Mild Cognitive Impairment Detection via Linear Discriminant Analysis of Picture Description Speech Features: A Cross Corpus Comparison

Yan-Lin Lai*, Erh-Yun Chang*, Yi-Wen Liu*, Jung Lung Hsu†, and Hui-Chuan Hsu‡
* Dept. Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan
E-mail: walterlai@gapp.nthu.edu.tw; jk10171213@gmail.com; ywliu@ee.nthu.edu.tw
† Dept. Neurology, New Taipei Municipal TuCheng Hospital (Build and Operated by Chang Gung Medical Foundation), New Taipei city, Taiwan
E-mail: tulu@ms36.hinet.net
‡ School of Public Health, Taipei Medical University, Taipei, Taiwan
E-mail: gingerhsu@tmu.edu.tw

***Abstract*—This study aims to investigate whether mild cognitive impairment (MCI) can be detected by analyzing the speech produced by elderly participants when they attempt to describe commonly used cartoon pictures in the field of dementia study. We collected and transcribed a mixed Mandarin and Southern Min speech dataset from elderly subjects in Taiwan. In addition, an English dataset was retrieved from DementiaBank for comparison purposes. Then, prosodic and linguistic features in six categories were computed – including acoustic, speaking rate, filled pause, part-of-speech, syntactic, and lexical richness features. Linear discriminant analysis (LDA) was applied to each of the six feature spaces to examine the data separability between MCI and the healthy control (HC) group. Subsequently, we selected the most discriminative feature categories for both datasets and trained support vector machine (SVM) classifiers on LDA-projected features. Performance of the classifiers is evaluated in terms of F1 score and accuracy via 3-fold cross validation. We discovered that the discriminative power of acoustic and linguistic features generally differ across the two datasets. This creates a challenge for early detection of MCI, and we suggest that the best classifier should be designed in language-specific manners.**

I. Introduction

Dementia is a progressive and currently incurable neurological disorder that primarily affects older adults. With the global aging population, the prevalence of dementia is expected to increase significantly, posing a major public health challenge. Caring for people with dementia often places a heavy emotional and logistical burden on families and caregivers. Early detection is crucial because medical intervention can help slow the progression of the disease and mitigate cognitive decline [1]. However, early diagnosis remains difficult due to patients' limited self-awareness, the reluctance of many older adults to seek medical attention, and constraints in healthcare accessibility [2].

In this research, we aim to explore the feasibility of detecting early signs of dementia by analyzing speech while cognitive impairment is still mild. This research goal is motivated by two general observations – First, signs of dementia can often be found in deterioration of speech usages [3]. Secondly, speech data acquisition is convenient and affordable, and its imposes relatively low privacy risk compared to image and video recordings. Thus, we argue that speech-based mild cognitive impairment (MCI) detection has potentials for fast dementia screening and deployment to the home.

II. Related Work

Voleti et al. [4] conducted a comprehensive review of speech and language features used in cognitive-linguistic assessment for MCI detection. They categorized commonly used features into audio-based and text-based domains, including prosodic features, articulation, pause-related measures, lexical diversity, syntactic complexity, and semantic coherence. The review work provided a systematic foundation for selecting and organizing multimodal features in our study.

Roark et al. [5] demonstrated that syntactic and timing-related features extracted from spoken narrative recall tasks could effectively differentiate individuals with MCI from healthy controls. In particular, they introduced syntactic measures such as Yngve score [6], Frazier score [7], and dependency distance [8], which were derived from syntactic parse trees and shown to be diagnostically informative.

Ablimit et al. [9] addressed the challenge of cross-corpus generalization in dementia detection. By analyzing acoustic and linguistic features across multiple datasets, they highlighted the variability introduced by different corpora and proposed the use of projection techniques to improve feature consistency. Their findings motivated our use of Linear Discriminant Analysis (LDA) for feature transformation and cross-dataset comparison.

III. Methodology

A. Datasets

1) The New Taipei Corpus: A total of 68 neurology patients for routine follow-up were recruited at New Taipei Municipal TuCheng Hospital, consisting of 35 healthy controls (HC), 16 individuals with MCI, and 17 individuals with Alzheimer's

979-8-3315-7206-8/25/$31.00

disease (AD). The AD data are not included since the present work focuses on early detection of MCI. Group labeling was based on Mini-Mental State Examination (MMSE) [10] scores, and Table I indicates a threshold adjustment according to years of education.

TABLE I: Criteria for HC vs. MCI group assignment

Education $\leq$ 6 years	**Education $>$ 6 years**
HC: MMSE $\geq$ 26	HC: MMSE $\geq$ 27
MCI: 23 $\leq$ MMSE $\leq$ 25	MCI: 24 $\leq$ MMSE $\leq$ 26

This dataset includes two types of tasks: self-introduction and picture description. In the self-introduction task, participants were asked to talk about their name, occupation, family background, hobbies, and other personal information. In the picture description task, participants were shown an image of Cookie Theft [11] and were asked to describe what they observed. In addition to the speech tasks, participants also completed the Montreal Cognitive Assessment (MoCA) [12] and MMSE cognitive assessments. This study focuses solely on the picture description task and uses MMSE scores as the basis for clinical group classification.

The exclusion criteria included (1) severe dementia, (2) inability to communicate (due to loss of consciousness, severe cognitive impairment, or mutism), and (3) lack of autonomy to provide informed consent. The demographic characteristics of the participants, including group, gender, age, education level, and MMSE score, are summarized in Table II. Data collection was approved by the Institution Review Board of Chang Gung Medical Foundation (No. 202301310B0C503).

The present data collection is part of an ongoing project called "Listen to Call for Help" (LCH). The purposes of LCH are to explore language and voice features, biomarkers, communication difficulties, caregiving burden, and health-related quality of life among older adults with cognitive impairment, and to develop tools to detect cognitive impairment and guidelines to alleviate communication difficulties. Questionnaires along with voice data are being collected from long-term care residents, community-based center attendees, and neurology patients.

2) The Delaware Corpus: We additionally retrieved the Delaware Corpus from the DementiaBank protocol [13], which includes audio recordings from participants with MCI and neurotypical controls, collected under a standardized protocol. To this date, the full dataset includes 95 individuals with MCI and 76 neurotypical participants, with data collection ongoing.

The dataset contains several speech elicitation tasks, including picture descriptions (e.g., Cookie Theft, Cat Rescue, Norman Rockwell's "Going and Coming"), narrative recall (e.g., Cinderella), procedural descriptions (e.g., making a peanut butter and jelly sandwich), and personal narratives (e.g., describing one's hometown). Cognitive assessments conducted as part of the protocol include the MoCA Blind/Telephone MoCA [14], the Boston Naming Test–Short Form (BNT-SF) [11], and the Multilingual Naming Test (MINT) [15], [16].

TABLE II: Demographic statistics (mean$\pm$ standard deviation)

Group (n)	Sex(M/F)	Age(y)	Education(y)	MMSE
HC (35)	18/17	70.80$\pm$4.73	9.14$\pm$3.03	28.20$\pm$1.28
MCI (16)	10/6	69.69$\pm$6.67	8.50$\pm$3.18	25.06$\pm$1.09
Total (51)	28/23	70.64$\pm$5.00	8.79$\pm$3.27	24.97$\pm$4.75

To enable analysis with a size comparable to the New Taipei dataset, we selected a subset of 70 participants from the Delaware dataset, consisting of 35 MCI and 35 neurotypical subjects. Only the Cookie Theft picture description task was used in subsequent analysis, as it aligns with the speech task used in our own recordings.

B. Task

We used the Cookie Theft picture description task, a standardized visual elicitation paradigm originally developed as part of the Boston Diagnostic Aphasia Examination [17]. This task has been widely adopted in dementia-related language research due to its ability to elicit spontaneous, syntactically rich narratives under minimal external constraints.

C. Annotation

From the interview recordings, we first partitioned each patient's speech into utterances, and the texts were manually transcribed in favor of accuracy. In our dataset, since elderly individuals in Taiwan often use both Mandarin and Southern Min (also called "Taiwanese") during interviews, all responses given in Taiwanese were manually translated into Mandarin. Then, the text transcripts were converted into TextGrid format using Praat software. Laughter and coughing were excluded from transcription. In addition to the speech content, we also annotated pauses between utterances, pauses within utterances, repeated segments, and filled pauses.

For the Delaware dataset, we used the official transcripts provided by DementiaBank and incorporated pause annotations to convert them into TextGrid format for consistency with our dataset.

D. Features Extraction

A total of 103 features were extracted, and they can be categorized into six feature types: Acoustic, Speaking Rate, Filled Pause, Part-of-Speech (POS), Syntactic Complexity, and Lexical Richness. A detailed list of features is given in Table III. Next, we describe the procedures for extracting each feature type.

Acoustic Features: We computed fundamental frequency (f0), the first three formants (f1–f3), harmonic-to-noise ratio (HNR), root mean square (RMS), zero-crossing rate (ZCR), and short-time magnitude. RMS and magnitude were also converted to the decibel (dB) scale. For each feature, five statistical descriptors were calculated at the utterance level [18], including maximum, minimum, mean, median, and standard deviation.

To mitigate gender-related pitch differences in the New Taipei dataset, f0 values were rescaled to semitones relative

TABLE III: Summary of extracted feature types

Feature Category	Subtypes (Feature Names)	Number of Features
Acoustic	F0 (max, min, maxdiff, mean, std), F1 (max, min, maxdiff, mean, median, std), F2 (max, min, maxdiff, mean, median, std), F3 (max, min, maxdiff, mean, median, std), HNR, RMS (max, min, maxdiff, mean, median, std), RMS_dB (max, min, maxdiff, mean, median), ZCR (max, min, maxdiff, mean, median, std), mag (max, mean, median, std), mag_dB (max, min, maxdiff, mean, median)	51
Speaking Rate	sen_num, sen_rate, word_num, word_rate, word_per_sen, pho_num, pho_rate, pho_per_sen	8
Filled Pause Features	total_dur, count_rate, sil_rate, sis_rate, fp_rate, re_rate, sil_dur_por, sis_dur_por, fp_dur_por, re_dur_por, sil_dur_mean, sis_dur_mean, fp_dur_mean, re_dur_mean, sil_dur_std, sis_dur_std, fp_dur_std, re_dur_std	18
Part-of-Speech (POS)	n_rate, pron_rate, v_rate, adj_rate, adv_rate, word_num, sen_num, n_num, pron_num, v_num, adj_num, adv_num, n_per, pron_per, v_per, adj_per, adv_per	17
Syntactic Complexity	w_avg_Frazier, w_avg_Yngve, w_avg_depend_dis	3
Lexical Richness	uni_richness, bi_richness, tri_richness, uni_richness_rate, bi_richness_rate, tri_richness_rate	6
Total		**103**

to average f0 of Taiwanese elderly men (137 Hz) and women (169 Hz) [19]. The conversion formulas are shown below:

$$f_{0\_\text{new}} = 12 \cdot \log_2 \left(\frac{f_0}{f_{0\_\text{gender}}} \right) \tag{1}$$

$$f_{0\_\text{gender}} = \begin{cases} 137, & \text{if gender = male} \\ 169, & \text{if gender = female.} \end{cases} \tag{2}$$

Speaking Rate Features: To extract speaking rate features, we used the pretrained "wav2vec2-xlsr-53-espeak-cv-ft" model [20] to automatically transcribe phoneme sequences from segmented audio at the sentence level. Based on the transcriptions and phoneme alignments, we computed word- and phoneme-related features from the picture description task.

Filled Pause Features: Based on the manual annotations from the picture description task, we extracted filled pause features using the Praat toolkit. Labeled events include inter-sentence silences ("sil"), intra-sentence silences ("sis"), filled pauses ("fp"), and repeated segments ("re"), all annotated with their corresponding durations ("dur"). For each participant, we calculated the mean and standard deviation of the durations, along with the event counts. To control for variability in total speaking time, we normalized these features by computing their rate per second and their duration proportion ("por") relative to the entire task. Thus, this feature set contains 18 dimensions.

Part-of-Speech (POS): Part-of-speech features were extracted from the transcripts to capture lexical usage patterns. For our dataset, which consists of Traditional Chinese transcripts, we used the CKIP-Transformers toolkit [21] for word segmentation and POS tagging. For the Delaware dataset, which contains English transcripts, we used the NLTK toolkit [22] to perform tokenization and POS tagging. In both datasets, 62 or more fine-grained POS tags were mapped to five major categories: nouns, pronouns, verbs, adjectives, and adverbs. To account for differences in task duration, we computed both the raw counts and normalized rates (per second and relative to total word count). This feature set contains 17 dimensions.

Syntactic Complexity Features: We incorporated two syntactic analysis techniques: constituency parsing and dependency parsing. For constituency parsing, we utilized NLTK CoreNLP parser to construct parse trees. For the New Taipei dataset, we first performed word segmentation using CKIP-Transformers; for the Delaware dataset, we tokenized the English transcripts using NLTK. Regardless of language, constituency parsing was conducted using the CoreNLP parser.

Fig. 1 shows an example of a constituency-based parse tree constructed from the sentence "she was a cook in a school cafeteria". The tree decomposes the sentence into hierarchical structures based on part-of-speech tags and grammar rules. From the nodes of the tree, we computed the Yngve score and Frazier score for each token. For sentence-level features, we summed the token-wise scores within each sentence.

For dependency parsing, we utilized Stanford NLP Group's Stanza toolkit [23]. The Chinese model was applied to the New Taipei dataset and the English model to the Delaware dataset. Fig. 2 shows a dependency parse tree of the same sentence "she was a cook in a school cafeteria". We computed dependency distance by measuring the linear distance between each word and its head in the sentence. The average dependency distance was used as a feature to quantify syntactic complexity.

Lexical Richness: For measuring lexical richness, it was necessary to conduct word segmentation first. For the New Taipei dataset, we used CKIP-Transformers for Traditional Chinese word segmentation; for the Delaware dataset, we used NLTK for English tokenization. After segmentation, n-grams ($n = 1, 2, 3$) were generated for each transcript.

The lexical richness is defined as the proportion of unique n-grams to the total number of n-grams within each n-gram level. Additionally, we computed the richness rate, defined as the proportion of unique n-grams divided by the task duration in seconds. This reflects the diversity of lexical use

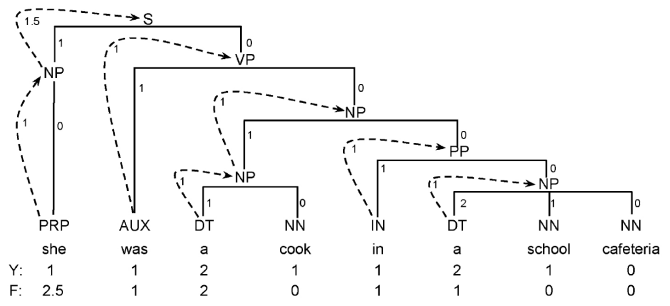


Fig. 1: A constituency-based parse tree example with Yngve scores and Frazier scores. The Yngve scores (Y) are summed up by the branch scores, and the Frazier scores (F) are summed up by the scores on upward dotted line. S=Sentence, NP=Noun Phrase, VP=Verb Phrase, PP= Prepositional Phrase, PRP=Personal Pronoun, AUX=Auxiliary Verb, DT=Determiner, NN=Noun(singular or mass), IN=Preposition or Subordinator.

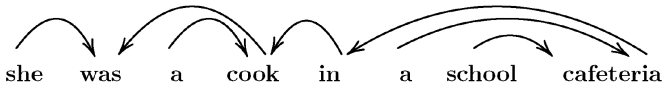


Fig. 2: A dependency-based parsing example. In this example, the 7 arrows represent the dependency relationship of each word, and the sum of the distances for each link is 11, thus, the average dependency distance can be computed as $\frac{11}{7}$.

per unit time, and serves as an indicator of expressive language capacity.

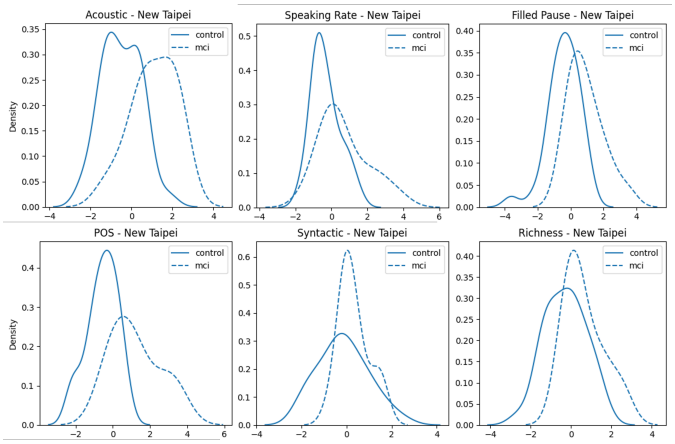
IV. Experiments and Results

A. Discriminative Power Evaluation

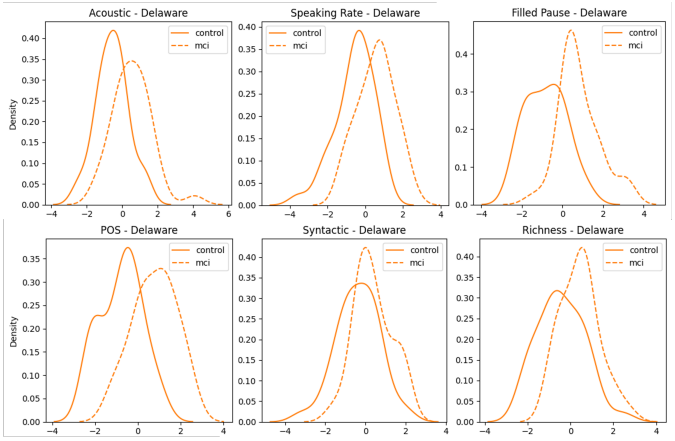
To evaluate the discriminative ability of the extracted features for HC/MCI classification, we applied LDA to each of the six feature sets. The LDA projections of the six feature sets for both datasets are visualized in Fig. 3, where the blue lines represent the New Taipei dataset and the orange lines represent the Delaware dataset. The corresponding Receiver Operating Characteristic (ROC) Area Under Curve (AUC) summarized in Table IV.

TABLE IV: ROC AUC of each LDA projection, with values greater than 0.6 highlighted in bold.

Feature	AUC	
	New Taipei	Delaware
Acoustic	**0.654**	0.598
Speaking Rate	0.488	**0.683**
Filled Pause	0.461	**0.620**
POS	0.538	**0.638**
Syntactic	0.511	**0.611**
Lexical Richness	**0.604**	**0.620**



(a) LDA projections of the New Taipei dataset.



(b) LDA projections of the Delaware dataset.

Fig. 3: LDA projections of each feature set in both datasets.

B. SVM Classification Using Fused Features

We used an AUC threshold of 0.6 as the selection criterion to identify informative feature sets. In the New Taipei dataset, the selected features included Acoustic and Lexical Richness. In the Delaware dataset, the selected features were Speaking Rate, Filled Pause, POS, Syntactic Complexity, and Lexical Richness. Each selected feature set was first individually projected onto a one-dimensional LDA axis. These LDA projections were concatenated to form the feature vectors for classification, then SVM classifiers were trained on the New Taipei and Delaware datasets, respectively. The SVM hyperparameters were optimized using GridSearchCV to automatically search for the parameter settings that yield the highest F1 score. The grid search explored values of penalty weight $C \in \{0.1, 1, 10, 20, 50\}$, kernel $\in$ {linear, polynomial, radial basis function, sigmoid}, and $\gamma \in$ {auto, scale}.

Note that the Acoustic feature set contains 51 dimensions, which happens to equal the number of speakers in the New Taipei dataset. To avoid overfitting, we applied Principal Component Analysis (PCA) [24] to reduce the dimensionality to 10 before performing LDA projection and classifier training. The classification performance of the two hybrid classifiers,

evaluated using three-fold cross-validation, is summarized in Table V.

TABLE V: Fusion SVM model performance, presented in mean±standard deviation based on 50 random rounds of 3-fold validation.

Dataset	F1-score (%)	Accuracy (%)
New Taipei	48.17±17.22	69.49±9.85
Delaware	56.91±11.41	57.94±9.25

V. Discussion

Comparing the two datasets, the discriminative power of the acoustic features notably differs. A contributing factor might be the sound level of speech. When calculated by utterances, the mean value of sound magnitude for MCI in the New Taipei dataset is –21.68±4.50 dB relative to the full-scale level; in contrast, the mean magnitude is –24.98±4.63 dB for HC, indicating a noticeable difference in vocal intensity between the two groups. In the Delaware dataset, the difference is not significant, with HC and MCI mean magnitudes of –18.64±3.43 dB and –17.88±3.43 dB, respectively.

Regarding the speaking rate, we found no clear difference in phoneme per sentence between HC and MCI participants in the New Taipei dataset, with mean values of 22.66±14.43 for HC and 22.15±6.51 for MCI. In contrast, participants in the Delaware dataset tend to speak faster, with mean values of 31.69±13.82 for HC and 27.32±10.93 for MCI. Notably, slower speakers in the Delaware dataset are more likely MCI patients. During the annotation process, we observed that slower speech does not necessarily indicate cognitive decline, and the statistics reflect the naturally slower speaking style of elderly individuals in New Taipei region.

For the Filled Pause, POS, and Syntactic Complexity feature sets, the New Taipei dataset shows lower discriminative power compared to the Delaware dataset. One possible explanation is that many elderly speakers in Taiwan tend to skip words for the sake of colloquial convenience. While these utterances may remain understandable to listeners, they often result in fragmented or incomplete sentences when transcribed, thereby reducing the effectiveness of part-of-speech and syntactic complexity features in distinguishing between HC and MCI groups.

In addition, filled pauses are frequently used as discourse markers in local speech, often appearing at the beginning of utterances either as a habitual speaking style or during moments of reflection. As a result, these features provide limited discriminative ability between HC and MCI participants in the New Taipei dataset.

We also conducted an experiment using a balanced subset of the New Taipei dataset by randomly selecting 16 HC participants to match the number of MCI participants. Using the same classification method, the F1-score increased to 60.80±14.66%, suggesting that class imbalance negatively impacts classification performance.

VI. Conclusion and Future Work

This study investigated the discriminative potential of six feature sets—Acoustic, Speaking Rate, Filled Pause, POS, Syntactic, and Lexical Richness—for distinguishing between HC and individuals with MCI. Through LDA-based projection and ROC AUC analysis, we identified different sets of informative features across the New Taipei and Delaware datasets, highlighting the influence of population-specific speech characteristics. In the New Taipei dataset, Acoustic and Lexical Richness features demonstrated stronger discriminative power (AUC > 0.6), while in the Delaware dataset, non-acoustic features such as Speaking Rate, Filled Pause, POS, and Syntactic measures showed higher utility.

SVM classifiers trained on concatenated LDA projections of the selected features achieved an F1-score of 48.17% and an accuracy of 69.49% for the New Taipei dataset, and an F1-score of 56.91% and an accuracy of 57.94% for the Delaware dataset. These results demonstrate that feature selection based on dataset-specific statistics can help to raise the performance of MCI detection mildly but nonetheless in a robust manner. The present results are thus encouraging for future deployment of the models to process previously unseen data.

For future work, in addition to speech, pause and duration, and linguistic features, we plan to incorporate semantic features by computing textual semantic similarity. This will allow for a more comprehensive representation of linguistic differences and enhance the detection of MCI. In the upcoming year, we also intend to adopt the "Cat Rescue" picture for a picture description task. Since this picture is also included in the Delaware corpus, it will enable us to analyze how different visual stimuli affect the performance of MCI detection.

Acknowledgment

This research was supported by the National Science Council of Taiwan under grant No. 112-2410-H-007-048 and 113-2410-H-038-007. We acknowledge the brainstorming that we received from Professor Ching-Ching Lu of the Institute of Taiwan Languages and Language Teaching at National Tsing Hua University. We also thank Ms. Yi-Shan Tsai for her significant contribution to data transcription and brainstorming.

This work utilized data from the DementiaBank protocol. Data collection was supported by the National Institute on Deafness and Other Communication Disorders under grant R01DC008524-13S1 awarded to Brian MacWhinney, and by the National Institute on Aging of the National Institutes of Health under award number RF1AG083823 (PIs: MacWhinney and Lanzi). Use of the dataset complies with TalkBank regulations. We gratefully acknowledge the authors of DementiaBank for making this resource available to the research community.

References

[1] J. Liss, S. Seleri Assunção, J. Cummings, *et al.*, "Practical recommendations for timely, accurate diagnosis of symptomatic alzheimer's disease (mci and dementia) in primary care: A review and synthesis," *Journal of Internal Medicine*, vol. 290, no. 2, pp. 310–334, 2021.

[2] T. Koch, S. Iliffe, and E.-E. project, "Rapid appraisal of barriers to the diagnosis and management of patients with dementia in primary care: A systematic review," *BMC Family Practice*, vol. 11, pp. 1–8, 2010.

[3] I. Martínez-Nicolás, T. E. Llorente, F. Martínez-Sánchez, and J. J. G. Meilán, "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*, vol. 12, p. 620 251, 2021.

[4] R. Voleti, J. M. Liss, and V. Berisha, "A review of language and speech features for cognitive-linguistic assessment," *arXiv preprint arXiv:1906.01157*, 2019.

[5] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2081–2090, Sep. 2011. DOI: 10.1109/TASL.2011.2112351.

[6] V. H. Yngve, "A model and an hypothesis for language structure," *Proceedings of the American philosophical society*, vol. 104, no. 5, pp. 444–466, 1960.

[7] L. Frazier, "Syntactic complexity," in *Natural Language Parsing*, D. Dowty, L. Karttunen, and A. Zwicky, Eds., Cambridge, U.K.: Cambridge University Press, 1985, pp. 129–189.

[8] D. M. Magerman, "Statistical decision-tree models for parsing," in *33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, Massachusetts, USA: Association for Computational Linguistics, Jun. 1995, pp. 276–283. DOI: 10.3115/981658.981695. [Online]. Available: https://aclanthology.org/P95-1037/.

[9] A. Ablimit, C. Botelho, A. Abad, T. Schultz, and I. Trancoso, "Exploring dementia detection from speech: Cross corpus analysis," in *Proc. IEEE Int.Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6472–6476. DOI: 10.1109/ICASSP43922.2022.9747167.

[10] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state": A practical method for grading the cognitive state of patients for the clinician," *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–198, 1975.

[11] E. Kaplan, H. Goodglass, and S. Weintraub, *Boston Naming Test*. Lippincott Williams & Wilkins, 2001.

[12] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, *et al.*, "The montreal cognitive assessment, moca: A brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[13] A. M. Lanzi, A. K. Saylor, D. Fromm, H. Liu, B. MacWhinney, and M. Cohen, "Dementiabank: Theoretical rationale, protocol, and illustrative analyses," *American Journal of Speech-Language Pathology*, vol. 32, no. 2, pp. 426–438, 2023.

[14] P. Dawes, A. Pye, D. Reeves, *et al.*, "Protocol for the development of versions of the montreal cognitive assessment (MoCA) for people with hearing or vision impairment," *BMJ Open*, vol. 9, no. 3, e026246, 2019.

[15] A. Stasenko, D. M. Jacobs, D. P. Salmon, and T. H. Gollan, "The multilingual naming test (MINT) as a measure of picture naming ability in alzheimer's disease," *Journal of the International Neuropsychological Society*, vol. 25, no. 8, pp. 821–833, 2019.

[16] T. H. Gollan, G. H. Weissberger, *et al.*, "Self-ratings of spoken language dominance: A multi-lingual naming test (mint) and preliminary norms for young and aging spanish-english bilinguals," *Bilingualism: Language and Cognition*, vol. 15, no. 3, pp. 594–615, 2012.

[17] H. Goodglass and E. Kaplan, *Boston Diagnostic Aphasia Examination Booklet*. Philadelphia: Lea & Febiger, 1983.

[18] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Communication*, vol. 52, no. 7-8, pp. 613–625, 2010.

[19] C.-Y. Cheng, "Acoustic characteristics of elderly voices (in Chinese)," in *Proceedings of the 31st Annual Conference of the Acoustical Society of Taiwan*, National Kaohsiung Normal University, Taiwan, 2018.

[20] Q. Xu, A. Baevski, and M. Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," *arXiv preprint arXiv:2109.11680*, 2021.

[21] CKIP Lab, *CKIP-transformers: Traditional Chinese transformers models and NLP tools*, version v0.3.4, Accessed: 2025-06-17, 2023. [Online]. Available: https://github.com/ckiplab/ckip-transformers.

[22] E. Loper and S. Bird, "NLTK: The natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[23] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, A. Celikyilmaz and T.-H. Wen, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 101–108. DOI: 10.18653/v1/2020.acl-demos.14. [Online]. Available: https://aclanthology.org/2020.acl-demos.14/.

[24] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (PCA)," *Computers Geosciences*, vol. 19, no. 3, pp. 303–342, 1993, ISSN: 0098-3004. DOI: https://doi.org/10.1016/0098-3004(93)90090-R. [Online]. Available: https://www.sciencedirect.com/science/article/pii/009830049390090R.