Lehr and colleagues predicted memory scores and detected MCI from story recall transcripts [81]. Fristed and colleagues deployed a remote smartphone-based AI system for daily story recall, achieving AUC 0.85 for MCI/AD detection and 0.78 for amyloid positivity prediction in the MCI/mild AD subgroup [82]; this is also an early example of home-based monitoring (Section 7.2). Automated test scoring is arguably the nearest-term clinical application, as it operates within a structured framework with established ground truth and does not require diagnostic reasoning *de novo*.

6. Conversational AI and Dialogue-Based Assessment

6.1. From Monologic to Dialogic Assessment

Traditional computational speech analysis relies on monologic samples: sustained phonation, reading passages, picture description, or verbal fluency tasks [11,37]. These are standardized and reproducible but capture only a narrow slice of communicative function. Clinical neurological assessment relies heavily on conversational interaction: the neurologist observes not only what the patient says but how they respond to questions, maintain conversational flow, repair misunderstandings, manage topic transitions, and adapt communication to the demands of dialogue [83,84]. Conversational AI capable of sustaining naturalistic dialogue creates the possibility of extending speech analysis from isolated task performance to interactive conversational assessment.

6.2. Chatbot-Based Cognitive Screening

Mirheidari and colleagues created an automated conversational system that administered questions from standard cognitive assessments within a naturalistic dialogue framework, analyzing both speech content and acoustic characteristics [85,86]. Their system achieved moderate accuracy in distinguishing dementia from healthy controls and from functional memory complaints, with acoustic and linguistic features contributing complementary information. Patients reported high acceptability, suggesting chatbot-based assessment may overcome barriers of clinician time and patient anxiety [86]. Tanaka and colleagues developed a computer-avatar agent for dementia screening that engaged elderly participants in unstructured dialogue and analyzed multimodal features (facial expression, eye gaze, head movement, acoustic features, linguistic content), outperforming unimodal baselines [87]. Takeshige-Amano and colleagues extended this approach using a chatbot to elicit speech and facial expressions from 99 controls and 93 individuals with AD or MCI, achieving an AUC of 0.94 with combined audiovisual features in a single-centre study requiring independent replication [88].

More recently, telephone-based and LLM-powered conversational agents have emerged. Serafimovska and colleagues showed that the TICS-M cognitive test could be autonomously administered by telephone with validity comparable to a psychologist's assessment [89]. König and colleagues validated automated telephone-based semantic verbal fluency testing in the PROSPECT-AD study, finding high agreement with manual gold-standard scoring [90]. Pakhomov and colleagues developed ANNA (Automated Neural Nursing Assistant), an LLM-based agent conducting brief cognitive assessments by telephone multiple times daily to detect neurotoxic symptoms in immunotherapy patients [91]. Ter Huurne and colleagues evaluated the user experience of automated phone-based cognitive testing in 67 memory clinic participants, finding satisfaction comparable to in-person assessment [92].

This body of work is growing but remains early-stage. The gap between conversational fluency and clinical reasoning is a fundamental concern: the system's ability to conduct an engaging conversation does not imply that its diagnostic inferences are reliable. Clinical validation of these systems has been limited to small, single-centre feasibility studies, and no system has shown diagnostic accuracy sufficient for clinical deployment.

6.3. Dialogue Structure Analysis

The structure of dialogue itself provides diagnostic information. Conversational analysis examines turn-taking, repair sequences, and response timing [93]. AD patients show increased self-repair, topic drift, and failure to respond to adjacency pairs; PD patients exhibit reduced turn-taking initiative and longer response latencies; right hemisphere damage produces tangential responses and failure to maintain conversational coherence [94,95].

Else and colleagues found that turn-taking patterns and repair sequences in routine medical consultations could detect cognitive impairment not captured by standard screening instruments [96]. Jones and colleagues showed that automated measures of turn-taking disruption correlated with clinical severity [97]. These studies suggest that the dialogue itself, not just the speech it contains, is a source of clinically actionable information.

6.4. Interactive Voice Assistants for Monitoring

The ubiquity of voice-enabled devices has prompted interest in their potential for ongoing neurological monitoring. Robin and colleagues showed digital speech assessments yield scores sensitive to MCI and early AD [98]. Daily smartphone monitoring in PD yielded acoustic measures with sufficient reliability to detect changes over weeks to months [11]. The appeal lies in ecological validity and scalability, though environmental noise, device variability, and user compliance introduce variance that may obscure the clinical signal [99].

6.5. Toward End-to-End Assessment

The convergence of speech foundation models, LLMs, and multimodal integration approaches the point where naturalistic clinical conversations, administered by AI, analyzed across acoustic and linguistic dimensions, and summarized as structured assessments, become technically feasible. Individual components exist: multimodal fusion [100], automated test administration [89,90], and LLM-based scoring [70]. No system yet combines them end-to-end. The distance between technical feasibility and clinical utility (discussed in Section 8) remains the field's central challenge.

7. Multimodal Integration and Longitudinal Monitoring

7.1. Multimodal Fusion

The distinction between motor speech analysis and language analysis has organized this review, but clinical reality is that motor and cognitive dimensions are frequently co-affected: PD patients exhibit both dysarthria and executive-mediated discourse deficits; frontotemporal dementia may affect both motor speech and language; stroke patients may have concurrent dysarthria and aphasia [101,102]. Multimodal approaches integrating acoustic, linguistic, and visual features represent the most complete analytical framework. Combining acoustic and linguistic features improved AD detection beyond either modality alone [37]. Rohanian and colleagues implemented a gating-based fusion architecture that learned to weight features adaptively, achieving state-of-the-art ADReSS performance [100]. Tanaka and colleagues showed that visual features (facial expression, eye gaze, head movement) provided additional discriminative information for dementia detection, particularly for patients whose speech remained superficially intact [87].

Multimodal foundation models (GPT-4o [103], Gemini [104]) offer unified processing of speech, text, and visual inputs without separate feature extraction pipelines, but have not been evaluated for neurological speech assessment.

Evaluating multimodal systems requires attention to whether each modality contributes independent clinical value or whether performance gains reflect redundancy. Ablation studies removing one modality at a time are needed to determine marginal contributions. Practical deployment must also address missing modalities: a system trained on audio, text, and video must degrade gracefully when video is unavailable or when ASR quality is poor due to severe dysarthria. Few studies have reported systematic ablation or robustness testing across modality-dropout conditions.

A growing class of native audio-language models can process speech input directly without requiring an intermediate ASR step. Models such as GPT-4o (OpenAI), Gemini (Google), SALMONN, and Qwen-Audio accept audio tokens alongside text, preserving

paralinguistic information (prosody, voice quality, timing, speech rate) that is discarded in text-only pipelines. For neurological assessment, where acoustic-motor features carry diagnostic weight independent of linguistic content, this architectural difference matters: a pipeline that first transcribes and then analyzes text cannot recover the breathy voice quality of early vocal fold paresis, the irregular articulatory breakdowns of cerebellar dysarthria, or the monotone prosody of Parkinson's disease.

These end-to-end models could reduce error propagation (ASR errors compound downstream NLP errors) and detect cross-modal patterns invisible to modular systems. The limitations are equally clear: clinical validation is absent, interpretability is reduced compared to modular pipelines where each component can be evaluated independently, computational requirements are substantial, and general-purpose pretraining on healthy speech may amplify domain mismatch for neurologically impaired speakers. Whether these models represent a practical advance over carefully engineered pipeline approaches or introduce new failure modes specific to disordered speech remains an open empirical question requiring targeted evaluation in neurological populations.

7.2. Home-Based and Remote Monitoring

Most neurodegenerative diseases are progressive, with changes unfolding over weeks to years, a temporal scale poorly captured by clinic visits every 3 to 12 months [105]. Home-based monitoring using consumer devices enables continuous tracking in the patient's natural environment. Tsanas and colleagues developed a telemonitoring system in which PD patients recorded phonation samples at home using a telephone, achieving remote estimation of UPDRS motor scores [106]. The Roche smartphone study showed passively collected speech features correlated with medication cycles and disease progression in a Phase 1 PD trial [107]. Connaghan and colleagues used a smartphone application for frequent remote speech sampling in ALS, showing automated acoustic measures tracked bulbar decline [108].

The analytical challenges of longitudinal monitoring differ fundamentally from cross-sectional classification. The relevant signal is the rate and pattern of change over time, requiring accounting for within-subject variability, diurnal fluctuations, medication effects, and disease natural history. Mixed-effects models and change-point detection are appropriate frameworks, but require substantially larger longitudinal datasets than currently exist [109].

7.3. Digital Phenotyping

Digital phenotyping extends speech analysis to data captured during routine device use (phone calls, voice commands) rather than to specific assessment tasks [110,111]. Fagherazzi and colleagues proposed a framework for voice as a broad biomarker of health [112]. The ethical implications (consent in cognitively impaired populations, privacy, algorithmic bias) are discussed in Section 8.4 [113,114].

7.4. Toward Integrated Neurological Speech Agents

The technological components reviewed in the preceding sections point toward a future direction: integrated AI agents capable of conducting a full neurological speech assessment through naturalistic interaction, combining real-time acoustic analysis, reliable transcription, linguistic and discourse analysis, adaptive conversational management, and clinical synthesis into a single pipeline [21,115]. No published system achieves this integrated functionality, and the individual components are at varying stages of maturity. We include this projection not as a review of existing systems but to frame the research agenda discussed in Section 8.

8. Discussion

8.1. Methodological Gaps

Despite technological progress, limitations persist across all stages (Figure 2). Clinical speech datasets remain small (DementiaBank ~300 participants [116]; most PD studies fewer than 100 [117]; ALS cohorts 20 to 65 speakers [40,118]). Standardized benchmarks exist only for AD (ADReSS [37,38]). The field remains overwhelmingly retrospective. Prospective clinical validation (deploying a system in a clinical workflow and measuring impact on diagnosis, monitoring, or treatment) is virtually absent. This is the most important barrier to translation.

Overfitting is a pervasive concern. Many studies, particularly those predating the ADReSS challenges, relied on k-fold cross-validation or leave-one-out designs without held-out test sets, creating a risk of inflated performance estimates. The ADReSS challenge [37,38] represents a positive exception: independent training and test partitions enabled more realistic evaluation. Studies using leave-one-out validation on cohorts of fewer than 50 participants should be interpreted with particular caution, as even modest data leakage through shared recording conditions or feature selection can inflate accuracy by 10 to 20 percentage points.

External validation, testing a model on an independent cohort from a different clinical site, recording environment, or demographic population, is nearly absent. Only a handful of studies evaluated models across sites, and none validated a speech-based AI system prospectively. Dataset biases compound this problem: most corpora are demographically homogeneous (predominantly white, educated, English-speaking participants), and models trained on read speech or sustained vowels may not generalize to spontaneous conversational speech. Recording conditions (clinical microphones versus smartphone audio versus telephone) introduce further variability rarely controlled for across studies.

A distinction must be drawn between statistical performance and clinical significance. High classification accuracy (e.g., AUC above 0.90) in a balanced research dataset does not imply clinical readiness. In clinical populations, where disease prevalence, cost of false positives and false negatives, and workflow integration determine real-world utility, even models with strong AUCs may produce unacceptable positive predictive values. No study in this review reported results according to the CLAIM checklist for medical imaging AI or the STARD diagnostic accuracy reporting guidelines, and few discussed clinical significance beyond classification metrics.

8.2. Cross-Linguistic Generalizability

The reviewed literature is dominated by English-speaking populations, limiting generalizability to the majority of the world's neurological patients. Speech biomarkers are not language-neutral: tonal languages (Mandarin, Cantonese) use pitch contours for lexical meaning, potentially confounding fundamental frequency measures used for PD assessment [119]; agglutinative languages (Turkish, Finnish) produce different pause and articulation patterns, affecting temporal measures for cognitive screening [120]. Linguistic features are even more language-dependent: lexical diversity, syntactic complexity, and discourse coherence scoring all require language-specific NLP tools and normative data. Only two studies in this review addressed cross-linguistic validity directly [34,53]. The DementiaBank corpus, ADReSS challenges, and most conversational AI studies are English-only. Cross-cultural pragmatic norms (turn-taking, directness, use of silence) further complicate conversational AI assessment. A system calibrated to Anglo-American norms could misinterpret culturally normative reticence as cognitive impairment. Until biomarkers are validated across typologically diverse languages, claims of clinical generalizability are not warranted.

8.3. The Fluency-Reasoning Gap

Current LLMs generate fluent, contextually appropriate text that can create an impression of clinical competence. An LLM analyzing a transcript can identify word-finding pauses, reduced lexical diversity, and simplified syntax, and generate an assessment that reads like a clinical note. But the system does not hear the voice, does not observe the patient, does not have access to medical history, and does not possess the clinical training that enables a neurologist to integrate speech observations with motor examination, cognitive testing, and neuroimaging [24,121]. Describing features of dysarthria from a transcript is pattern description; integrating those features into a diagnostic formulation is clinical reasoning. The risk of conflating these capabilities is not theoretical: premature deployment could produce false reassurance or unnecessary anxiety [122,123]. The field must maintain a clear distinction between the capacity to analyze speech data and the capacity to make clinical decisions.

Beyond the fluency-reasoning distinction, LLMs pose specific risks in clinical speech analysis. Hallucination, the generation of plausible but factually incorrect content, could produce fabricated clinical observations or unsupported diagnostic inferences from a speech transcript. Overconfident explanations may present uncertain findings with unwarranted certainty, misleading clinicians who lack the technical background to evaluate the model output. Mitigation strategies under investigation in other clinical AI domains include retrieval-augmented generation to ground outputs in validated sources structured prompting to constrain the response format, uncertainty estimation to flag low-confidence outputs, and clinician-in-the-loop review to ensure that no automated assessment reaches the patient or medical record without expert oversight. These strategies remain largely untested in speech-based neurological assessment.

8.4. Ethical Considerations

Speech recordings are biometric data that can identify speakers, reveal emotional states, and encode demographic characteristics (age, sex, accent, socioeconomic background) beyond the targeted clinical information [113,114]. In neurological conditions that may impair decision-making capacity, data collection requires particular attention to informed consent and data protection [124]. Algorithmic bias is a specific concern: speech models trained on healthy adult English speech degrade on accented, aged, and neurologically impaired speech, potentially disadvantaging the populations most in need [23,54,125]. The privacy implications of passive speech monitoring in homes of patients with dementia raise issues of autonomy and surveillance beyond standard data protection [126]. A related concern is interpretability: SSL embeddings and LLM-generated assessments function as black boxes, and without mechanistic transparency, clinicians cannot judge whether an assessment reflects a genuine clinical pattern or a spurious artifact. Governance must precede deployment.

8.5. Future Directions

Foundational infrastructure: large, diverse, longitudinal clinical speech datasets encompassing multiple conditions, languages, and demographic groups are the single most important investment. The ADReSS benchmarking model should be extended to other conditions and expanded to include acoustic, linguistic, and conversational dimensions. Technical directions: multimodal architectures jointly processing speech audio and linguistic content represent the strongest technical direction. Clinically grounded evaluation frameworks analogous to the CLAIM checklist for medical imaging AI [127] or SPIRIT-AI guidelines [128] would establish minimum reporting standards. Translational requirements: prospective validation studies must evaluate not only accuracy but impact on

clinical workflows: whether AI assessments reduce time to diagnosis, improve monitoring, or change treatment decisions. Regulatory readiness (FDA SaMD framework, EU MDR) and ethical governance (consent in cognitively impaired populations, privacy-preserving methods, bias auditing) must advance in parallel. None of these priorities can be achieved by the speech technology community alone; closing the gap requires partnership with neurologists, healthcare systems, and patients [129,130].

8.6. Regulatory Pathways and Clinical Implementation

Translation of speech-based AI tools into clinical neurology requires a regulatory pathway that does not yet exist for this class of biomarker. Under the FDA Software as a Medical Device (SaMD) framework, a speech analysis tool intended for cognitive screening would likely be classified as Class II (moderate risk), requiring a 510(k) or De Novo pathway with clinical performance data. A tool intended to inform diagnostic decisions (e.g., differentiating AD from frontotemporal dementia based on speech patterns) would face higher scrutiny. The EU Medical Device Regulation (MDR) imposes similar tiered requirements. FDA guidance on predetermined change control plans for adaptive AI/ML-based devices is relevant here, as speech models that update with new data or population shifts would need prespecified update protocols. These regulatory expectations differ across intended uses: a screening tool for population-level triage, a symptom monitoring system for longitudinal tracking, an automated scoring aid for cognitive tests, and a clinical decision support system. Each carries distinct evidence requirements, risk classifications, and deployment constraints that must be addressed individually.

Clinical validation must extend beyond discrimination metrics. Validation against established clinical endpoints (specialist-confirmed diagnosis, standardized cognitive scores such as MoCA and MMSE, disease progression on UPDRS for PD or ALSFRS-R for ALS, neuroimaging biomarkers) is a minimum requirement. Concurrent validity with these endpoints, followed by prospective predictive validation demonstrating that the tool improves outcomes or clinical efficiency, would constitute the evidence base needed for regulatory clearance and clinical adoption.

Implementation science considerations are absent from the current literature. Where in the clinical pathway would a speech-based AI tool be deployed: primary care screening, specialist triage, or remote monitoring between visits? How would results be presented to clinicians, and what actions would they trigger? How would false positives (unnecessary referrals, patient anxiety) and false negatives (missed diagnoses) be managed? No reimbursement codes exist for AI-assisted speech assessment, and integration with electronic health records remains an unsolved engineering and governance challenge. Until these practical barriers are addressed, clinical deployment will remain theoretical regardless of technical performance.

8.7. Limitations

This narrative review has four limitations that should be considered when interpreting its findings. First, narrative reviews are inherently subject to selection bias. Although we describe the search strategy, screening process, and inclusion criteria in Section 2, study selection was performed by a single author without independent duplicate screening, and the prioritization of studies relied in part on subjective judgments of relevance and methodological influence. Studies that were not indexed in the searched databases, published in languages other than English, or presented in venues outside the searched conference proceedings may have been missed. Second, publication bias is likely to affect the evidence base: studies reporting positive or high-accuracy results are more likely to be published than null or negative findings, and this review reflects the published literature as it stands.

The performance metrics reported across studies should therefore be interpreted as upper-bound estimates. Third, we did not apply a formal quality assessment or risk-of-bias tool (e.g., QUADAS-2 for diagnostic accuracy studies) to the included studies. While methodological limitations of individual studies are discussed throughout the text, the absence of a standardized quality appraisal means that studies of varying rigour are presented alongside one another without formal weighting. Fourth, the rapid pace of development in foundation models and LLMs means that some findings may already be superseded by newer work not captured within our search window.

9. Conclusions

The progression from handcrafted acoustic features through self-supervised speech representations to large language models and conversational AI has expanded what can be extracted from the speech signal. Each stage has advanced analytical capability but introduced new limitations: domain mismatch for SSL models, text-only access for LLMs, and absent clinical validation for conversational AI.

The most important gap identified by this review is not technological but evaluative. Whether these capabilities translate into improved diagnosis, earlier detection, more sensitive monitoring, or better patient outcomes has not been tested in the clinical settings where such tools would be used. No speech-based AI system has yet shown the evidence base required for deployment in neurological clinical practice.

The convergence of speech foundation models and large language models creates the technical possibility of integrated neurological speech agents: systems that can conduct naturalistic clinical conversations, analyze both the voice and its content, and generate structured clinical assessments. Whether this possibility is realized in a clinically responsible manner will depend less on further technical advances than on the willingness of the field to invest in foundational infrastructure (diverse datasets, standardized benchmarks), translational partnerships (clinicians, patients, healthcare systems), and governance frameworks (ethical oversight, regulatory engagement) needed to distinguish genuine utility from impressive demonstration.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/computation14070160/s1>. Supplementary Table S1. Selected modern studies in computational speech analysis for neurological disorders, organized by condition and analytical stage. Supplementary Table S2. Speech features by neurological condition, primary analytical stage, and key limitations. Supplementary Table S3. Clinical speech datasets used in the reviewed literature. Supplementary Table S4. Speech and language foundation models applied to neurological speech analysis. Supplementary Table S5. Search strategy: exact search strings, database-specific modifications, and applied filters. Figure S1. Simplified PRISMA-style flow diagram of the study selection process.

Funding: This research received no external funding.

Institutional Review Board Statement: This study is a narrative review of published literature and did not involve human participants, animal subjects, or identifiable personal data. No ethical approval was required.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: Figures 1 and 2 were generated using the AI tool Google's Gemini 3 Pro Image version based on text prompts defined by the authors. Artificial intelligence (AI) tools were used to assist with language editing, improve clarity and readability, provide feedback on manuscript structure, and generate visual materials based on author-defined prompts. All AI-generated outputs

were reviewed, verified, and revised by the authors, who take full responsibility for the final content of the manuscript.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

AD	Alzheimer’s Disease
AI	Artificial Intelligence
ALS	Amyotrophic Lateral Sclerosis
ASR	Automatic Speech Recognition
AUC	Area Under the Receiver Operating Characteristic Curve
BERT	Bidirectional Encoder Representations from Transformers
FTD	Frontotemporal Dementia
GPT	Generative Pre-trained Transformer
HuBERT	Hidden-Unit BERT
ICASSP	International Conference on Acoustics, Speech, and Signal Processing
Interspeech	Annual Conference of the International Speech Communication Association
LLM	Large Language Model
MCI	Mild Cognitive Impairment
NLP	Natural Language Processing
PD	Parkinson’s Disease
RMSE	Root Mean Squared Error
SSL	Self-Supervised Learning
TTR	Type-Token Ratio
MLU	Mean Length of Utterance

References

1. Duffy, J.R. *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*, 4th ed.; Elsevier: Amsterdam, The Netherlands, 2019.
2. Levelt, W.J.M. *Speaking: From Intention to Articulation*; MIT Press: Cambridge, MA, USA, 1989.
3. Rusz, J.; Cmejla, R.; Ruzickova, H.; Ruzicka, E. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson’s disease. *J. Acoust. Soc. Am.* **2011**, *129*, 350–367. [[CrossRef](#)] [[PubMed](#)]
4. Cummings, L. *Clinical Linguistics*; Edinburgh University Press: Edinburgh, UK, 2008.
5. Darley, F.L.; Aronson, A.E.; Brown, J.R. Differential diagnostic patterns of dysarthria. *J. Speech Hear. Res.* **1969**, *12*, 246–269. [[CrossRef](#)] [[PubMed](#)]
6. Goodglass, H.; Kaplan, E. *The Assessment of Aphasia and Related Disorders*, 2nd ed.; Lea & Febiger: Philadelphia, PA, USA, 1983.
7. Tsanas, A.; Little, M.A.; McSharry, P.E.; Spielman, J.; Ramig, L.O. Novel speech signal processing algorithms for high-accuracy classification of Parkinson’s disease. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 1264–1271. [[CrossRef](#)] [[PubMed](#)]
8. Meilan, J.J.G.; Martinez-Sanchez, F.; Carro, J.; Lopez, D.E.; Millian-Morell, L.; Arana, J.M. Speech in Alzheimer’s disease: Can temporal and acoustic parameters discriminate dementia? *Dement. Geriatr. Cogn. Disord.* **2014**, *37*, 327–334. [[PubMed](#)]
9. Fraser, K.C.; Meltzer, J.A.; Rudzicz, F. Linguistic features identify Alzheimer’s disease in narrative speech. *J. Alzheimers Dis.* **2016**, *49*, 407–422. [[PubMed](#)]
10. Le, D.; Licata, K.; Provost, E.M. Automatic quantitative analysis of spontaneous aphasic speech. *Speech Commun.* **2018**, *100*, 1–12. [[CrossRef](#)]
11. Robin, J.; Harrison, J.E.; Kaufman, L.D.; Rudzicz, F.; Simpson, W.; Yancheva, M. Evaluation of speech-based digital biomarkers: Review and recommendations. *Digit. Biomark.* **2020**, *4*, 99–108. [[CrossRef](#)] [[PubMed](#)]
12. de la Fuente Garcia, S.; Ritchie, C.W.; Luz, S. Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer’s disease: A systematic review. *J. Alzheimers Dis.* **2020**, *78*, 1547–1574. [[CrossRef](#)] [[PubMed](#)]
13. 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.* **2017**, *124*, 303–334. [[CrossRef](#)] [[PubMed](#)]
14. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
15. Hsu, W.N.; Bolte, B.; Tsai, Y.H.; Lakhota, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [[CrossRef](#)]

16. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. [\[CrossRef\]](#)
17. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. *Proc. ICML* **2023**, *202*, 28492–28518.
18. OpenAI. GPT-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
19. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Roziere, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
20. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Moor, M.; Banerjee, O.; Abad, Z.S.H.; Krumholz, H.M.; Leskovec, J.; Topol, E.J.; Rajpurkar, P. Foundation models for generalist medical artificial intelligence. *Nature* **2023**, *616*, 259–265. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Violeta, L.P.; Huang, W.-C.; Toda, T. Investigating self-supervised pretraining frameworks for pathological speech recognition. *arXiv* **2022**, arXiv:2203.15431.
24. Nori, H.; King, N.; McKinney, S.M.; Carignan, D.; Horvitz, E. Capabilities of GPT-4 on medical challenge problems. *arXiv* **2023**, arXiv:2303.13375.
25. Tu, T.; Azizi, S.; Driess, D.; Schaeckermann, M.; Amin, M.; Chang, P.-C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. Towards generalist biomedical AI. *NEJM AI* **2024**, *1*, A10a2300138. [\[CrossRef\]](#)
26. GBD 2021 Nervous System Disorders Collaborators. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: A systematic analysis for the Global Burden of Disease Study 2021. *Lancet Neurol.* **2024**, *23*, 344–381. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Green, B.N.; Johnson, C.D.; Adams, A. Writing narrative literature reviews for peer-reviewed journals: Secrets of the trade. *J. Chiropr. Med.* **2006**, *5*, 101–117. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Baumeister, R.F.; Leary, M.R. Writing narrative literature reviews. *Rev. Gen. Psychol.* **1997**, *1*, 311–320. [\[CrossRef\]](#)
29. Ho, A.K.; Iansek, R.; Marigliani, C.; Bradshaw, J.L.; Gates, S. Speech impairment in a large sample of patients with Parkinson's disease. *Behav. Neurol.* **1999**, *11*, 131–137. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Ramig, L.O.; Fox, C.; Sapir, S. Speech treatment for Parkinson's disease. *Expert. Rev. Neurother.* **2008**, *8*, 297–309. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Little, M.A.; McSharry, P.E.; Hunter, E.J.; Spielman, J.; Ramig, L.O. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Trans. Biomed. Eng.* **2009**, *56*, 1015–1022. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Skodda, S.; Visser, W.; Schlegel, U. Vowel articulation in Parkinson's disease. *J. Voice* **2011**, *25*, 467–472. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Rusz, J.; Hlavnicka, J.; Tykalova, T.; Buskova, J.; Ulmanova, O.; Ruzicka, E.; Sonka, K. Quantitative assessment of motor speech abnormalities in idiopathic rapid eye movement sleep behaviour disorder. *Sleep Med.* **2016**, *19*, 141–147. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Orozco-Arroyave, J.R.; Arias-Londono, J.D.; Vargas-Bonilla, J.F.; Gonzalez-Rativa, M.C.; Noth, E. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In *LREC*; European Language Resources Association (ELRA): Luxembourg, 2014; pp. 342–347.
35. Szatloczki, G.; Hoffmann, I.; Vincze, V.; Kalman, J.; Pakaski, M. Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Front. Aging Neurosci.* **2015**, *7*, 195. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Ahmed, S.; Haigh, A.M.; de Jager, C.A.; Garrard, P. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain* **2013**, *136*, 3727–3737. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Luz, S.; Haider, F.; de la Fuente Garcia, S.; Fromm, D.; MacWhinney, B. Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. *Front. Comput. Sci.* **2021**, *3*, 780169. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Luz, S.; Haider, F.; de la Fuente Garcia, S.; Fromm, D.; MacWhinney, B. Detecting cognitive decline using speech only: The ADReSSo Challenge. *arXiv* **2021**, arXiv:2104.09356.
39. Green, J.R.; Yunusova, Y.; Kuruvilla, M.S.; Wang, J.; Pattee, G.L.; Synhorst, L.; Zinman, L.; Berry, J.D. Bulbar and speech motor assessment in ALS: Challenges and future directions. *Amyotroph. Lateral Scler. Front. Degener.* **2013**, *14*, 494–500. [\[CrossRef\]](#)
40. Stegmann, G.M.; Hahn, S.; Liss, J.; Shefner, J.; Rutkove, S.; Shelton, K.; Duncan, C.J.; Berisha, V. Early detection and tracking of bulbar changes in ALS via frequent and remote speech analysis. *npj Digit. Med.* **2020**, *3*, 132. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Fusaroli, R.; Lambrechts, A.; Bang, D.; Bowler, D.M.; Gaigg, S.B. Is voice a marker for autism spectrum disorder? A systematic review and meta-analysis. *Autism Res.* **2017**, *10*, 384–407. [\[PubMed\]](#)
42. Bone, D.; Lee, C.-C.; Black, M.P.; Williams, M.E.; Lee, S.; Levitt, P.; Narayanan, S. The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody. *J. Speech Lang. Hear. Res.* **2014**, *57*, 1162–1177. [\[CrossRef\]](#) [\[PubMed\]](#)

43. Feijó, A.V.; Parente, M.A.; Behlau, M.; Haussen, S.; de Veccino, M.C.; Martignago, B.C. Acoustic analysis of voice in multiple sclerosis patients. *J. Voice* **2004**, *18*, 341–347. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Noffs, G.; Perera, T.; Kolbe, S.C.; Shanahan, C.J.; Boonstra, F.M.C.; Evans, A.; Butzkueven, H.; van der Walt, A.; Vogel, A.P. What speech can tell us: A systematic review of dysarthria characteristics in multiple sclerosis. *Autoimmun. Rev.* **2018**, *17*, 1202–1209. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Rusz, J.; Saft, C.; Schlegel, U.; Hoffmann, R.; Skodda, S. Phonatory dysfunction as a preclinical symptom of Huntington disease. *PLoS ONE* **2014**, *9*, e113412. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Coelho, C.A.; Liles, B.Z.; Duffy, R.J. Impairments of discourse abilities and executive functions in traumatically brain-injured adults. *Brain Inj.* **1995**, *9*, 471–477. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Mohamed, A.; Lee, H.-Y.; Borgholt, L.; Havtorn, J.D.; Edin, J.; Igel, C.; Kirchhoff, K.; Li, S.-W.; Livescu, K.; Maaloe, L.; et al. Self-supervised speech representation learning: A review. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1179–1210. [\[CrossRef\]](#)
48. Yang, S.-W.; Chi, P.-H.; Chuang, Y.-S.; Lai, C.-I.J.; Lakhotia, K.; Lin, Y.Y.; Liu, A.T.; Shi, J.; Chang, X.; Lin, G.-T.; et al. SUPERB: Speech processing universal PERformance benchmark. *arXiv* **2021**, arXiv:2105.01051.
49. Bayerl, S.P.; Wagner, D.; Noth, E.; Riedhammer, K. Detecting Dysfluencies in Stuttering Therapy Using wav2vec 2.0. *arXiv* **2022**, arXiv:2204.03417.
50. Javanmardi, F.; Tirronen, S.; Kodali, M.; Alku, P. wav2vec-based detection and severity level classification of dysarthria from speech. In *Icassp 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (Icassp)*; IEEE: New York, NY, USA, 2023; pp. 1–5.
51. Balagopalan, A.; Eyre, B.; Rudzicz, F.; Novikova, J. To BERT or not to BERT: Comparing speech and language-based approaches for Alzheimer’s disease detection. *arXiv* **2020**, arXiv:2008.01551.
52. Pappagari, R.; Cho, J.; Joshi, S.; Moro-Velazquez, L.; Zelasko, P.; Villalba, J.; Dehak, N. Automatic detection and assessment of Alzheimer disease using speech and language technologies in low-resource scenarios. In *Proceedings of Interspeech 2021*; ISCA: Antwerp, Belgium, 2021; pp. 3825–3829.
53. Favaro, A.; Tsai, Y.-T.; Butala, A.; Thebaud, T.; Villalba, J.; Dehak, N.; Moro-Velazquez, L. Interpretable speech features vs. DNN embeddings: What to use in the automatic assessment of Parkinson’s disease in multi-lingual scenarios. *Comput. Biol. Med.* **2023**, *166*, 107559. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Rowe, H.P.; Gutz, S.E.; Maffei, M.F.; Tomanek, K.; Green, J.R. Characterizing dysarthria diversity for automatic speech recognition: A tutorial from the clinical perspective. *Front. Comput. Sci.* **2022**, *4*, 770210. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Xiong, F.; Barker, J.; Yue, Z.; Christensen, H. Source domain data selection for improved transfer learning targeting dysarthric speech recognition. In *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2020; pp. 7424–7428.
56. Rudzicz, F.; Namasivayam, A.K.; Wolff, T. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Lang. Resour. Eval.* **2012**, *46*, 523–541.
57. Kim, H.; Hasegawa-Johnson, M.; Perlman, A.; Gunderson, J.; Huang, T.S.; Watkin, K.; Frame, S. Dysarthric speech database for universal access research. In *Proceedings of Interspeech 2008*; ISCA: Antwerp, Belgium, 2008; pp. 1741–1744.
58. Haulcy, R.; Glass, J. Classifying Alzheimer’s disease using audio and text-based representations of speech. *Front. Psychol.* **2021**, *11*, 624137. [\[CrossRef\]](#) [\[PubMed\]](#)
59. Pasad, A.; Chou, J.C.; Livescu, K. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*; IEEE: New York, NY, USA, 2021; pp. 914–921.
60. Bucks, R.S.; Singh, S.; Cuerden, J.M.; Wilcock, G.K. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology* **2000**, *14*, 71–91. [\[CrossRef\]](#)
61. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 2019; pp. 4171–4186.
62. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [\[PubMed\]](#)
63. Alsentzer, E.; Murphy, J.R.; Boag, W.; Weng, W.-H.; Jindi, D.; Naumann, T.; McDermott, M. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 2019; pp. 72–78.
64. Balagopalan, A.; Novikova, J. Comparing acoustic-based approaches for Alzheimer’s disease detection. *arXiv* **2021**, arXiv:2106.01555.
65. 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: New York, NY, USA, 2021.

66. Agbavor, F.; Liang, H. Predicting dementia from spontaneous speech using large language models. *PLoS Digit. Health* **2022**, *1*, e0000168. [\[CrossRef\]](#) [\[PubMed\]](#)
67. Balamurali, B.T.; Chen, J.-M. Performance assessment of ChatGPT versus Bard in detecting Alzheimer's dementia. *Diagnostics* **2024**, *14*, 817. [\[CrossRef\]](#)
68. Taherinezhad, F.; Momeni Nezhad, M.J.; Karimi, S.; Rashidi, S.; Zolnour, A.; Dadkhah, M.; Haghbin, Y.; Azadmaleki, H.; Zolnoori, M. Large language model adaptation strategies in speech-based cognitive screening: Systematic evaluation. *JMIR AI* **2026**, *5*, e82608. [\[CrossRef\]](#) [\[PubMed\]](#)
69. Chlasta, K.; Struzik, P.; Wojcik, G.M. Enhancing dementia and cognitive decline detection with large language models and speech representation learning. *Front. Neuroinform.* **2025**, *19*, 1679664. [\[CrossRef\]](#) [\[PubMed\]](#)
70. Kurland, J.; Varadharaju, V.; Liu, A.; Stokes, P.; Gupta, A.; Hudspeth, M.; O'Connor, B. Large language models' ability to assess main concepts in story retelling: A proof-of-concept comparison of human versus machine ratings. *Am. J. Speech Lang. Pathol.* **2025**, *34*, 3636–3646. [\[CrossRef\]](#) [\[PubMed\]](#)
71. Fei, X.; Tang, Y.; Zhang, J.; Zhou, Z.; Yamamoto, I.; Zhang, Y. Evaluating cognitive performance: Traditional methods vs. ChatGPT. *Digit. Health* **2024**, *10*, 20552076241264639. [\[CrossRef\]](#) [\[PubMed\]](#)
72. Laine, M.; Laakso, M.; Vuorinen, E.; Rinne, J. Coherence and informativeness of discourse in two dementia types. *J. Neurolinguist.* **1998**, *11*, 79–87. [\[CrossRef\]](#)
73. Ash, S.; Moore, P.; Antani, S.; McCawley, G.; Work, M.; Grossman, M. Trying to tell a tale: Discourse impairments in progressive aphasia and frontotemporal dementia. *Neurology* **2006**, *66*, 1405–1413. [\[CrossRef\]](#) [\[PubMed\]](#)
74. Iter, D.; Yoon, J.; Jurafsky, D. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 2018; pp. 136–146.
75. Mota, N.B.; Vasconcelos, N.A.P.; Lemos, N.; Pieretti, A.C.; Kinouchi, O.; Cecchi, G.A.; Copelli, M.; Ribeiro, S. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS ONE* **2012**, *7*, e34928. [\[CrossRef\]](#) [\[PubMed\]](#)
76. Ilias, L.; Askounis, D. Multimodal deep learning models for detecting dementia from speech and transcripts. *Front. Aging Neurosci.* **2022**, *14*, 830943. [\[CrossRef\]](#) [\[PubMed\]](#)
77. Hobbs, J.R. Coherence and coreference. *Cogn. Sci.* **1979**, *3*, 67–90. [\[CrossRef\]](#)
78. Kaplan, E.; Goodglass, H.; Weintraub, S. *The Boston Naming Test*, 2nd ed.; Lea & Febiger: Philadelphia, PA, USA, 1983.
79. Troyer, A.K.; Moscovitch, M.; Winocur, G. Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology* **1997**, *11*, 138–146. [\[CrossRef\]](#) [\[PubMed\]](#)
80. Toth, L.; Hoffmann, I.; Gosztolya, G.; Vincze, V.; Sztaloczk, G.; Banreti, Z.; Pakaski, M.; Kalman, J. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Curr. Alzheimer Res.* **2018**, *15*, 130–138. [\[CrossRef\]](#) [\[PubMed\]](#)
81. Lehr, M.; Prud'hommeaux, E.; Shafran, I.; Roark, B. Fully automated neuropsychological assessment for detecting mild cognitive impairment. In *Proceedings of Interspeech 2012*; ISCA: Antwerp, Belgium, 2012.
82. Fristed, E.; Skirrow, C.; Meszaros, M.; Lenain, R.; Meepegama, U.; Cappa, S.; Aarsland, D.; Weston, J. A remote speech-based AI system to screen for early Alzheimer's disease via smartphones. *Alzheimers Dement.* **2022**, *14*, e12366. [\[CrossRef\]](#)
83. Heritage, J.; Maynard, D.W. *Communication in Medical Care: Interaction Between Primary Care Physicians and Patients*; Cambridge University Press: Cambridge, UK, 2006.
84. Perkins, L.; Whitworth, A.; Lesser, R. Conversing in dementia: A conversation analytic approach. *J. Neurolinguist.* **1998**, *11*, 33–53. [\[CrossRef\]](#)
85. Mirheidari, B.; Blackburn, D.; Walker, T.; Reuber, M.; Christensen, H. Dementia detection using automatic analysis of conversations. *Comput. Speech Lang.* **2019**, *53*, 65–79. [\[CrossRef\]](#)
86. Mirheidari, B.; Blackburn, D.; Walker, T.; Reuber, M.; Christensen, H. Diagnosing people with dementia using automatic conversation analysis. In *Proceedings of Interspeech 2016*; ISCA: Antwerp, Belgium, 2016; pp. 1220–1224.
87. Tanaka, H.; Adachi, H.; Ukita, N.; Ikeda, M.; Kazui, H.; Kudo, T.; Nakamura, S. Detecting dementia through interactive computer avatars. *IEEE J. Transl. Eng. Health Med.* **2017**, *5*, 2200111. [\[CrossRef\]](#) [\[PubMed\]](#)
88. Takeshige-Amano, H.; Oyama, G.; Ogawa, M.; Fusegi, K.; Kambe, T.; Shiina, K.; Ueno, S.-I.; Okuzumi, A.; Hatano, T.; Motoi, Y.; et al. Digital detection of Alzheimer's disease using smiles and conversations with a chatbot. *Sci. Rep.* **2024**, *14*, 26309. [\[CrossRef\]](#) [\[PubMed\]](#)
89. Serafimovska, A.; Swavley, K.; Zhang Qian Ao, A.; Challinor, K.L.; Florio, T. Cognitive status assessment of older adults: Test administration by conversational artificial intelligence (AI) chatbot: Proof-of-concept investigation. *J. Clin. Exp. Neuropsychol.* **2025**, *47*, 472–484. [\[CrossRef\]](#) [\[PubMed\]](#)
90. Konig, A.; Kohler, S.; Troger, J.; Duzel, E.; Glanz, W.; Butryn, M.; Mallick, E.; Priller, J.; Altenstein, S.; Spottke, A.; et al. Automated remote speech-based testing of individuals with cognitive decline: Bayesian agreement of transcription accuracy. *Alzheimers Dement.* **2024**, *16*, e70011.

91. Pakhomov, S.; Solinsky, J.; Michalowski, M.; Bachanova, V. A conversational agent for early detection of neurotoxic effects of medications through automated intensive observation. *Pac. Symp. Biocomput.* **2024**, *29*, 24–38. [PubMed]
92. Ter Huurne, D.; Ramakers, I.; Possemis, N.; Konig, A.; Linz, N.; Troger, J.; Langel, K.; Verhey, F.; de Vugt, M. User experience of a (semi-) automated cognitive phone-based assessment within a memory clinic population. *Arch. Clin. Neuropsychol.* **2025**, *40*, 319–329. [PubMed]
93. Sacks, H.; Schegloff, E.A.; Jefferson, G. A simplest systematics for the organization of turn-taking for conversation. *Language* **1974**, *50*, 696–735. [CrossRef]
94. Orange, J.B.; Lubinski, R.B.; Higginbotham, D.J. Conversational repair by individuals with dementia of the Alzheimer's type. *J. Speech Hear. Res.* **1996**, *39*, 881–895. [CrossRef] [PubMed]
95. McNamara, P.; Durso, R. Pragmatic communication skills in patients with Parkinson's disease. *Brain Lang.* **2003**, *84*, 414–423. [CrossRef] [PubMed]
96. Elsey, C.; Drew, P.; Jones, D.; Blackburn, D.; Wakefield, S.; Harkness, K.; Venneri, A.; Reuber, M. Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics. *Patient Educ. Couns.* **2015**, *98*, 1071–1077. [CrossRef] [PubMed]
97. Jones, D.; Drew, P.; Elsey, C.; Blackburn, D.; Wakefield, S.; Harkness, K.; Reuber, M. Conversational assessment in memory clinic encounters: Interactional profiling for differentiating dementia from functional memory disorders. *Aging Ment. Health* **2016**, *20*, 500–509. [PubMed]
98. Robin, J.; Xu, M.; Kaufman, L.D.; Simpson, W. Using digital speech assessments to detect early signs of cognitive impairment. *Front. Digit. Health* **2021**, *3*, 749758. [CrossRef] [PubMed]
99. Sabbagh, M.N.; Boada, M.; Borson, S.; Chilukuri, M.; Dubois, B.; Ingram, J.; Iwata, A.; Porsteinsson, A.P.; Possin, K.L.; Rabinovici, G.D.; et al. Early detection of mild cognitive impairment (MCI) in primary care. *J. Prev. Alzheimers Dis.* **2020**, *7*, 165–170. [CrossRef] [PubMed]
100. Rohanian, M.; Hough, J.; Purver, M. Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech. *arXiv* **2020**, arXiv:2106.09668.
101. Gorno-Tempini, M.L.; Hillis, A.E.; Weintraub, S.; Kertesz, A.; Mendez, M.; Cappa, S.F.; Ogar, J.M.; Rohrer, J.D.; Black, S.; Boeve, B.F.; et al. Classification of primary progressive aphasia and its variants. *Neurology* **2011**, *76*, 1006–1014. [CrossRef] [PubMed]
102. Duffy, J.R.; Strand, E.A.; Clark, H.M.; Machulda, M.M.; Whitwell, J.L.; Josephs, K.A. Primary progressive apraxia of speech: Clinical features and acoustic and neurologic correlates. *Am. J. Speech Lang. Pathol.* **2015**, *24*, 88–100. [CrossRef] [PubMed]
103. OpenAI. GPT-4o System Card. 2024. Available online: <https://openai.com/index/gpt-4o-system-card/> (accessed on 15 June 2026).
104. Gemini Team Google. Gemini: A family of highly capable multimodal models. *arXiv* **2023**, arXiv:2312.11805.
105. Dorsey, E.R.; Glidden, A.M.; Holloway, M.R.; Birbeck, G.L.; Schwamm, L.H. Teleneurology and mobile technologies: The future of neurological care. *Nat. Rev. Neurol.* **2018**, *14*, 285–297. [CrossRef] [PubMed]
106. Tsanas, A.; Little, M.A.; McSharry, P.E.; Ramig, L.O. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Trans. Biomed. Eng.* **2010**, *57*, 884–893. [CrossRef] [PubMed]
107. Lipsmeier, F.; Taylor, K.I.; Kilchenmann, T.; Wolf, D.; Scotland, A.; Schjodt-Eriksen, J.; Cheng, W.-Y.; Fernandez-Garcia, I.; Siebourg-Polster, J.; Jin, L.; et al. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. *Mov. Disord.* **2018**, *33*, 1287–1297. [CrossRef] [PubMed]
108. Connaghan, K.P.; Green, J.R.; Paganoni, S.; Chan, J.; Weber, H.; Collins, E.; Richburg, B.; Eshghi, M.; Onnela, J.-P.; Berry, J.D. Use of Beiwe smartphone application to identify and track speech decline in amyotrophic lateral sclerosis. In *Proceedings of Interspeech 2019*; ISCA: Antwerp, Belgium, 2019; pp. 4504–4508.
109. Laird, N.M.; Ware, J.H. Random-effects models for longitudinal data. *Biometrics* **1982**, *38*, 963–974. [CrossRef] [PubMed]
110. Torous, J.; Kiang, M.V.; Lorme, J.; Onnela, J.P. New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. *JMIR Ment. Health* **2016**, *3*, e16. [CrossRef] [PubMed]
111. Insel, T.R. Digital phenotyping: Technology for a new science of behavior. *JAMA* **2017**, *318*, 1215–1216. [CrossRef] [PubMed]
112. Fagherazzi, G.; Fischer, A.; Ismael, M.; Despotovic, V. Voice for health: The use of vocal biomarkers from research to clinical practice. *Digit. Biomark.* **2021**, *5*, 78–88. [CrossRef] [PubMed]
113. Martinez-Martin, N.; Kreitmair, K. Ethical issues for direct-to-consumer digital psychotherapy apps: Addressing accountability, data protection, and consent. *JMIR Ment. Health* **2018**, *5*, e32. [CrossRef] [PubMed]
114. Gerke, S.; Minssen, T.; Cohen, G. Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial Intelligence in Healthcare*; Bohr, A., Memarzadeh, K., Eds.; Academic Press: New York, NY, USA, 2020; pp. 295–336.
115. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. [CrossRef] [PubMed]
116. Becker, J.T.; Boller, F.; Lopez, O.L.; Saxton, J.; McGonigle, K.L. The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Arch. Neurol.* **1994**, *51*, 585–594. [CrossRef] [PubMed]

117. Wroge, T.J.; Ozkanca, Y.; Demiroglu, C.; Si, D.; Atkins, D.C.; Ghomi, R.H. Parkinson's disease diagnosis using machine learning and voice. In *Proceedings of the 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*; IEEE: New York, NY, USA, 2018; pp. 1–7.
118. Shellikeri, S.; Green, J.R.; Kulkarni, M.; Rong, P.; Martino, R.; Zinman, L.; Yunusova, Y. Speech movement measures as markers of bulbar disease in amyotrophic lateral sclerosis. *J. Speech Lang. Hear. Res.* **2016**, *59*, 887–899. [[CrossRef](#)] [[PubMed](#)]
119. Ma, J.K.; Whitehill, T.L.; So, S.Y. Intonation contrast in Cantonese speakers with hypokinetic dysarthria associated with Parkinson's disease. *J. Speech Lang. Hear. Res.* **2010**, *53*, 836–849. [[CrossRef](#)] [[PubMed](#)]
120. Weiner, J.; Herff, C.; Schultz, T. Speech-based detection of Alzheimer's disease in conversational German. In *Proceedings of Interspeech 2016*; ISCA: Antwerp, Belgium, 2016; pp. 1938–1942.
121. Lee, P.; Bubeck, S.; Petro, J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N. Engl. J. Med.* **2023**, *388*, 1233–1239. [[CrossRef](#)] [[PubMed](#)]
122. Bates, D.W.; Levine, D.; Syrowatka, A.; Kuznetsova, M.; Thomas Craig, K.J.; Rui, A.; Jackson, G.P.; Rhee, K. The potential of artificial intelligence to improve patient safety: A scoping review. *npj Digit. Med.* **2021**, *4*, 54. [[CrossRef](#)] [[PubMed](#)]
123. Coiera, E. The last mile: Where artificial intelligence meets reality. *J. Med. Internet Res.* **2019**, *21*, e16323. [[CrossRef](#)] [[PubMed](#)]
124. Ienca, M.; Wangmo, T.; Jotterand, F.; Kressig, R.W.; Elger, B. Ethical design of intelligent assistive technologies for dementia: A descriptive review. *Sci. Eng. Ethics* **2018**, *24*, 1035–1055. [[PubMed](#)]
125. Koenecke, A.; Nam, A.; Lake, E.; Nudell, J.; Quartey, M.; Mengesha, Z.; Toups, C.; Rickford, J.R.; Jurafsky, D.; Goel, S. Racial disparities in automated speech recognition. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 7684–7689. [[CrossRef](#)] [[PubMed](#)]
126. Mittelstadt, B.D.; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data Soc.* **2016**, *3*, 2053951716679679. [[CrossRef](#)]
127. Mongan, J.; Moy, L.; Kahn, C.E., Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. *Radiol. Artif. Intell.* **2020**, *2*, e200029. [[CrossRef](#)] [[PubMed](#)]
128. Cruz Rivera, S.; Liu, X.; Chan, A.-W.; Denniston, A.K.; Calvert, M.J. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nat. Med.* **2020**, *26*, 1351–1363. [[CrossRef](#)] [[PubMed](#)]
129. Bhati, D.; Neha, F.; Bandaru, D.S.; Weber, M.; Gajera, I.D. Mapping the LLM landscape: A cross-family survey of architectures, alignment methods, and benchmark performance. *AI* **2026**, *7*, 142. [[CrossRef](#)]
130. Neha, F.; Bhati, D.; Shukla, D.K. Retrieval-augmented generation (RAG) in healthcare: A comprehensive review. *AI* **2025**, *6*, 226. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.