

The PARLO Dementia Corpus: A German Multi-Center Resource for Alzheimer’s Disease

Franziska Braun*, Christopher Witzl*, Florian Hönig†,
Elmar Nöth‡, Tobias Bocklet*, Korbinian Riedhammer*

*Technische Hochschule Nürnberg, Germany

†KST Institut GmbH Bad Emstal, Germany

‡Friedrich-Alexander-Universität Erlangen, Germany

franziska.braun@th-nuernberg.de, korbinian.riedhammer@th-nuernberg.de

Abstract

Early and accessible detection of Alzheimer’s disease (AD) remains a major challenge, as current diagnostic methods often rely on costly and invasive biomarkers. Speech and language analysis has emerged as a promising non-invasive and scalable approach to detecting cognitive impairment, but research in this area is hindered by the lack of publicly available datasets, especially for languages other than English. This paper introduces the PARLO Dementia Corpus (PDC), a new multi-center, clinically validated German resource for AD collected across nine academic memory clinics in Germany. The dataset comprises speech recordings from individuals with AD-related mild cognitive impairment and mild to moderate dementia, as well as cognitively healthy controls. Speech was elicited using a standardized test battery of eight neuropsychological tasks, including confrontation naming, verbal fluency, word repetition, picture description, story reading, and recall tasks. In addition to audio recordings, the dataset includes manually verified transcriptions and detailed demographic, clinical, and biomarker metadata. Baseline experiments on ASR benchmarking, automated test evaluation, and LLM-based classification illustrate the feasibility of automatic, speech-based cognitive assessment and highlight the diagnostic value of recall-driven speech production. The PDC thus establishes the first publicly available German benchmark for multi-modal and cross-lingual research on neurodegenerative diseases.

Keywords: dementia screening, pathological speech, neuropsychological tests

1. Introduction

Alzheimer’s disease (AD) is the most prevalent cause of dementia worldwide and poses major challenges to healthcare systems in aging societies. The progressive decline in memory, language and executive functions profoundly affects the autonomy and quality of life of those affected. Early detection is essential to enable timely intervention, treatment planning, and inclusion in clinical trials for emerging disease-modifying therapies. However, current diagnostic procedures rely largely on costly and invasive biomarkers such as Positron Emission Tomography (PET) imaging or cerebrospinal fluid analysis, which are not available in many clinical or community settings.

In recent years, speech and language analysis has gained increasing attention as a non-invasive, cost-effective, and scalable approach to detecting cognitive impairment. A growing body of work shows that acoustic and linguistic features extracted from spontaneous or semi-structured speech can serve as early indicators of mild cognitive impairment and dementia (Ramanarayanan, 2024). Speech is a particularly appealing biomarker, as it can be recorded remotely, repeatedly, and unobtrusively using consumer devices. Despite rapid methodological advances, speech resources are still very limited. Most pub-

licly available corpora, such as the Pitt Corpus (DementiaBank) (Becker et al., 1994), are small, English-centric, and often limited to single elicitation tasks (e.g., picture description) as opposed to standardized tests. As a result, cross-lingual generalization and reproducibility remain limited. In particular, German resources for investigating cognitive impairment based on speech are rare and not publicly available, and only a few datasets combine clinically verified diagnoses with standardized neuropsychological tasks and human transcriptions.

This paper introduces the PARLO Dementia Corpus (PDC), a new German multi-center corpus for Alzheimer’s disease, which was collected at nine academic memory clinics across Germany as part of a study conducted by the PARLO Institute for Research and Teaching in Speech Therapy. The dataset includes recordings from 208 German-speaking participants, including AD patients with mild cognitive impairment (MCI) and mild to moderate Dementia (DEM), as well as cognitively healthy controls (HC) of comparable age. Speech data was elicited using a standardized iPad-based test battery with eight cognitive tasks, including story reading, verbal fluency (animal naming), confrontation naming, word repetition, picture description, and recall tasks. All recordings were collected under uniform clinical conditions and manually transcribed according to the extended scientific transcription

rules of Dresing & Pehl (Dresing and Pehl, 2010). The resulting multi-modal resource includes audio recordings and manually verified transcriptions, enriched with detailed demographic and clinical metadata, including Mini-Mental State Examination (MMSE) scores, diagnostic labels, and biomarkers from cerebrospinal fluid and medical imaging. The PDC’s design enables research across multiple domains, from automatic speech recognition (ASR) and speech biomarker extraction to automated dementia assessment, as well as cross-lingual and multi-modal studies on cognitive impairment. To demonstrate the applicability of the dataset, we present three baseline experiments:

- ASR benchmarking using three open-source systems to evaluate transcription quality under different cognitive and task conditions.
- Automatic test evaluation for the *Confrontation Naming* and *Animal Naming* tasks using rule-based scoring to validate the agreement between ASR-based and manual scores.
- Vision-LLM zero-shot classification of cognitive impairment to explore the potential of generative models for automatic dementia assessment.

2. Related Work and Data

The automatic detection of Alzheimer’s disease based on speech has become an active area of research. Early systematic reviews and meta-analyses (de la Fuente Garcia et al., 2020; Martínez-Nicolás et al., 2021; Vigo et al., 2022) summarize heterogeneous approaches (e.g., acoustic, linguistic, temporal) and report promising discriminatory performance in small clinical cohorts, but they also point to methodological heterogeneity, limited external validation, and biases in the datasets that hinder clinical implementation. The publicly available Pitt Corpus (Dementia-Bank) (Becker et al., 1994) and derived benchmark sets have been widely used for picture description tasks (“Cookie Theft”) and verbal fluency analyses, enabling comparisons between different studies. Building on these, the ADRess, ADRessO, and TAUkADIAL Challenges (Luz et al., 2020, 2021; Barrera-Altuna et al., 2024) provided acoustically preprocessed, age- and gender-balanced datasets with spontaneous speech and tasks such as MCI and AD classification, regression of cognitive scores (MMSE), and prediction of disease progression. Work in this field generally follows two complementary strategies: (1) pipelines that extract acoustic and linguistic features, followed by classical classifiers (e.g., SVM, RF, gradient boosting) (Ben Ammar and Ben Ayed, 2018; Yuan et al.,

2020b; Calzà et al., 2021; Braun et al., 2022a); and (2) representation learning approaches that fine-tune large pre-trained audio or text encoders (e.g., wav2vec, transformer-based language models) on AD tasks (Yuan et al., 2020a; Balagopalan et al., 2020, 2021; Pappagari et al., 2021). Recent studies show that both acoustic and linguistic signals contain disease-relevant information and that multi-modal fusion often improves robustness (Rohanian et al., 2020; Pompili et al., 2020; Braun et al., 2024; Lin and Washington, 2024). Since English-centric studies do not necessarily generalize to other languages, multilingual studies and challenges (Pérez-Toro et al., 2022; Luz et al., 2023) have investigated cross-lingual approaches and shown that some acoustic markers (e.g., wav2vec) can be transferred between languages, while linguistic markers may require language-aware modeling and evaluation. This motivates language-specific data collection and validation, especially for German, where large, balanced corpora with standardized collection are relatively rare and furthermore not publicly available (Weiner et al., 2016; Braun et al., 2022b).

The PARLO Dementia Corpus fills this gap by providing the first German multi-center, clinically annotated resource for AD, enabling reproducible multi-modal and cross-lingual research on cognitive decline. Subsets of the corpus have already been used in (Pérez-Toro et al., 2023; Braun et al., 2023, 2024), which provided baselines for speech-based classification of HC, MCI, and DEM. Cross-corpus results have shown that disease-based classification across other German datasets (Braun et al., 2023), as well as across English and Spanish datasets (Pérez-Toro et al., 2023) is possible using acoustic, linguistic, and emotional embeddings from several subtests of the PDC (e.g., animal naming, confrontation naming, word repetition, and picture description). Moreover, (Braun et al., 2024) established a baseline for the early detection and monitoring of AD using pause-enriched cross-attention with text and audio data from picture description and animal naming tasks of the PDC.

3. The PARLO Dementia Corpus

Group	<i>n</i>	Age	Gender
HC	83	55-87 (69.0±7.9)	32m/51f
MCI	59	55-85 (70.7±8.4)	32m/27f
DEM	66	55-85 (71.7±8.8)	32m/34f
ALL	208	55-87 (70.4±8.4)	96m/112f

Table 1: Demographics for healthy control (HC), mild cognitive impairment (MCI), and dementia (DEM) groups.

The PARLO Dementia Corpus (PDC) consists of audio recordings from a total of 208 German-speaking subjects (96 men, 112 women) aged between 55 and 87 ($\mu = 70.4 \pm 8.4$). Subjects were recruited as part of a multi-center, cross-sectional, open-label, controlled, parallel group clinical study in nine academic memory clinics across Germany. Research was approved by the Institutional Review Boards (IRBs) responsible for each study site; each subject provided informed consent prior to recording. The demographic distribution of the study population, including healthy controls (HC, $n = 83$), and individuals with AD-related mild cognitive impairment (MCI, $n = 59$) and dementia (DEM, $n = 66$), is given in Table 1.

3.1. Inclusion and Exclusion Criteria

The study population consists of male and female participants aged 55 years or older, including patients diagnosed with Alzheimer’s disease, ranging from MCI to mild and moderate dementia, as well as cognitively healthy controls of comparable age. All participants must have adequate vision and hearing, fluent German language skills, and sufficient verbal abilities to perform the study tasks. Legal competence and written informed consent were required for inclusion. Participation in registries was permitted for all study subjects. For AD patients, an established diagnosis of AD according to the National Institute on Aging-Alzheimer’s Association (NIA-AA) criteria (Jack Jr. et al., 2018) was required prior to study enrollment. For controls, cognitive test scores must fall within normal limits of the CERAD test battery (Morris et al., 1988) (≤ 1.2 standard deviations below age-adjusted norms). Controls must have at least six years of education, report no subjective cognitive complaints, and female participants must be of non-childbearing potential (≥ 1 year postmenopausal or surgically sterilized).

Subjects were excluded if they had any medical or psychiatric condition interfering with cognitive functioning, including clinical signs of depression, or any condition prohibiting the use of a tablet-based application. Other exclusion criteria include ongoing clinically significant metabolic or systemic disease that could impair memory or speech (e.g., untreated thyroid disease, vitamin deficiency, or chronic kidney and liver disease), cognitive impairment attributable to conditions other than early-phase neurodegenerative disease, Parkinson’s disease, history of brain tumor, intracranial lesion, disturbance of cerebrospinal fluid circulation (e.g., normal pressure hydrocephalus), significant head trauma, or brain surgery. Further exclusion criteria include major cerebrovascular disease, as indicated by clinical history or neuroimaging, the use of psychoactive medications likely to interfere

with cognitive testing (e.g., benzodiazepines, sedatives, or antipsychotics), and participation in any parallel clinical trial or investigation.

3.2. PARLO Dementia Test Battery



Figure 1: Picture Description Task “Mountain Scene”

The PARLO Dementia Test Battery (PDTB) comprises eight neuropsychological speech tasks that assess multiple cognitive and linguistic domains, as described below in the order of examination:

1. **Story Reading:** Subjects read aloud a short story (called “Johanna Subway”). This task ensures uniform stimulus exposure and provides a baseline for prosodic, articulation, and fluency analysis. The ground truth story text is included in the corpus.
2. **Confrontation Naming:** Subjects successively view 15 objects (line drawings) and name them. This is a variant of the Boston Naming Test (Kaplan et al., 1978), which assesses lexical access, confrontation naming, and semantic competence. Deficits in naming can be an indicator of neurodegenerative diseases.
3. **Animal Naming:** In one minute, subjects name as many animals as possible. This is a semantic verbal fluency test that assesses lexical retrieval, semantic network integrity, and executive control (switching between subcategories) (Isaacs and Kennie, 1973). It is widely used in dementia research due to its cross-lingual applicability and sensitivity to AD-related cognitive decline.
4. **Picture Description:** Subjects describe the picture of a “Mountain Scene” (shown in Figure 1¹). This elicits spontaneous descriptive discourse, combining visual-to-verbal mapping, semantic content generation, and syntax planning (Borod et al., 1980). Innovations in speech graph and semantic path modeling can

¹PARLO GmbH, www.parlo-institut.de

be used to derive features from such descriptions. The original picture is included in the corpus.

5. **Word Repetition “pataka”**: Subjects repeat the pseudo-word “pataka” ([ˈpataka]) as often as possible within 10 seconds. This tests phonological processing, speech-motor planning, and phonological loop functioning under minimal lexical interference.
6. **Word Repetition “sischafu”**: Analogous to the “pataka” task, subjects repeat another pseudo-word, “sischafu” ([sɪʃaˈfu]), within 10 seconds. The use of multiple pseudo-words reduces item-specific learning or familiarity, increasing the generalizability of phonological processing measures.
7. **Story Recall**: After previous distraction, subjects retell the story in their own words. This tests episodic memory, narrative coherence, and spontaneous language generation, capturing both semantic and syntactic retrieval processes.
8. **Picture Recall**: After previous distraction, subjects describe the previously shown picture from memory. This tests episodic memory, reconstruction of spatial and semantic elements, and narrative consistency under memory load.

3.3. Recordings

Task	ALL	HC	MCI	DEM
Confrontation Naming	204	81	58	65
Animal Naming	205	82	58	65
Subset Naming	204	81	58	65
Pataka	201	82	56	63
Sischafu	202	82	57	63
Subset Repetition	201	82	56	63
Picture Description	195	75	54	66
Picture Recall	200	81	56	63
Subset Picture	189	74	52	63
Story Reading	199	82	55	62
Story Recall	185	80	50	55
Subset Story	182	80	48	54
Subset All	171	73	45	53

Table 2: Number of samples for the PDC data subsets by diagnostic groups: healthy control (HC), mild cognitive impairment (MCI), dementia (DEM).

The audio recordings were collected using a standardized iPad-based application and the same iPad model across all study sites to ensure consistency. The recordings include the subjects’ speech and, where applicable, brief interventions by the examiner during the performance of the PDTB tasks. Participants were given sufficient time to complete the tasks in a quiet, controlled environment. All iPad data was stored in pseudonymized and encrypted form before being transferred for central processing. The corpus comprises ≈20 hours of raw audio material, with recording durations ranging from 4 seconds to 4.5 minutes, excluding empty or damaged files. The audio files were recorded as 16-bit mono waveforms with a sampling rate of 44.1 kHz. Since not all subjects completed all of the eight tasks, the corpus provides task-specific subsets (see Table 2).

3.4. Transcription

All speech recordings were manually transcribed by a professional transcription service according to the extended scientific transcription rules of Dresing & Pehl (Dresing and Pehl, 2010), preserving verbal and nonverbal features of spoken interaction.

The transcriptions are verbatim, capturing the structure of the spoken language while adapting colloquial speech and dialects to standard German. For example, reductions such as “so’n Zettel” are rendered as “so ein Zettel,” and “hamma” as “haben wir.” The transcription protocol includes the following paralinguistic and structural markers reflecting aspects of speech production and conversation organization:

- word and sentence breaks (as /)
- pauses in seconds (e.g., (.), (..), (...), (4), (5), etc.)
- filler words and backchannel signals (e.g., “ähm”, “hm”, “ja”, “aha”)
- emphases (as VERSALS)
- nonverbal utterances (e.g., (lacht), engl. (laughs))
- incomprehensible utterances (e.g., (unv.), (unv., with context), (guess?))
- speaker changes (as paragraphs beginning with A: or B: and ending with timestamps)
- speaker overlaps (as //)
- reception signals with intonation (e.g., “mhm (bejahend)”, engl. “mhm (affirmative)”)

The resulting transcripts form the ground truth to which ASR and NLP methods can be measured. Furthermore, the level of detail enables linguistic, acoustic, and interactional analyses of neurodegenerative speech features such as hesitation, repair phenomena, pause patterns, prosodic intonation, and turn-taking. A detailed description of the twelve transcription rules and their extensions can be found in (Dresing and Pehl, 2010).

3.5. Metadata

Short	Education Level	Years
education 1: school degree (SD)		
SDAL	A-Level	13
SDPS	Polytechnical School	12
SDSS	Secondary School	10
SDMS	Main Schooling	9
SDWO	Without	9
SDNC	No Comment	9
education 2: degree (D)		
DPHD	PhD	21
DUNI	University	17
DVTR	Vocational Training	= SD + 3
DWTR	Without Vocational Training	= SD

Table 3: Metadata abbreviations for educational degrees 1 and 2 with average years of education.

During the study visit, a detailed set of metadata was collected for each subject to ensure accurate characterization of the study population and adherence to the study protocol. General information includes demographic data such as age, gender, and educational level, as well as session parameters such as date, time, state, clinic ID, subject ID, investigator ID, and the iOS version of the iPad used. Educational level is reported as average years spent in education and the type of first and second educational degree (see Table 3). Verification of all inclusion and exclusion criteria listed in Section 3.1 was also documented for each subject. Diagnosis-related metadata includes the diagnostic group (HC, MCI, DEM) and the Mini Mental State Examination (MMSE) score. For MCI and DEM subjects, the NIA-AA criteria for the confirmed AD diagnosis were recorded, including values for the Clinical Dementia Rating (CDR) scale, biomarkers in cerebrospinal fluid (β -Amyloid and p-Tau), and diagnostic confirmation by imaging biomarkers (AD-related MRI T1/T2/FLAIR or PET). For more clinical details, refer to the original source of the NIA-AA criteria (Jack Jr. et al., 2018).

4. Directions of Research

The PARLO Dementia Corpus opens a broad range of research opportunities at the intersection of speech technology, linguistics, and clinical neuroscience. Its multi-center design, standardized task structure, and detailed clinical annotations make it a valuable resource for computer-based and clinical investigations of cognitive decline.

From a speech technology perspective, the corpus provides a benchmark for the development and evaluation of ASR systems for disordered and atypical speech. The recordings capture variations in speech fluency, articulation, and lexical access typical of MCI and dementia, enabling studies on robustness, domain adaptation, and bias mitigation in multilingual ASR systems. Beyond ASR, the data enable the extraction of acoustic (e.g., pause duration, prosodic variability, and voice quality) and linguistic biomarkers (e.g., lexical, syntactic, and semantic changes) associated with the progression of neurodegenerative diseases. The structured form of the PDTB allows for a systematic comparison of speech and language production under different cognitive loads, ranging from constrained naming to phonological production to open-ended narrative recall. This diversity enables the investigation of task-dependent and cross-task features, as well as multi-modal modeling approaches that combine text and audio signals.

From a clinical perspective, the corpus provides an empirical basis for the development and validation of automated screening tools for the early detection and monitoring of AD. By linking speech-based measurements with established clinical scales (e.g., MMSE) and biomarkers (e.g., cerebrospinal fluid), the dataset supports clinically interpretable modeling of cognitive impairments.

Beyond computational approaches, the PDC establishes a benchmark for speech-based dementia research in the German language. Its compatibility with established English corpora such as Pitt Corpus supports cross-lingual comparisons and contributes to the broader goal of developing reliable speech technologies for early cognitive screening.

To ensure the comparability and reproducibility of future studies using the PDC, we recommend standardized evaluation procedures that account for speaker (and site) variability, such as speaker-independent cross-validation (and leave-one-site-out validation).

5. Baseline Experimental Design

5.1. Automatic Speech Recognition

We automatically transcribe all speech recordings in the PDC using three state-of-the-art ASR

systems: `whisper-large-v3` by OpenAI (Radford et al., 2023), `parakeet-tdt-0.6b-v3` by NVIDIA, and `owsm_ctc_v4_1B` by Peng et al. (Peng et al., 2025). All models are openly available from HuggingFace²³⁴. The full raw audio recordings of the PDTB tasks (see Section 3.2) are transcribed with each ASR system using the default decoding parameters and the German language set. The word repetition tasks (i.e., “pataka” and “sischafu”) are excluded from ASR evaluation because their output requires phonetic rather than orthographic evaluation. Transcripts are normalized regarding case-sensitivity, punctuation, fillers, and numerical tokens prior to evaluation.

5.2. Automatic Test Evaluation

For the automatic evaluation of the standardized cognitive tests included in the PDTB, rule-based scoring is applied to the automatic transcriptions of the *Confrontation Naming* and *Animal Naming* tests. In the *Confrontation Naming* test, performance is evaluated as the number of correctly named objects among the 15 presented stimuli, while in the *Animal Naming* test, the score corresponds to the number of uniquely named animals within one minute. To compute these scores automatically, the ASR transcripts are normalized and compared against predefined task-specific lexical sets. The reference lists comprise the target object names (*Confrontation Naming*) and animal lexicon (*Animal Naming*), which are derived from GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) and the German OpenThesaurus⁵ for synonym and morphological variant matching. Word matches are identified using Levenshtein distance with a minimum threshold of 85% relative to word length. After filtering out duplicates, the word count represents the final score.

5.3. Large Language Models

For language modeling and classification experiments, we employ the open-source Vision Large Language Model (Vision-LLM) `Mistral-3.1`⁶, which is based on the Mistral 24B architecture. To enable efficient inference on a single GPU (NVIDIA L40S, 48 GB VRAM), the model is deployed using FP8 quantization and executed via the vLLM inference engine (Kwon et al., 2023). Although FP16 inference can yield slightly higher numerical precision, FP8 was chosen to bal-

ance computational efficiency and model usability. Model generation follows the parameters recommended by the authors (`temperature=0.15`, `max_tokens=4096`) to ensure reproducibility and prevent excessive output generation. To ensure consistent and structured output, all completions use constrained decoding based on a predefined JSON schema that specifies the output format. All prompts are written in German to match the input language, while English translations are provided in this article. A consistent system prompt is applied across all experiments:

“You are an expert in Alzheimer’s dementia who classifies the cognitive status of individuals based on their speech performance. You must return a JSON object that strictly follows this schema:”
+ JSON schema

This is followed by a task-specific user prompt containing a brief description of the cognitive task (e.g., picture description), the definition of the classification target (HC vs. MCI vs. DEM), and the corresponding transcription input from the subject’s task execution. We additionally input the source picture (Figure 1) for the *Picture Description* task and the ground truth text for the *Story Reading* task. Original prompts are given in the supplementary material A.1.

6. Results

6.1. ASR Performance

Automatic transcription quality for the ASR models described in Section 5.1 is evaluated based on the word error rate (WER) relative to normalized human reference transcriptions (see Section 3.4). The average WER (in %) is calculated per cognitive task and diagnostic group using normalized transcripts from `whisper-large-v3` (`whisper`), `parakeet-tdt-0.6b-v3` (`parakeet`), and `owsm_ctc_v4_1B` (`OWSM`), respectively.

The ASR results are summarized in Table 4. Across all tasks, `whisper` consistently achieves the lowest WER, followed by `parakeet` and `OWSM`. On average, `whisper` reaches an overall WER of 9–18% depending on the task, substantially outperforming the other models (typically 5–10 pp lower). Performance differences between the diagnostic groups are consistent across all models, with error rates increasing from healthy controls (HC) to mild cognitive impairment (MCI) to dementia (DEM). This trend reflects the increasing prevalence of speech disorders, pauses, and articulation irregularities in cognitively impaired speakers, which pose challenges for current end-to-end ASR systems. The task-specific speech production context

²www.hf.co/openai/whisper-large-v3

³www.hf.co/nvidia/parakeet-tdt-0.6b-v3

⁴www.hf.co/espnet/owsm_ctc_v4_1B

⁵www.openthesaurus.de

⁶www.hf.co/RedHatAI/Mistral-Small-3.1-24B-Instruct-2503-FP8-dynamic

Task	Group	whisper	OWSM	parakeet
Story Reading	HC	5.0	9.7	6.8
	MCI	8.2	14.2	9.4
	DEM	15.2	19.8	14.0
	ALL	9.0	14.1	9.7
Confrontation Naming	HC	10.9	26.8	11.1
	MCI	20.2	35.3	20.7
	DEM	23.6	38.5	23.8
	ALL	17.6	32.9	17.9
Animal Naming	HC	13.9	35.4	20.7
	MCI	21.6	39.8	22.9
	DEM	18.9	38.8	24.2
	ALL	17.7	37.7	22.4
Picture Description	HC	7.6	13.4	7.8
	MCI	10.5	15.9	10.4
	DEM	18.5	26.4	17.1
	ALL	12.1	18.5	11.7
Story Recall	HC	10.8	15.5	11.7
	MCI	17.7	23.3	19.1
	DEM	25.2	35.8	37.4
	ALL	16.9	23.7	21.4
Picture Recall	HC	8.4	12.1	10.9
	MCI	15.1	23.8	17.5
	DEM	16.9	28.8	21.5
	ALL	12.9	20.6	16.1

Table 4: Average Word Error Rate (in %) by diagnostic group (HC: healthy control, MCI: mild cognitive impairment, DEM: dementia) and cognitive task for `whisper-large-v3`, `parakeet-dt-0.6b-v3`, and `owsm_ctc_v4_1B`.

also strongly influences transcription accuracy. The lowest WERs are obtained for *Story Reading* and *Picture Description*, which involve structured and continuous speech. In contrast, spontaneous and memory-intensive tasks such as *Story Recall* and *Animal Naming* show markedly higher WERs (up to 25% for DEM). These increased error rates are partly attributable to the intrinsic task characteristics: Subjects with cognitive impairment often produce shorter, fragmentary responses, hesitations, and prolonged pauses, leading to higher WERs. OWSM exhibits the largest degradation under these conditions, suggesting lower robustness to non-native and disordered speech compared to `whisper`'s large multilingual training data coverage.

6.2. Automatic Test Evaluation

These experiments aim to automatically evaluate the standardized subtests of the PDTB, which are

the *Confrontation Naming* and the *Animal Naming* tasks, by applying the rule-based scoring approach from Section 5.2. We use transcriptions generated by `whisper`, which achieved the best ASR performance for both tasks (see Table 4). Scoring performance is measured using the Pearson correlation coefficient between the automatic scores derived from ASR transcripts and those derived from human reference transcripts. In addition, the calculated test scores are correlated with the expert-assigned MMSE scores to verify their diagnostic validity.

The results show nearly perfect correlations between ASR-based and reference-based scores, reaching a value of 0.92 for *Confrontation Naming* and 0.99 for *Animal Naming*. These findings confirm that automatic scoring based on ASR transcripts can reproduce the manual evaluation outcomes. As expected, the correlations with global MMSE scores are lower, at 0.43 for *Confrontation Naming* and 0.65 for *Animal Naming*. However, the combination (addition) of both subtest scores achieves a correlation of 0.70 with the MMSE, suggesting that the automatic evaluation of combined verbal tasks can capture clinically relevant deviations with global scales.

6.3. LLM Zero-Shot Classification

		Story Reading			+ Recall		
Actual	HC	50	30	0	50	23	7
	MCI	20	30	1	8	13	30
	DEM	12	44	4	2	4	54
		HC	MCI	DEM	HC	MCI	DEM
		Picture Description			+ Recall		
Actual	HC	70	9	1	70	9	1
	MCI	34	20	3	22	30	5
	DEM	17	38	9	3	24	37
		HC	MCI	DEM	HC	MCI	DEM
		Predicted			Predicted		

Figure 2: Confusion matrices for LLM zero-shot classification of healthy control (HC), mild cognitive impairment (MCI), and dementia (DEM) groups.

Our classification experiments aim to evaluate the ability of a Vision-LLM to determine the cognitive status of individuals in a zero-shot setting without task-specific fine-tuning. The objective is a 3-class classification that distinguishes between

healthy controls (HC), and individuals with AD-related mild cognitive impairment (MCI) and dementia (DEM). The speech samples are taken from the *Story* and *Picture* subsets of the corpus (see Table 2), including both stimulus tasks (*Story Reading*, *Picture Description*) and the corresponding recall tasks (*Story Recall*, *Picture Recall*). Automatic transcriptions are generated using `whisper`, which achieved the best overall ASR performance in the tasks (Table 4). Prompts and system configurations correspond to the settings described in Section 5.3. Two variants are tested: (1) baseline prompts using only the stimulus task, and (2) extended prompts that additionally include the corresponding recall transcription. Model performance is evaluated using the Unweighted Average Recall (UAR) to ensure equal weighting of all diagnostic classes.

Results are shown as confusion matrices in Figure 2. For both the story tasks and the picture tasks, the UAR increases significantly from 42.7% and 45.6% (stimulus only) to 59.3% and 66.0% when recall speech is included. These improvements indicate that incorporating recall data enhances discriminative performance, particularly by improving the model's sensitivity to intermediate impairment (MCI). As shown in Figure 2, MCI samples were frequently confused with neighboring classes in the baseline setting and misclassified as HC in *Story Reading* and as DEM in *Picture Description*. This may be because *Picture Description* is more cognitively demanding than *Story Reading*. Adding recall inputs reduces these confusions and yields a more balanced class distribution, reflecting the diagnostic significance of memory-driven, spontaneous speech for cognitive differentiation.

7. Conclusion

We introduced the PARLO Dementia Corpus (PDC), the first multi-center, clinically validated German resource addressing AD-related mild cognitive impairment and dementia. The corpus includes audio recordings of AD patients and healthy controls performing eight established neuropsychological tasks covering speech and language production under various cognitive loads, from constrained naming to phonological production to open-ended narrative recall. Enriched with detailed demographic, clinical and biomarker metadata, as well as human-verified transcriptions, the PDC provides a comprehensive resource for research into speech and cognition in neurodegenerative diseases. With its standardized data collection protocol and open evaluation framework, the corpus supports a wide range of research directions, including automatic speech recognition for pathological and atypical speech, extraction of acoustic and linguistic biomarkers, multi-

modal modeling (audio, text, medical biomarker), and automated cognitive assessment. Initial benchmark experiments demonstrate the feasibility of LLM-based zero-shot classification of cognitive impairment groups and the automatic evaluation of standardized dementia tests, providing a basis for future research.

8. Data Availability and Licensing

Researchers interested in obtaining access to the PDC are invited to contact PARLO at `info@parlo-institut.de` for commercial and non-commercial licensing. Requests will be reviewed in accordance with ethical and data protection requirements.

9. Bibliographical References

- Aparna Balagopalan, Benjamin Eyre, Jessica Robin, Frank Rudzicz, and Jekaterina Novikova. 2021. Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech. volume Volume 13 - 2021.
- Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. [To BERT or not to BERT: Comparing Speech and Language-Based Approaches for Alzheimer's Disease Detection](#). In *Interspeech 2020*, pages 2167–2171. ISCA.
- Benjamin Barrera-Altuna, Daeun Lee, Zaima Zarnaz, Jinyoung Han, and Seungbae Kim. 2024. [The interspeech 2024 taukadial challenge: Multilingual mild cognitive impairment detection with multimodal approach](#). In *Interspeech 2024*, pages 967–971.
- J. T. Becker, F. Boller, O. L. Lopez, J. Saxton, and K. L. McGonigle. 1994. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.
- Randa Ben Ammar and Yassine Ben Ayed. 2018. [Speech processing for early alzheimer disease diagnosis: Machine learning based approach](#). In *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8.
- Joan C. Borod, Harold Goodglass, and Edith Kaplan. 1980. [Normative data on the boston diagnostic aphasia examination, parietal lobe battery, and the boston naming Test](#). *Journal of Clinical Neuropsychology*, 2(3):209–215.

- Franziska Braun, Sebastian P. Bayerl, Florian Hönig, Hartmut Lehfeld, Thomas Hillemacher, Tobias Bocklet, and Korbinian Riedhammer. 2024. [Infusing acoustic pause context into text-based dementia assessment](#). In *Interspeech 2024*, pages 1980–1984.
- Franziska Braun, Sebastian P. Bayerl, Paula A. Pérez-Toro, Florian Hönig, Hartmut Lehfeld, Thomas Hillemacher, Elmar Nöth, Tobias Bocklet, and Korbinian Riedhammer. 2023. [Classifying Dementia in the Presence of Depression: A Cross-Corpus Study](#). In *Proc. INTERSPEECH 2023*, pages 2308–2312.
- Franziska Braun, Andreas Erzigkeit, Hartmut Lehfeld, Thomas Hillemacher, Korbinian Riedhammer, and Sebastian P. Bayerl. 2022a. Going Beyond the Cookie Theft Picture Test: Detecting Cognitive Impairments Using Acoustic Features. In *Text, Speech, and Dialogue*, pages 437–448. Springer International Publishing.
- Franziska Braun, Markus Förstel, Bastian Oppermann, Andreas Erzigkeit, Hartmut Lehfeld, Thomas Hillemacher, and Korbinian Riedhammer. 2022b. [Automated Evaluation of Standardized Dementia Screening Tests](#). In *Proc. Interspeech 2022*, pages 2478–2482.
- Laura Calzà, Gloria Gagliardi, Rema Rossini Favretti, and Fabio Tamburini. 2021. [Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia](#). *Computer Speech & Language*, 65:101113.
- Sofia de la Fuente Garcia, Craig W. Ritchie, and Saturnino Luz. 2020. Artificial intelligence, speech, and language processing approaches to monitoring alzheimer’s disease: A systematic review. *Journal of Alzheimer’s Disease*, 78(4):1547–1574.
- Thorsten Dresing and Thorsten Pehl. 2010. [Transkription](#). In Günter Mey and Katja Mruck, editors, *Handbuch Qualitative Forschung in der Psychologie*, pages 723–733. VS Verlag für Sozialwissenschaften, Wiesbaden, Germany.
- Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.
- Verena Henrich and Erhard Hinrichs. 2010. GernEdiT - the GermaNet editing tool. In *Proceedings of LREC*.
- B. Isaacs and A. T. Kennie. 1973. [The Set test as an aid to the detection of dementia in old people](#). *The British Journal of Psychiatry: The Journal of Mental Science*, 123(575):467–470.
- Clifford R. Jack Jr., David A. Bennett, Kaj Blennow, Maria C. Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M. Holtzman, William Jagust, Frank Jessen, Jason Karlawish, Enchi Liu, Jose Luis Molinuevo, Thomas Montine, Creighton Phelps, Katherine P. Rankin, Christopher C. Rowe, Philip Scheltens, Eric Siemers, Heather M. Snyder, Reisa Sperling, Contributors, Cerise Elliott, Eliezer Masliah, Laurie Ryan, and Nina Silverberg. 2018. Nia-aa research framework: Toward a biological definition of alzheimer’s disease. *Alzheimer’s & Dementia*, 14(4):535–562.
- EF Kaplan, H Goodglass, and S Weintraub. 1978. *The Boston Naming Test*. Veterans Administration Medical Center, Boston; MA.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Kaiying Lin and Peter Y. Washington. 2024. Multimodal deep learning for dementia classification using text and audio. *Scientific Reports*, 14(1):13887.
- Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and Brian MacWhinney. 2023. [Multilingual alzheimer’s dementia recognition through spontaneous speech: a signal processing grand challenge](#).
- Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. [Alzheimer’s Dementia Recognition Through Spontaneous Speech: The ADRess Challenge](#). In *Interspeech 2020*, pages 2172–2176. ISCA.
- Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021. [Detecting Cognitive Decline Using Speech Only: The ADRessSo Challenge](#). In *Interspeech 2021*, pages 3780–3784. ISCA.
- Israel Martínez-Nicolás, Thide E. Llorente, Francisco Martínez-Sánchez, and Juan José G. Meilán. 2021. Ten years of research on automatic voice and speech analysis of people with alzheimer’s disease and mild cognitive impairment: A systematic review article. *Frontiers in Psychology*, Volume 12 - 2021.

- J. C. Morris, R. C. Mohs, H. Rogers, G. Fillenbaum, and A. Heyman. 1988. Consortium to establish a registry for Alzheimer's disease (CERAD) clinical and neuropsychological assessment of Alzheimer's disease. *Psychopharmacology Bulletin*, 24(4):641–652.
- Raghavendra Pappagari, Jaejin Cho, Sonal Joshi, Laureano Moro-Velázquez, Piotr Żelasko, Jesús Villalba, and Najim Dehak. 2021. [Automatic detection and assessment of alzheimer disease using speech and language technologies in low-resource scenarios](#). In *Interspeech 2021*, pages 3825–3829.
- Yifan Peng, Muhammad Shakeel, Yui Sudo, William Chen, Jinchuan Tian, Chyi-Jiunn Lin, and Shinji Watanabe. 2025. [OWSM v4: Improving Open Whisper-Style Speech Models via Data Scaling and Cleaning](#). In *Interspeech 2025*, pages 2225–2229.
- Paula Andrea Pérez-Toro, Philipp Klumpp, Abner Hernandez, Tomas Arias, Patricia Lillo, Andrea Slachevsky, Adolfo Martín García, María Schuster, Andreas K. Maier, Elmar Noeth, and Juan Rafael Orozco-Arroyave. 2022. [Alzheimer's Detection from English to Spanish Using Acoustic and Linguistic Embeddings](#). In *Interspeech 2022*, pages 2483–2487. ISCA.
- Anna Pompili, Thomas Rolland, and Alberto Abad. 2020. [The inesc-id multi-modal system for the adress 2020 challenge](#). In *Interspeech 2020*, pages 2202–2206.
- Paula A. Pérez-Toro, Tomás Arias-Vergara, Franziska Braun, Florian Hönic, Carlos A. Tobón-Quintero, David Aguillón, Francisco Lopera, Liliana Hincapié-Henao, María Schuster, Korbinian Riedhammer, Andreas Maier, Elmar Nöth, and Juan Rafael Orozco-Arroyave. 2023. [Automatic assessment of alzheimer's across three languages using speech and language features](#). In *Interspeech 2023*, pages 1748–1752.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org.
- Vikram Ramanarayanan. 2024. Multimodal technologies for remote assessment of neurological and mental health. *Journal of Speech, Language, and Hearing Research*, 67(11):4233–4245.
- Morteza Rohanian, Julian Hough, and Matthew Purver. 2020. [Multi-modal fusion with gating using audio, lexical and disfluency features for alzheimer's dementia recognition from spontaneous speech](#). In *Interspeech 2020*, pages 2187–2191.
- Inês Vigo, Luis Coelho, and Sara Reis. 2022. Speech- and language-based classification of alzheimer's disease: A systematic review. *Bio-engineering*, 9(1).
- Jochen Weiner, Christian Herff, and Tanja Schultz. 2016. [Speech-based detection of alzheimer's disease in conversational german](#). In *Interspeech 2016*, pages 1938–1942.
- Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. 2020a. [Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease](#). In *Interspeech 2020*, pages 2162–2166. ISCA.
- Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. 2020b. [Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease](#). In *Interspeech 2020*, pages 2162–2166. ISCA.

10. Ethics Statement

All data included in the PARLO Dementia Corpus were collected in compliance with national and institutional ethical standards for medical research involving human participants. The study protocol was approved by the responsible IRBs at all participating clinical centers in Germany, and all participants (or their legal representatives) provided written informed consent prior to inclusion. Participants were informed about the study's aims, the nature of the speech recordings, and their right to withdraw consent at any time without negative consequences. All recordings were pseudonymized at the point of collection, and no personally identifiable information is contained in the released dataset. Data transfer, storage, and processing followed the principles of the EU General Data Protection Regulation. Access to the corpus requires a signed data-use and license agreement ensuring compliance with ethical and legal standards. The PARLO Dementia Corpus has been curated with particular attention to participant dignity, data security, and scientific transparency, aiming to enable reproducible and ethically responsible research on speech and cognitive health.

A. Appendix

A.1. Prompts

The following prompts (in German to match the input language) are used in the LLM zero-shot classification experiments from Section 5.3.

A unified system prompt with JSON schema:

```
"Sie sind ein Experte für Alzheimer's Demenz,
der den kognitiven Status von Personen anhand
ihrer Sprachleistung klassifiziert.
Sie müssen ein JSON-Objekt zurückgeben, das
sich strikt an dieses Schema hält:
{'properties':
{'cognitive_status':
{'enum': ['HC', 'MCI', 'DEM'],
'title': 'Cognitive Status',
'type': 'string'}},
'required': ['cognitive_status'],
'title': 'CognitiveAssessment',
'type': 'object'}
```

A prompt for the *Picture Description* task with the extension for *Picture Recall* task (in parentheses):

```
"Die Person beschreibt das gegebene Bild
(Bildbeschreibung).
Nach Ablenkung muss die Person das Bild
aus dem Gedächtnis beschreiben (Verzögerter
Recall).
Das Bild kann mit acht Konzepten beschrieben
werden:
- Eine Berglandschaft
- Wanderer suchen den Weg
- Eine Brücke führt über den Fluss
- Kinder sitzen auf einem Baumstamm
- Ein Junge ruft seine Eltern
- Ein Mädchen fällt ins Wasser
- Gewitterwolken mit Blitzen
- Ein Haus in der Ferne

Ihre Aufgabe darin, anhand der Bildbeschreibung
und des verzögerten Recalls den kognitiven
Status der Person in eine der drei Kategorien zu
klassifizieren:
- HC = Healthy Control
- MCI = Mild Cognitive Impairment
- DEM = Mild to Severe Dementia

Hier ist die Bildbeschreibung:
... {picture_description} ...

(Hier ist der verzögerte Recall:)
(... {picture_recall} ...)

Geben Sie die Bewertung NUR im JSON-
Format entsprechend des gegebenen Schemas
an. Keine anderen Ausgaben."
```

A prompt for the *Story Reading* task with the extension for *Story Recall* task (in parentheses):

```
"Die Person liest eine Geschichte vor (Vorlesung).
Nach Ablenkung erzählt die Person die
Geschichte aus dem Gedächtnis mit eige-
nen Worten nach (Verzögerter Recall).

Ihre Aufgabe besteht darin, anhand der
Vorlesung und des verzögerten Recalls den
kognitiven Status der Person in eine der drei
Kategorien zu klassifizieren:
- HC = Healthy Control
- MCI = Mild Cognitive Impairment
- DEM = Mild to Severe Dementia

Hier ist die Vorlesung:
... {story_reading} ...

(Hier ist der verzögerte Recall:)
(... {story_recall} ...)

Geben Sie die Bewertung NUR im JSON-
Format entsprechend des gegebenen Schemas
an. Keine anderen Ausgaben."
```