Detecting Mild Cognitive Impairment Across Languages: An Analysis of Speech Features in Chinese and English

Hung-Wei Lee and Ya-Ning Chang[1]
Miin Wu School of Computing, National Cheng Kung University

Abstract

Speech analysis offers significant potential for the early, cross-linguistic detection of Mild Cognitive Impairment (MCI), but the crucial features for this remain unclear. Our study investigated a classification model for MCI detection in both English and Chinese, using three interpretable acoustic feature sets: time-domain (TD), eGeMAPS (EGE), and short-time Fourier transform (STFT). We found that integrating multi-domain features yielded the best performance in combined language conditions. Specifically, robust cross-linguistic acoustic markers were linked to energy variation, voicing regularity, fine-grained temporal and spectral dynamics, and amplitude envelope features, as identified by group-based SHAP analysis.

Introduction

Automated methods using machine learning and deep learning show growing promise for analyzing speech biomarkers for Mild Cognitive Impairment (MCI), a crucial stage for dementia prevention. However, most research has focused on single languages, limiting the broad applicability of these findings. Recent studies, like those from the TAUKADIAL Challenge (Luz et al., 2024; Perez-Toro et al., 2024), have begun exploring cross-linguistic MCI detection using Mandarin Chinese and English speech. While these highlight the potential of multilingual approaches, it remains unclear which specific features are critical for cross-linguistic MCI detection.

Our study systematically investigated the efficacy of a classification model for early MCI detection in both English and Chinese. We used three key acoustic feature sets: statistical time-domain (TD), the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS, EGE) (Eyben et al., 2015), and short-time Fourier transform (STFT). These were chosen for their broad acoustic coverage, interpretability for clinicians and linguists, and consistent extractability across languages. We also conducted model interpretation analyses to identify the crucial features for cross-linguistic MCI detection.

Methods

Data and Participants

The study utilised the TAUKADIAL Challenge corpus (Luz et al., 2024), consisting of 507 picture description recordings: 260 in Mandarin (ZH) and 247 in English (EN). All recordings were labeled as either MCI or cognitively normal control (NC). Speakers of both languages were between 61 and 87 years of age (mean = 73.4, SD = 6.4). The gender distribution was 39% male and 61% female, with comparable proportions in both the MCI and NC cohorts. Participants performed a picture description task in their native language, with each recording averaging approximately one minute in duration (total speech time $<$ 528 minutes). To ensure speaker independence, the corpus was divided into 80% training and 20% test partitions using StratifiedGroupKFold (Pedregosa et al., 2011). The stratification key was the diagnostic label, while grouping by a unique speaker identifier prevented recordings from the same individual from appearing in both partitions.

Feature Extraction

All recordings were denoised, resampled to 16 kHz, and silence-trimmed. Three sets of acoustic features were extracted and grouped into clinically meaningful categories for subsequent explainability analysis. Specifically, we computed 30 TD features related to voicing, macro-prosody, attack–decay–sustain–release (ADSR) envelope, autocorrelation harmonics, teager energy, and fine statistical moments; 88 EGE features extracted using openSMILE (Eyben et al., 2015), including pitch, energy, spectral shape, mel-frequency cepstral coefficients (MFCCs), formants, and voice quality; and 201 STFT features, calculated as time-averaged spectral energies from a 400-point FFT (25 ms window, 10 ms hop). These STFT features were further grouped into five frequency bands: Band1 (0-1000 Hz), Band2 (1000-2000 Hz), Band3 (2000-4000 Hz), Band4 (4000-6000 Hz), and Band5 (6000-8000 Hz). All features were z-score normalized.

Classifier and Training

A two-layer fully connected neural network (512 units, ReLU, batch normalization, dropout=0.5) was trained with Adam (learning rate=1e-5, weight decay=1e-4). We used five-fold speaker-independent cross-validation and early stopping (validation F1 score) to evaluate seven feature configurations (single, pairwise, and all combinations). Final models were then retrained on the full training data and tested on the held-out set.

Model Explainability

We used Shapley Additive Explanations (SHAP) with the KernelExplainer (Lundberg & Lee, 2017) for model

[1]Corresponding author: Ya-Ning Chang, Email: yaningchang@gs.ncku.edu.tw

Table 1: Test-set F1-score for each feature set and language.

Feature	Language	F1
All	ZH	0.67
	EN	0.51
	ZH+EN	**0.81**
EGE	ZH	0.68
	EN	0.74
	ZH+EN	0.54
STFT	ZH	0.61
	EN	0.43
	ZH+EN	0.66
TD	ZH	0.58
	EN	0.58
	ZH+EN	0.61
EGE+STFT	ZH	0.62
	EN	0.80
	ZH+EN	0.59
TD+EGE	ZH	**0.70**
	EN	**0.88**
	ZH+EN	0.65
TD+STFT	ZH	0.65
	EN	0.32
	ZH+EN	0.80

interpretation. For each feature configuration, we randomly selected 100 training samples as a background set. These were aligned to the shortest segment length and then mean-pooled along the time axis to create fixed-size representations. KernelExplainer then calculated SHAP values for these pooled vectors using its default sampling strategy. To improve readability, we summed absolute SHAP values within predefined acoustic groups (TD, EGE, and STFT) to generate a group-level importance score. The ten most influential groups per experiment were then ranked and visualized.

Results

Table 1 reports F1-scores for all feature sets. TD+STFT+EGE fusion performed best on the combined dataset (F1 = 0.81); TD+EGE was optimal for English (0.88) and Mandarin (0.70), highlighting the benefit of combining time-domain and prosodic features. Table 2 shows two groups consistently ranked highest across languages—TD_Voicing and EGE_Energy—while TD_ADSR, TD_AutocorrHarmonics, and EGE_MFCC were additional key groups in combined analysis. Language-specific patterns emerged clearly. Mandarin emphasized TD_FineStats, TD_AutocorrHarmonics and TD_TeagerAmpDynamics, reflecting sensitivity to syllabic amplitude variations. English prioritized high-frequency spectral energy (STFT_Band5), EGE_Pitch, and TD_FineStats, indicating spectral tilt and fine temporal cues as critical in stress-timed speech.

Table 2: Top-5 SHAP Feature Groups for each language grouping. Mean SHAP values are the average absolute SHAP importances from KernelSHAP analysis.

Language	**Feature Group**	**Mean SHAP**
Mandarin (ZH)	TD_FineStats	0.00720
	EGE_Energy	0.00506
	TD_AutocorrHarmonics	0.00441
	TD_Voicing	0.00425
	TD_TeagerAmpDynamics	0.00361
English (EN)	TD_FineStats	0.00429
	TD_Voicing	0.00260
	STFT_Band5	0.00164
	EGE_Pitch	0.00116
	EGE_Energy	0.00110
Combined (ZH+EN)	TD_Voicing	0.00271
	TD_ADSR	0.00191
	TD_AutocorrHarmonics	0.00136
	EGE_MFCC	0.00131
	EGE_Energy	0.00121

Discussion

This study examined a classification model for early MCI detection in English and Chinese, identifying shared and unique acoustic features. Results showed feature fusion consistently enhanced model generalizability. SHAP analyses revealed robust common features—primarily TD_Voicing and EGE_Energy—and clear language-specific markers. Notably, TD_FineStats ranked highest in both single-language models but dropped outside the top-5 in the combined analysis, suggesting that micro-level temporal statistics (e.g., short-window variance, skewness) strongly differentiate MCI within each language yet manifest differently across languages, reducing their discriminative power when pooled. Specifically, energy variation (EGE_Energy), voicing regularity (TD_Voicing), temporal envelope (TD_ADSR & TD_AutocorrHarmonics), and fine-band spectral features (EGE_MFCC) were reliable cross-lingual MCI indicators. These align with clinical findings (Agbavor & Liang, 2024; Fraser, Meltzer, & Rudzicz, 2016) on reduced prosodic control and laryngeal stability in cognitive decline. Language-specific markers were also found: Mandarin relied more on formant dynamics and macro-prosodic pitch modulation, while English depended on spectral tilt and fine statistical fluctuations. This highlights the importance of language-tailored models.

Conclusion

In summary, our study demonstrated that while key acoustic features for MCI vary with linguistic structures, common characteristics could also be identified and proved useful for MCI detection. Future work can consider automatic group discovery, integration of linguistic features, and broader multilingual datasets to further advance fair, explainable clinical speech screening.

Acknowledgement

This research was supported by grants from the National Science Technology Council (NSTC 113-2224-E-006-003 and NSTC 112-2927-I-006-507 to YNC). The authors have no conflicts of interest to declare.

References

Agbavor, F., & Liang, H. (2024). Multilingual prediction of cognitive impairment with large language models and speech analysis. *Brain Sciences*, *14*(12), 1292. doi: 10.3390/brainsci14121292

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... Truong, K. P. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202.

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, *49*(2), 407–422. doi: 10.3233/JAD-150520

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774.

Luz, S., Garcia, S. D. L. F., Haider, F., Fromm, D., MacWhinney, B., Lanzi, A., ... Liu, Y. (2024). *Connected speech-based cognitive assessment in Chinese and English.* arXiv preprint arXiv:2406.10272. doi: 10.48550/arXiv.2406.10272

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perez-Toro, P. A., Arias-Vergara, T., Klumpp, P., Weise, T., Schuster, M., Nöth, E., ... Maier, A. (2024). Multilingual speech and language analysis for the assessment of mild cognitive impairment: Outcomes from the TAUKADIAL challenge. In *Proceedings of INTERSPEECH 2024* (pp. 982–986). Kos, Greece. doi: 10.21437/Interspeech.2024-2115