Language-agnostic Speech Biomarker Exploration for Early Dementia Screening

Josh Ashik*, Zongxing Xie†, Ying Chen‡, Chenhan Xu*, Huining Li*
**Department of Computer Science, North Carolina State University, USA*
Email: {juashik, cxu34, hli83}@ncsu.edu
†*Department of Computer Science, Kennesaw State University, Marietta, USA*
Email: zxie1@kennesaw.edu
‡*College of Information Science and Technology, Pennsylvania State University, USA*
Email: yfc5578@psu.edu

Abstract—**Early dementia detection is a global healthcare priority in diverse populations. In this study, we propose a language-agnostic screening pipeline for dementia detection in the early stage. First, we use speaker diarization to isolate the speech of the target subject from a conversational recording. From the extracted speech segments, we derive a set of acoustic features (e.g., spectral centroid, pitch mean, mel-frequency cepstral coefficients) and linguistic features (e.g., normalized tone contrast, articulation clarity coefficient, articulatory effort coefficient). These features are used to train a ResNet-based binary classifier to distinguish between Healthy Controls (HC) and individuals with Mild Cognitive Impairment (MCI). We evaluated the trained model on a held-out test set comprising speakers of previously unseen languages, achieving an accuracy of 70%. This cross-lingual transfer performance highlights the potential of our approach for scalable, language-independent dementia screening.**

Index Terms—**Speech Biomarker, Language-agnostic, Artificial Intelligence, Dementia Screening**

I. Introduction

Dementia affects over 55 million people globally, with nearly 10 million new cases each year, posing a growing challenge to public health systems worldwide [1]. Early detection is critical: interventions at the stage of Mild Cognitive Impairment (MCI) can delay progression and improve quality of life. However, traditional cognitive assessments remain resource-intensive, requiring in-person clinical visits, trained professionals, and subjective interpretation. These constraints limit the scalability of early dementia detection, especially in diverse or low-resource settings. To address this, researchers have increasingly turned to machine learning-based tools to monitor early cognitive decline from daily life activities (e.g., conversation), with the goal of creating accessible, software-based systems capable of notifying clinicians or caregivers when signs of decline emerge.

Recent research has increasingly recognized language impairments as early markers of cognitive decline, which positions speech and language characteristics as important biomarkers for the early diagnosis of dementia [2]. Fraser et al. [3] highlighted lexical and syntactic impairments as markers of dementia, while Luz et al. [4] investigated conversational speech, reporting promising but language-specific performance. However, these studies often rely on single-language datasets and lack evaluation on unseen languages. Recently, Bertini et al. (2022) [5] proposed a cross-language dementia classifier, which trained an end-to-end deep model on multilingual datasets. Their approach treated the model as a black box and offers limited interpretability, which makes it challenging to extend the framework to more diverse languages.

To address these limitations, we propose a language-agnostic system for early dementia detection using machine learning models trained on acoustic and linguistic features extracted from spontaneous speech. Our goal is to create a robust system that can generalize across languages, enabling dementia detection in previously unseen linguistic groups.

Several technical challenges arise in building such a system. First, there is model uncertainty due to language mismatch between training and test data, especially when transferring to previously unseen languages such as Mandarin [6]. To mitigate this, we incorporate statistical interpretation of classification probabilities to better infer the likelihood of MCI. Second, there is the challenge of capturing consistent features across different languages. We address this by extracting cross-linguistically relevant linguistic features, such as lexical richness, syntactic complexity, and speech timing, alongside acoustic features like pitch, jitter, and shimmer. Our pipeline included data preprocessing, speaker diarization, feature extraction, ResNet-based classification, and evaluation on unseen languages.

We selected Mandarin for zero-shot evaluation because it represents one of the most widely spoken languages globally, with over a billion native speakers. Yet, it remains underrepresented in publicly available clinical speech datasets, especially those with validated dementia annotations. By demonstrating cross-lingual performance on Mandarin without retraining, we highlight the model's potential for deployment in large, underserved populations where labeled clinical data is scarce.

Our contributions include: (1) a full audio preprocessing pipeline including robust speaker diarization for conversations, (2) extraction of cross-linguistically valid acoustic and linguistic features, (3) development and training of a deep neural model for dementia classification, and (4) cross-lingual eval-

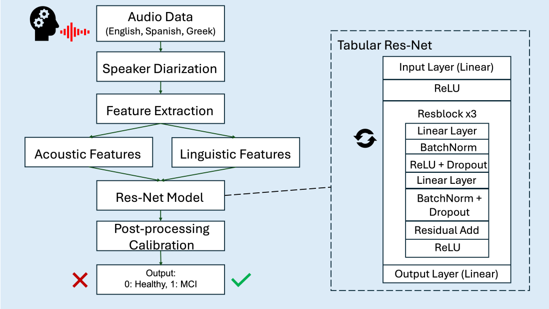

Fig. 1. System pipeline for language-agnostic dementia detection.

uation of the model's generalization capability on an unseen language. By enabling dementia screening across languages without retraining for each population, this work represents a critical step toward accessible, scalable, and language-independent dementia detection.

II. Method

This section describes our dementia screening pipeline. The full system workflow is depicted in Fig. 1.

A. Audio Preprocessing and Quality-Aware Data Curation

The audio recordings are preprocessed using noise reduction, speaker diarization, and voice activity detection. Only continuous speech segments produced solely by the target subject are retained. To maintain high audio quality, we apply signal-to-noise ratio (SNR) criteria and exclude segments containing overlapping speech, excessive background noise, or non-speech vocalizations. This rigorous filtering process ensures the reliability and accuracy of subsequent feature extraction.

B. Acoustic Feature Extraction

To characterize the prosodic and spectral structure of each subject's speech, we extracted a set of frame-level acoustic features that reflect fundamental cognitive and motor aspects of vocal production. These features were computed over 25 ms windows with a 10 ms hop size and aggregated using statistical pooling to produce a fixed-length feature vector for each recording.

- **PitchMean and PitchStd**: These features represent the first-order statistical descriptors (mean and standard deviation) of the fundamental frequency (F0) trajectory over voiced regions. Pitch is extracted from the autocorrelation of short-time frames in the time domain, where the lag of maximum correlation approximates the glottal cycle length. PitchMean reflects vocal register; PitchStd captures prosodic variability, often reduced in cognitive decline.
- **Formant Frequencies (F1, F2, F3)**: Formants are resonant frequencies of the vocal tract, and their values correspond to articulatory configurations. Formants are estimated via LPC-based all-pole vocal tract modeling. The roots of the LPC polynomial yield formant candidates via their angular frequencies. F1 to F3 represent vowel height, backness, and lip rounding, and their shifts may indicate articulatory deficits.
- **Spectral Centroid**: This feature is computed as the weighted mean frequency of the power spectrum and corresponds perceptually to the brightness of the speech signal. Mathematically, it is given by $\mu = \frac{\sum_k f_k S(f_k)}{\sum_k S(f_k)}$, where $S(f_k)$ is the spectral magnitude at frequency bin f_k.
- **Spectral Spread**: This represents the second central moment (variance) of the spectrum around the spectral centroid, quantifying spectral dispersion. It is calculated as $\sigma^2 = \frac{\sum_k (f_k - \mu)^2 S(f_k)}{\sum_k S(f_k)}$.
- **Spectral Flux**: This measures the frame-to-frame Euclidean distance between normalized spectra, reflecting dynamic spectral change. It is sensitive to abrupt transitions and is linked to articulatory agility.
- **MFCC1–MFCC3**: The Mel-Frequency Cepstral Coefficients (MFCCs) capture the spectral envelope of speech on a perceptually motivated frequency scale. These are obtained by applying the Discrete Cosine Transform (DCT) to the log-filterbank energies. The first three coefficients reflect low-order spectral shape variations linked to vocal effort and clarity.

All features are statistically pooled using mean and variance operators to summarize temporal dynamics over the entire recording. This produces a compact representation of global and local prosodic features.

C. Linguistic and Prosodic Feature Extraction

To model language-agnostic articulatory and phonological patterns, we extract features from phoneme sequences derived through forced alignment and phonemization [7]. The underlying goal is to quantify timing, clarity, and effort in speech production, independent of language-specific lexical content.

In general, phoneme sequences are obtained by aligning audio with transcripts using Whisper (large-v2) [8] and then converting aligned tokens into International Phonetic Alphabet (IPA) symbols using espeak-ng [9]. For Mandarin, phonotactic rules are applied to classify phonemes into "initials" (typically consonants) and "finals" (typically vowels or syllabic nuclei), enabling structured analysis of phonological transitions.

The following features are computed:

- **PhonemeTranscription**: we calculate a sequence of IPA-based phonemes per utterance, which can be used as the foundation for higher-order linguistic features.
- **Mean Phonation Duration (MDP)**: It is calculated by averaging the lengths of continuous voiced segments (i.e., periods of vocal fold vibration) within a speech sample. It captures temporal control in vowel production, which can indicate the decline with motor impairment.
- **Normalized Tonal Contrast (NTC)**: It quantifies the frequency of transitions between initials and finals, normalized by sequence length. This reflects the rhythm and tonal variation of speech, which are often reduced in MCI.
- **Articulation Clarity Coefficient (ACC)**: It is defined as the average run-length of initial segments, indicating consistency in consonant articulation. A higher ACC reflects stable motor planning in speech onset.

- **Articulatory Effort Coefficient (ACE)**: It refers to the average run-length of final segments that captures the subject's sustained vowel effort. It indirectly reflects vocal stamina and breath support.
- **Log-Cepstral Coefficient (LCC)**: It denotes the maximum run-length of consecutive initial phonemes, reflecting burst-like consonant articulation that may become fragmented as cognitive decline progresses.
- **Log-Cepstral Energy (LCE)**: It represents the maximum run-length of final segments, approximating vocal energy over extended vowel sounds. It captures speech smoothness and fluency.

These features are designed to generalize across languages by analyzing speech production at the phonetic level, avoiding dependency on lexicon or grammar. They capture articulatory timing, segmental clarity, and rhythmic stability—properties shown to degrade with early cognitive [3].

D. Tabular ResNet and Optimization

While simpler models like SVMs offer lower computational cost, they underperformed in cross-lingual settings. To effectively work with structured feature matrix (i.e., tabular data), we use a tabular ResNet model to better capture non-linear feature interactions, offering improved generalization with minimal added complexity. It contained approximately 113,665 trainable parameters (for 100 input features), offering a balance between expressivity and computational efficiency. The model is optimized using binary cross-entropy loss with a class-weighted formulation and Adam optimization.

E. Post-Processing Calibration

Temperature scaling is a post-processing method used to improve probability calibration by adjusting the confidence of the softmax outputs. Specifically, it introduces a scalar temperature parameter $T > 0$ that rescales the logits before applying the softmax function. Higher temperatures flatten the output distribution, reducing overconfidence, while lower temperatures sharpen it. Temperature scaling helps ensure that predicted probabilities are better aligned with actual outcome frequencies without affecting classification accuracy [10].

III. Experiments & Results

A. Dataset and Implementation

Data Preparation. We use the DementiaBank dataset [11]. The dataset consists of over 200 participants, offering a rich source of conversational speech. Participants include individuals diagnosed with Mild Cognitive Impairment (MCI) as well as Healthy Controls (HC), with detailed metadata including age (typically 45–90 years), gender, and education level. This corpus contains audio recordings and transcripts of clinical interviews, primarily involving the Cookie Theft picture description task from the Boston Diagnostic Aphasia Examination. The ResNet classifier is trained using a combined dataset of English, Spanish, and Greek speech recordings. Speech data is divided into overlapping clips of 15–30 seconds, with each segment retaining full utterance boundaries. In total, we get around 900 samples. We perform five-fold cross-validation to evaluation the performance.

Software Implementation. Experiments are seeded for reproducibility. Feature scaling is applied using a standard scaler, and training is performed using PyTorch with the Adam optimizer. A batch size of 64 and an initial learning rate of 0.001 are used across all training epochs.

B. Evaluation Metrics

We evaluate model performance using the following standard classification metrics: accuracy, F1 score, receiver operating characteristic (ROC) and precision-recall (PR) curves.

C. Results

1) Overall Performance: The ResNet classifier achieves a final validation accuracy of 82.6% on the in-language validation set, with an F1-score of 0.81. The model shows balanced performance across HC and MCI, as evidenced by the validation confusion matrix presented in Table I. To contextualize ResNet performance, we trained baseline classifiers on the same features: Support Vector Machine (SVM) with RBF kernel, Logistic Regression, and Random Forest (100 trees). On cross-validation, Logistic Regression achieved 69.4% accuracy, Random Forest 73.1%, and SVM 75.6%. The ResNet's 82.6% validation accuracy thus reflects a statistically meaningful improvement over conventional baselines.

TABLE I
Validation Set Confusion Matrix

	Predicted HC	**Predicted MCI**
Actual HC	92 (84.4%)	17 (15.6%)
Actual MCI	14 (20.3%)	55 (79.7%)

2) Performance on Unseen Language: We further evaluate the trained model on Mandarin speech to assess generalization to an unseen language. A calibrated decision threshold of 0.5493 is applied, obtained from the temperature scaling procedure using the validation set. Specifically, this threshold corresponds to the probability at which the calibrated model's predicted scores achieve the best alignment with true labels. Using this threshold, the model is tested on a Mandarin dataset comprising healthy controls and individuals with MCI.

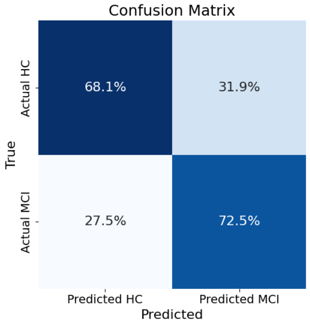

Fig. 2. Confusion matrix for the Mandarin evaluation set.

Fig. 2 presents the confusion matrix heatmap for the Mandarin evaluation set, an unseen language during training. The classifier correctly identifies 72.5% of individuals with MCI and 68.1% of HC, indicating strong performance despite the language mismatch. The false positive rate (31.9%), where HC individuals are misclassified as MCI, is slightly higher than the false negative rate (27.5%). It suggests a tendency toward over-detection, which may be preferable in a screening context to minimize missed MCI cases.

Fig. 3 illustrates the ROC and PR curves for the Mandarin evaluation set. The ROC curve (AUC = 0.70) suggests a reasonably good trade-off between the true positive rate and false positive rate, indicating that the classifier can distinguish between HC and MCI cases with moderate confidence. The PR curve further contextualizes this performance in the presence of class imbalance. It shows a gradual decline in precision as recall increases, which is an expected behavior when positive (MCI) cases are less frequent or more difficult to classify. The AUC of 0.72 for the PR curve indicates that the model maintains reasonable precision across a range of recall values, reinforcing the utility of post-temperature scaling calibration in improving probabilistic outputs for reliable downstream decision-making in multilingual screening applications.

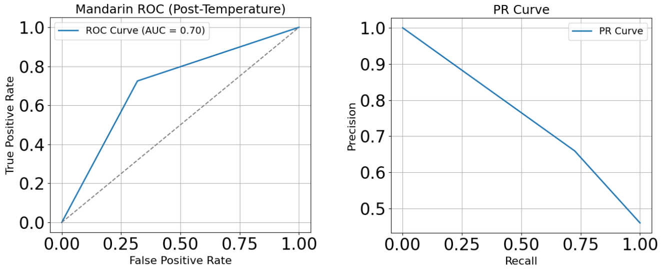

Fig. 3. ROC and Precision-Recall curves for the Mandarin evaluation set.

3) Feature Contribution Analysis: To understand feature relevance, we computed permutation-based feature importance using a random forest surrogate model. Although 18 features were originally extracted, only 13 appear in the importance plot because the remaining features contributed negligibly and were pruned during permutation analysis for clarity. As shown in Fig. 4, pitch-related features along with the MFCC, emerged as the top contributors to model performance. These features are closely associated with prosody, reflecting variations in vocal tone and intonation, which are known to be affected by cognitive decline. Additionally, spectral features such as SpectralFlux and SpectralSpread, and articulatory-related MFCCs (e.g., MFCC3) also showed notable importance. Interestingly, higher-order features like Normalized Tone Contrast (NTC) and Mean Phonation Duration (MPD) ranked lower, possibly due to their dependence on longer or more complex speech contexts. Overall, the results emphasize that low-level acoustic cues, particularly those related to pitch and spectral shape, are the most reliable indicators for distinguishing between HCs and individuals with MCI in a language-agnostic framework.

IV. Conclusion

We presented a language-agnostic speech-based dementia screening pipeline using acoustic and linguistic features extracted from dialogue. A ResNet-based model was trained on data from English, Spanish, and Greek speakers, then evaluated on Mandarin in a zero-shot setting. The system achieved a final validation accuracy of 82.6% on in-language data and a cross-lingual accuracy of 70.0% on Mandarin.

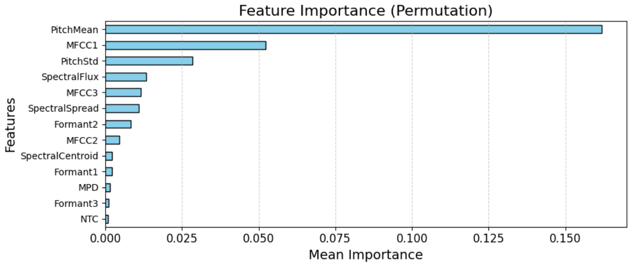

Fig. 4. Permutation-based feature importance rankings.

This work demonstrates that cognitive impairment manifests in speech through language-independent prosodic and articulatory features. Our system enables scalable MCI screening without language-specific retraining. Key contributions include a full audio preprocessing pipeline, cross-linguistic feature design, and probabilistic calibration via temperature scaling.

A key limitation of our study is the relatively small and imbalanced dataset, particularly in Spanish, Greek, and Mandarin. This constrains model generalization. Future research could leverage large-scale multilingual pretraining, semi-supervised approaches, or transfer learning from automatic speech recognition to mitigate data scarcity.

References

[1] World Health Organization. (2025, Mar.) Dementia. Accessed: 2025-06-07. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] C.-J. Chou, C.-T. Chang, Y.-N. Chang, C.-Y. Lee, Y.-F. Chuang, Y.-L. Chiu, W.-L. Liang, Y.-M. Fan, and Y.-C. Liu, "Screening for early alzheimer's disease: enhancing diagnosis with linguistic features and biomarkers," *Frontiers in Aging Neuroscience*, vol. 16, 2024. [Online]. Available: https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2024.1451326

[3] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, 2015.

[4] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The adresso challenge," *Frontiers in Computer Science*, 2021.

[5] F. Bertini, I. Palmisano, and M. Poesio, "A cross-language dementia classifier: a preliminary study," in *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics, 2022, pp. 512–517.

[6] Z. Wang, J. Xie, R. Xu, Y. Yang, G. Neubig, and J. Carbonell, "Cross-lingual alignment vs joint training: A comparative study and a simple unified framework," *arXiv preprint arXiv:1910.04708*, 2019.

[7] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's disease*, vol. 49, no. 2, pp. 407–422, 2015.

[8] T. Amorese, C. Greco, M. Cuciniello, R. Milo, O. Sheveleva, and N. Glackin, "Automatic speech recognition (asr) with whisper: Testing performances in different languages." in *S3C@ CHItaly*, 2023, pp. 1–8.

[9] espeak-ng contributors, "espeak ng text-to-speech," 2021, https://github.com/espeak-ng/espeak-ng.

[10] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[11] 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*, 2023. [Online]. Available: https://doi.org/10.1044/2022$_{A}JSLP-22-$00281