Diagnosis of dementia in the early stages of medical preventive research

A Preprint

Dmitry S.Slobodin*
Computational Mathematics and Cybernetics
Lomonosov Moscow State University
Moscow, Russia
`email@domain`

Oleg V.Senko†
Computational Mathematics and Cybernetics
Lomonosov Moscow State University
Moscow, Russia
`email@domain`

Abstract

This study investigates the comparative effectiveness of two speech tasks—standardized text reading and spontaneous picture description—for early dementia screening. While both approaches leverage acoustic biomarkers in speech, their relative performance and clinical utility remain unclear. We conducted a systematic comparison using parallel datasets where the same participants completed both tasks under identical recording conditions. Acoustic features were extracted using the eGeMAPS parameter set and evaluated through multiple machine learning models for regression and classification of established clinical scales (MMSE, MoCA, CDR). Contrary to initial expectations, spontaneous picture description consistently outperformed standardized reading across most evaluation metrics, achieving higher accuracy in differentiating between healthy individuals, those with mild cognitive impairment, and early-stage dementia patients. These findings provide empirical evidence for task selection in clinical speech-based screening protocols and demonstrate the superior sensitivity of natural language production for capturing cognitive-linguistic deficits associated with dementia progression.

Keywords Early dementia detection · Speech-based screening · Standardized reading task

1 Introduction

Dementia screening requires tools that are fast, low-cost, and deployable in everyday clinics where non-specialists see patients first. Speech offers a noninvasive digital biomarker that can be captured in minutes with ubiquitous hardware, enabling a triage workflow: flag suspicious cases during a routine visit and refer them to specialists for definitive assessment. Earlier detection improves care planning, supports timely interventions, and can reduce downstream healthcare costs, while remote collection enables longitudinal monitoring between visits.

Two primary speech paradigms dominate dementia screening research: standardized reading tasks and spontaneous description tasks. Each approach offers distinct advantages and limitations, yet their comparative effectiveness remains inadequately explored. Standardized reading controls linguistic content, potentially reducing variability and focusing on acoustic-prosodic markers of motor speech control. Spontaneous description engages natural language production, potentially capturing broader cognitive-linguistic processes but introducing greater variability.

Prior research predominantly analyzes spontaneous speech tasks such as picture description or interviews, combining acoustic–prosodic features (e.g., eGeMAPS/ComParE, MFCCs, pause and tempo statistics, F0 variability) with linguistic features extracted from ASR transcripts (lexical richness, syntactic complexity,

---

*Student Research Project (NIR).
†Supervisor; DSc (Phys.–Math.), Professor.

semantic coherence, and disfluencies) (Luz et al., 2020, 2021; de la Fuente Garcia et al., 2020; Fraser et al., 2016; Karlekar et al., 2018; Warnita et al., 2018; Eyben et al., 2016). Challenge datasets that control for confounders (e.g., ADReSS/ADReSSo) report around 75–80

However, spontaneous speech poses several challenges that limit clinical translation. High linguistic variability and topic effects can lead models to learn content rather than cognitive markers, reducing cross-domain and cross-language generalization (de la Fuente Garcia et al., 2020; Luz et al., 2020). Expert-labeled corpora are small, and imperfect validation (e.g., speaker leakage) inflates reported performance; dependence on ASR quality introduces variable noise by accent, recording, and pathology, precisely where robust signals are needed (de la Fuente Garcia et al., 2020; Luz et al., 2021). Channel and demographic confounders (age, sex, microphone) can further bias results if not explicitly controlled, and calibration and interpretability are often underreported (Luz et al., 2020, 2021; de la Fuente Garcia et al., 2020).

Standardized reading tasks offer potential solutions to these challenges by fixing lexical and syntactic content, reducing variability and ASR dependence. This allows measurement of acoustic–prosodic, timing, and pronunciation markers that more directly reflect motor–cognitive control. Audio can be aligned to canonical text via forced alignment to compute precise word/phoneme durations, speech versus articulation rate, silent and filled pause statistics and their positions, distributions of vowel/consonant durations (including VOT), as well as jitter/shimmer and F0 dynamics (Eyben et al., 2016; McAuliffe et al., 2017; Baevski et al., 2020).

This study addresses a critical gap in the literature by systematically comparing both paradigms using identical participant cohorts, acoustic feature sets, and machine learning pipelines. We evaluate their relative effectiveness for predicting multiple clinical scales (MMSE, MoCA, CDR) through both regression and classification frameworks. Our comparative approach enables direct assessment of each task type's strengths and limitations for clinical deployment.

Our contributions are as follows.

- Comparative framework: We develop and validate a systematic methodology for comparing standardized reading and spontaneous description tasks using parallel datasets from the same participant cohort.

- Feature analysis: We identify the most discriminative acoustic features for each task type using identical eGeMAPS parameter sets and statistical validation procedures.

- Multi-scale validation: We evaluate both regression (predicting raw clinical scores) and classification (cognitive status detection) performance across three established clinical instruments.

- Clinical insights: We provide empirical evidence for task selection in clinical practice, identifying contexts where each paradigm offers optimal sensitivity and specificity.

- Methodological transparency: We release reproducible pipelines for both task types to facilitate cross-study comparisons and clinical translation of speech-based dementia screening.

2 Related Work

A large body of research has investigated speech as a biomarker for cognitive impairment, with most studies focusing on either spontaneous speech tasks or standardized protocols, but rarely comparing them directly. Understanding the relative strengths of each approach is crucial for developing optimized clinical screening tools.

2.1 Spontaneous Speech Paradigms

The dominant approach in computational dementia detection has been spontaneous speech tasks, particularly picture description. Community challenges ADReSS and ADReSSo curated balanced subsets of DementiaBank, controlling for age, sex, and recording channel, and established rigorous speaker-held-out evaluations (Luz et al., 2020, 2021). Systems in these challenges typically fused acoustic–prosodic features (eGeMAPS/ComParE, MFCCs, pitch and energy statistics, pause and tempo measures) with linguistic features obtained from ASR, such as lexical diversity, syntactic complexity, semantic coherence, and disfluency counts, achieving about 75–80% accuracy for Alzheimer's vs. control and moderate error for MMSE regression (Luz et al., 2020, 2021). These results highlighted that temporal structure—speech rate, articulation rate, and pause distributions—carries strong signal for dementia detection.

Hand-crafted acoustic feature sets have served as robust baselines across paralinguistic tasks. The Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) was designed for physiological interpretability and low overfitting risk and has repeatedly delivered strong performance in health and affective computing (Eyben et al., 2016). Beyond low-level descriptors, research has emphasized prosodic and timing markers: ratio and distribution of silence, inter-word pauses, maximum pause length, variability of F0, jitter and shimmer, and stability of articulation rate (de la Fuente Garcia et al., 2020; Luz et al., 2017). Such features are particularly attractive in clinical contexts due to interpretability and hardware robustness.

Linguistic approaches exploit the idea that neurodegeneration affects lexical access, syntactic planning, and semantic coherence. Classical models employ hundreds of features spanning type–token ratios, part-of-speech patterns, parse-tree depth, and discourse measures (Fraser et al., 2016). Neural NLP architectures (CNN/LSTM/Transformers) trained on transcripts have matched or surpassed feature-engineered baselines, while interpretation studies point to simplified syntax, increased repetitions, and self-repairs as discriminative cues (Karlekar et al., 2018). However, transcript-based methods hinge on ASR quality, which varies with accent, noise, and pathology; this dependency limits portability across settings and languages (de la Fuente Garcia et al., 2020).

2.2   Standardized Speech Tasks

Standardized speech tasks have a long clinical tradition but have received less computational attention than spontaneous speech. Verbal fluency tests quantify clustering and switching behavior and predict outcomes in MCI, linking detailed temporal patterns to brain structure (Pakhomov et al., 2016). Reading aloud provides a complementary standardized probe: by fixing the lexical and syntactic content, it isolates motor–prosodic and planning components of speech production while enabling precise measurement through forced alignment (McAuliffe et al., 2017). Prior work suggests that alignment-based timing measures (word/phoneme durations, pause positions, vowel and stop consonant timing, VOT) and prosodic stability (F0 dynamics, jitter, shimmer) are sensitive to cognitive decline, and their interpretability facilitates clinician trust (de la Fuente Garcia et al., 2020; Luz et al., 2017).

The theoretical advantages of standardized reading include reduced linguistic variability, elimination of ASR dependency, and improved cross-site comparability. By controlling lexical content, reading tasks potentially isolate motor-speech and prosodic markers from linguistic planning deficits. However, the relative diagnostic sensitivity of reading versus spontaneous speech remains poorly characterized in the literature.

2.3   Methodological Advances and Clinical Translation

End-to-end acoustic models and self-supervised speech representations have gained traction for low-resource medical audio. CNN/LSTM architectures on log-mel or MFCC inputs demonstrated competitive accuracy on DementiaBank (Warnita et al., 2018). More recently, self-supervised models like wav2vec 2.0 and HuBERT provide powerful embeddings that improve over traditional features and mitigate data scarcity, especially when paired with simple classifiers and strong regularization (Baevski et al., 2020; Hsu et al., 2021). Nonetheless, their black-box nature complicates clinical interpretation, and domain shifts (microphone, room acoustics) still degrade performance without explicit adaptation.

Clinical realism motivates studies beyond picture description, including dialogues in memory clinics, telephone calls, and multilingual corpora. Work on real-world clinical conversations has shown that turn-taking behavior, response latency, and dialogue-level timing offer discriminative signals, with moderate-to-high AUCs under realistic noise and topic variability (Mirheidari et al., 2019). Cross-language studies report that pause structure and articulation-related markers are relatively language-agnostic compared to lexical features, but strict speaker-held-out splits and confound control are essential for reliable estimates (Gosztolya et al., 2019; de la Fuente Garcia et al., 2020). Across this literature, common limitations include small labeled datasets, potential leakage between training and test speakers or sessions, and underreporting of calibration and robustness analyses.

2.4   Research Gap and Our Contribution

While both spontaneous and standardized speech tasks show promise for dementia detection, few studies have directly compared their relative effectiveness using identical participant cohorts, feature sets, and evaluation frameworks. This gap is significant because task selection has major implications for clinical implementation: spontaneous tasks may capture richer cognitive-linguistic markers but introduce variability, while standardized tasks offer better control but potentially miss important diagnostic information.

Our study addresses this gap through systematic comparison of standardized reading and spontaneous picture description using parallel datasets from the same participants. We employ identical acoustic feature extraction (eGeMAPS), machine learning pipelines, and evaluation metrics to provide direct evidence for task selection in clinical screening contexts. This comparative approach enables us to identify the specific contexts where each paradigm offers optimal sensitivity and practical advantages for real-world deployment.

3 Problem Formulation and Method

We conduct a comparative study of early dementia assessment using two distinct speech modalities: standardized reading and spontaneous picture description. Our framework enables direct comparison of these approaches using identical participant cohorts, feature extraction methods, and evaluation protocols. The study addresses both regression (predicting continuous clinical scores) and classification (cognitive status detection) tasks to provide comprehensive insights into each modality's diagnostic utility.

3.1 Problem Formulation

Let $\mathcal{D}_R = \{(x_i^R, y_i, c_i)\}_{i=1}^n$ represent the reading task dataset, where x_i^R is the audio waveform of participant i reading a standardized passage, $y_i \in \mathbb{R}$ is the target clinical score (MMSE, MoCA, or CDR), and c_i denotes confounders (age, sex, recording device). Similarly, let $\mathcal{D}_D = \{(x_i^D, y_i, c_i)\}_{i=1}^n$ represent the picture description dataset from the same participants. We assume speaker-level independence with joint distribution $P(X, Y, C)$ and employ strict speaker-disjoint splits for evaluation.

For each modality, we construct feature mappings Φ_R and Φ_D that extract comprehensive acoustic-prosodic representations. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) serves as our primary feature space, providing 88 clinically interpretable acoustic parameters across frequency, amplitude, spectral, and temporal domains (Eyben et al., 2016). The per-recording feature vectors are defined as:

$$z_i^R = \Phi_R(x_i^R) = A\left(\Phi_{\text{eGeMAPS}}(x_i^R)\right) \in \mathbb{R}^{88}$$

$$z_i^D = \Phi_D(x_i^D) = A\left(\Phi_{\text{eGeMAPS}}(x_i^D)\right) \in \mathbb{R}^{88}$$

where A aggregates frame-level acoustic descriptors using statistical functionals (means, percentiles, variances, extremes). This identical feature extraction ensures direct comparability between modalities.

The prediction functions for each modality are:

$$f_R = h_R \circ \Phi_R : \mathcal{X}_R \to \mathbb{R}, \quad f_D = h_D \circ \Phi_D : \mathcal{X}_D \to \mathbb{R}$$

with h_R and h_D being machine learning models trained separately for each task type.

3.2 Evaluation Framework

We employ a comprehensive evaluation strategy assessing both regression and classification performance:

3.2.1 Regression Evaluation

For clinical score prediction, we report speaker-held-out estimates of:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|,$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Higher R^2 (maximum 1) and lower MAE/RMSE indicate better performance, with negative R^2 indicating the model underperforms the mean predictor.

3.2.2   Classification Evaluation

For binary cognitive status detection, we evaluate:

- ROC AUC: Area under Receiver Operating Characteristic curve
- Precision: $\frac{TP}{TP+FP}$
- Recall: $\frac{TP}{TP+FN}$
- F1-score: $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

3.3   Machine Learning Methodology

We compare four model classes for both regression and classification:

3.3.1   Gradient Boosting (CatBoost)

Our primary model class employs gradient boosting on decision trees, which builds an additive ensemble through iterative refinement:

$$F_m(x) = F_{m-1}(x) + \nu \cdot \gamma_m \cdot h_m(x)$$

where h_m is a weak learner fitted to residuals from F_{m-1}, $\nu \in (0, 1]$ is the learning rate, and γ_m is the optimal step size. CatBoost specifically handles categorical features and reduces prediction shift through ordered boosting.

3.3.2   Random Forest

Ensemble method combining multiple decorrelated decision trees using bagging and random feature selection, providing robust performance and feature importance estimates.

3.3.3   Decision Trees

Single decision tree models serving as interpretable baselines, though prone to overfitting.

3.3.4   Linear Models

Linear Regression for continuous outcomes and Logistic Regression for classification tasks, providing simple, interpretable baselines.

The empirical risk minimized during training is the mean squared error (regression) or cross-entropy (classification) with regularization:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(h_{\theta}(z_i), y_i) + \lambda \Omega(\theta)$$

3.4   Experimental Pipeline

3.4.1   Data Preprocessing

- Audio normalization: Peak amplitude normalization to -1 dB FS
- Voice activity detection: Preserve authentic pause structure while removing non-speech segments
- Quality control: Manual verification of audio integrity and transcription accuracy
- Confounder recording: Systematic documentation of age, sex, recording device, and environmental conditions

3.4.2   Feature Extraction

We extract the complete eGeMAPS feature set comprising:

- Frequency parameters: F0 semitone range, formant frequencies (F1-F3), harmonic-to-noise ratio
- Amplitude/energy parameters: Loudness percentiles (20th, 50th, 80th), shimmer, jitter

- Spectral parameters: Spectral flux, slope, centroid, entropy, sharpness
- Temporal parameters: Speech rate, pause duration statistics, voice activity ratio

All features are z-score standardized within training folds to prevent data leakage.

3.4.3 Model Training and Selection

- Cross-validation: Leave-one-speaker-out (LOSO) validation ensuring strict speaker independence
- Hyperparameter tuning: Grid search over comprehensive parameter spaces:
  - CatBoost: `n_estimators` $\in \{200, 400, 800\}$, `learning_rate` $\in \{0.01, 0.05, 0.1\}$, `max_depth` $\in \{2, 3, 4\}$
  - Random Forest: `n_estimators` $\in \{100, 200, 500\}$, `max_features` $\in \{\text{sqrt}, \log2\}$
  - Regularization: `reg_alpha`, `reg_lambda` $\in \{0, 10^{-3}, 10^{-2}, 10^{-1}\}$
- Early stopping: Based on validation R^2 (regression) or AUC (classification) with 50-epoch patience

3.4.4 Statistical Analysis

- Feature significance: Mann-Whitney U-tests with Benjamin-Hochberg correction for multiple testing
- Confidence intervals: Bootstrap sampling (1000 iterations) for performance metrics
- Modality comparison: Paired statistical tests (Wilcoxon signed-rank) for cross-modal performance differences

3.4.5 Interpretability and Robustness

- Feature importance: Gain-based importance (tree models) and SHAP values for model interpretation
- Ablation studies: Systematic evaluation of feature category contributions
- Confounder analysis: Partial correlation analysis to isolate cognitive effects from demographic influences
- Cross-modality feature analysis: Identification of modality-specific versus generalizable biomarkers

This comprehensive methodological framework enables rigorous comparison of reading and description tasks while maintaining clinical interpretability and statistical robustness. The identical treatment of both modalities ensures that performance differences reflect genuine task characteristics rather than methodological inconsistencies.

4 Experiments

4.1 Experimental Setup

We conducted a comprehensive comparative analysis of two speech modalities for dementia screening, using parallel datasets from the same 95 participants to ensure direct comparability:

- Dataset R (Standardized Reading): Participants read identical literary passages, controlling linguistic content to isolate acoustic-prosodic and timing features while minimizing lexical and syntactic variability.
- Dataset D (Spontaneous Description): Participants described complex visual scenes, capturing natural language production with inherent variability in lexical choice, syntactic complexity, and discourse organization.

This dual-task design enabled us to test our initial hypothesis that standardized reading would provide superior diagnostic performance due to reduced variability, while simultaneously evaluating the alternative possibility that spontaneous speech might capture richer cognitive-linguistic markers of impairment.

For both modalities, we conducted exhaustive experiments spanning regression (predicting continuous MMSE, MoCA, and CDR scores) and binary classification (distinguishing cognitively impaired from healthy participants). All evaluations employed leave-one-out cross-validation with strict speaker-level separation, comparing four model families:

1. CatBoost: Gradient boosting with categorical feature handling
2. Random Forest: Ensemble of decorrelated decision trees
3. Decision Tree: Single tree models for interpretability
4. Linear/Logistic Regression: Linear baselines for performance benchmarking

Evaluation metrics included R^2, RMSE, and MAE for regression tasks, and ROC AUC, precision, recall, and F1-score for classification, providing comprehensive assessment across multiple performance dimensions.

4.2 Feature Analysis and Biomarker Discovery

Our feature extraction employed the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), capturing 88 clinically interpretable parameters across four domains:

- Frequency characteristics: Fundamental frequency dynamics, formant structures, voicing patterns
- Amplitude-energy profiles: Loudness distributions, intensity contours, energy fluctuations
- Spectral properties: Harmonic-to-noise ratios, spectral flux, formant dispersion
- Temporal patterns: Pause structures, speech rhythm, articulation timing

Statistical analysis using Mann-Whitney U-tests with Benjamin-Hochberg correction revealed distinct feature significance patterns across modalities (Table 1).

Table 1: Most significant eGeMAPS features for dementia vs. MCI classification across speech modalities

Feature	Description	p-value
Spontaneous Description (Dataset D)		
loudness_sma3_percentile50.0	50th percentile of loudness	8.55E-05
loudness_sma3_percentile20.0	20th percentile of loudness	9.32E-05
spectralFluxUV_sma3nz_amean	Spectral flux on unvoiced segments	0.00038
loudness_sma3_stddevNorm	Normalized loudness std. deviation	0.00043
mfcc3_sma3_stddevNorm	Normalized MFCC3 std. deviation	0.00061
Standardized Reading (Dataset R)		
loudness_sma3_percentile20.0	20th percentile of loudness	7.33E-05
loudnessPeaksPerSec	Loudness peaks per second	0.00062
mfcc3V_sma3nz_stddevNorm	MFCC3 std. deviation (voiced)	0.00158
loudness_sma3_meanFallingSlope	Mean falling slope of loudness	0.00164
loudness_sma3_stddevNorm	Normalized loudness std. deviation	0.00197

The feature analysis yielded several critical insights. First, loudness distribution parameters emerged as consistently discriminative across both modalities, suggesting that vocal intensity control represents a core deficit in cognitive impairment that transcends task demands. Second, spontaneous description revealed richer spectral signatures (spectral flux on unvoiced segments), potentially reflecting more varied articulatory gestures and breath control patterns. Third, reading tasks showed greater sensitivity to dynamic loudness changes (slopes, peaks), possibly indicating challenges in prosodic modulation during constrained speech production.

4.3 Regression Performance: Clinical Score Prediction

Table 2 presents the comprehensive regression results, revealing nuanced performance patterns across modalities and clinical scales.

Contrary to our initial hypothesis, spontaneous description demonstrated superior overall regression performance, particularly for MoCA and CDR prediction where it achieved substantially higher R^2 values (0.14 and 0.15 vs. -0.04 and 0.04 for reading). This pattern suggests that spontaneous speech captures cognitive-linguistic processes more aligned with these clinical instruments' assessment domains.

The modality-performance relationship varied by clinical scale. For MMSE, reading tasks showed a slight advantage ($R^2 = 0.15$ vs. 0.07), possibly because MMSE emphasizes orientation and basic attention, which

Table 2: Comparative LOO regression performance across speech modalities and clinical scales

Modality	Scale	Best Model	RMSE	R^2	MAE	n
Reading (R)	MMSE	CatBoost	3.36	0.15	2.35	95
Reading (R)	MoCA	CatBoost	4.93	-0.04	3.78	90
Reading (R)	CDR	CatBoost	0.42	0.04	0.31	95
Description (D)	MMSE	CatBoost	3.50	0.07	2.54	95
Description (D)	MoCA	Random Forest	4.50	0.14	3.60	90
Description (D)	CDR	Random Forest	0.40	0.15	0.28	95

may be reflected in structured task performance. However, for MoCA and CDR—scales that assess executive function, complex attention, and functional impairment—spontaneous description proved markedly more predictive.

The consistent underperformance of linear models across both modalities ($R^2 < 0$ in most cases) highlights the complex, non-linear relationships between acoustic features and cognitive status, reinforcing the value of tree-based ensembles for this domain.

4.4 Classification Performance: Cognitive Status Detection

Binary classification results (Table 3) revealed even more pronounced modality differences, with spontaneous description demonstrating superior robustness across multiple evaluation metrics.

Table 3: Comparative LOO classification performance for cognitive status detection

Modality	Task	Best Model	ROC AUC	Precision	Recall	F1
Reading (R)	t_MMSE	Logistic Reg.	0.79	0.42	0.50	0.46
Reading (R)	t_MoCA	CatBoost	0.39	0.73	0.89	0.80
Reading (R)	t_CDR	Logistic Reg.	0.77	0.46	0.52	0.49
Description (D)	t_MMSE	Logistic Reg.	0.77	0.35	0.38	0.36
Description (D)	t_MoCA	Random Forest	0.60	0.74	0.98	0.84
Description (D)	t_CDR	Random Forest	0.74	0.67	0.26	0.38

The classification results revealed several important patterns. First, MoCA-based classification achieved the highest F1-scores across both modalities (0.80-0.84), suggesting that acoustic features capture processes particularly relevant to the cognitive domains assessed by MoCA. Second, the dramatic performance difference in t_MoCA classification (ROC AUC 0.39 for reading vs. 0.60 for description) underscores spontaneous speech's superior sensitivity to mild cognitive impairment.

Notably, while reading tasks showed competitive ROC AUC values for t_MMSE and t_CDR classification (0.79 and 0.77), they suffered from precision-recall tradeoffs that reduced their practical utility. In contrast, spontaneous description models demonstrated more balanced performance characteristics, particularly for the clinically crucial task of MCI detection.

4.5 Comparative Analysis and Clinical Implications

Our comprehensive comparison yielded several key findings that challenge initial assumptions and inform clinical implementation:

Modality-Specific Strengths: Rather than one modality universally dominating, each demonstrated distinct advantages. Standardized reading showed value for basic cognitive screening (MMSE-based tasks) and provided cleaner acoustic signals for fundamental prosodic analysis. Spontaneous description excelled for detecting subtle cognitive-linguistic impairments (MoCA-based tasks) and offered richer feature signatures for complex cognitive assessment.

Feature Interpretability: The consistent prominence of loudness-related features across both modalities suggests vocal intensity control as a fundamental marker of cognitive-motor integration deficits. However,

spontaneous speech revealed additional spectral and temporal features that may reflect higher-order cognitive processes like lexical access, syntactic planning, and discourse organization.

Clinical Scale Alignment: The varying modality performance across clinical scales indicates that task selection should consider the specific cognitive domains of interest. For executive function and complex attention assessment (MoCA), spontaneous description is clearly superior. For basic cognitive screening (MMSE), standardized reading offers practical advantages.

Implementation Considerations: While spontaneous description demonstrated superior diagnostic performance overall, standardized reading retains important practical benefits including reduced administration time, eliminated ASR dependency, and improved cross-site standardization. These factors remain crucial for large-scale screening implementation.

The performance patterns observed suggest that spontaneous speech engages a broader range of cognitive processes—including lexical retrieval, syntactic formulation, and discourse planning—that are particularly vulnerable in early cognitive decline. This cognitive-linguistic complexity, while introducing variability, appears to provide richer diagnostic information than the more constrained motor-speech processes emphasized in reading tasks.

Future work should explore hybrid approaches that leverage both modalities' complementary strengths, potentially using standardized reading for initial screening and spontaneous description for more detailed assessment of suspicious cases. Additionally, the development of modality-specific normative databases could enhance the clinical utility of both approaches by accounting for their inherent performance characteristics.

4.6 Figures

[10] See Figure ??. Here is how you add footnotes. [3] [11]

4.7 Tables

See awesome Table 4.

The documentation for `booktabs` ('Publication quality tables in LaTeX') is available from:

$$https://www.ctan.org/pkg/booktabs$$

4.8 Lists

- Lorem ipsum dolor sit amet
- consectetur adipiscing elit.
- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

References

S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. In Proc. Interspeech, 2020. URL https://arxiv.org/abs/2004.06833.

S. Luz, F. Haider, S. de la Fuente Garcia, et al. Detecting cognitive decline using speech only: The adresso challenge. medRxiv, 2021. doi: 10.1101/2021.03.24.21254263. URL https://www.medrxiv.org/content/10.1101/2021.03.24.21254263v2.

S. de la Fuente Garcia, C. W. Ritchie, and S. Luz. Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: A systematic review. Journal of Alzheimer's Disease, 2020. doi: 10.3233/JAD-200888. URL https://doi.org/10.3233/JAD-200888.

K. C. Fraser, J. A. Meltzer, and F. Rudzicz. Linguistic features identify alzheimer's disease in narrative speech. Journal of Alzheimer's Disease, 49(2):407–422, 2016. doi: 10.3233/JAD-160532. URL https://content.iospress.com/articles/journal-of-alzheimers-disease/jad160532.

J. Karlekar, T. Niu, and M. Bansal. Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models, 2018. URL https://arxiv.org/abs/1804.06440.

---

[3]Sample of the first footnote.

T. A. Warnita, Y. Inoue, and K. Shinoda. Detecting alzheimer's disease using gated convolutional neural network and lstm, 2018. URL `https://arxiv.org/abs/1803.11344`.

F. Eyben, K. R. Scherer, B. W. Schuller, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2):190–202, 2016. doi: 10.1109/TAFFC.2015.7160715. URL `https://ieeexplore.ieee.org/document/7160715`.

M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In Proc. Interspeech, 2017. doi: 10.21437/Interspeech.2017-1386. URL `https://montreal-forced-aligner.readthedocs.io/`.

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems (NeurIPS), 2020. URL `https://arxiv.org/abs/2006.11477`.

S. Luz et al. Longitudinal monitoring and detection of alzheimer's type dementia from spontaneous speech data, 2017. URL `https://ieeexplore.ieee.org/document/8104154`. IEEE Xplore.

S. V. Pakhomov, S. E. Marino, et al. Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. Alzheimer's Dementia: Diagnosis, Assessment Disease Monitoring, 2016. URL `https://www.sciencedirect.com/science/article/pii/S2352872916000051`.

W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. In Proc. ICASSP, 2021. URL `https://arxiv.org/abs/2106.07447`.

B. Mirheidari, D. Blackburn, M. Reuber, et al. Detecting cognitive impairment by machine learning analysis of conversations in memory clinics. PLOS ONE, 2019. URL `https://journals.plos.org/plosone/`.

G. Gosztolya, L. Tóth, M. Pákáski, I. Hoffmann, J. Kálmán, et al. Identifying mild cognitive impairment and dementia by acoustic and lexical features of spontaneous speech. In Proc. Interspeech, 2019. URL `https://www.isca-speech.org/archive/interspeech_2019/`.