An LLM-based Temporal-spatial Data Generation and Fusion Approach for Early Detection of Late Onset Alzheimer's Disease (LOAD) Stagings Especially in Chinese and English-speaking Populations

Yang Han, Jacqueline CK Lam*, Victor OK Li* The University of Hong Kong {yhan, jcklam, vli}@eee.hku.hk

Lawrence YL Cheung The Chinese University of Hong Kong yllcheung@cuhk.edu.hk

Abstract

Alzheimer's Disease (AD), the 7th leading cause of death globally, demands scalable methods for early detection. While speech-based diagnostics offer promise, existing approaches struggle with temporal-spatial (T-S) challenges in capturing subtle linguistic shifts across different disease stages (temporal) and in adapting to cross-linguistic variability (spatial). This study introduces a novel Large Language Model (LLM)-driven T-S fusion framework that integrates multilingual LLMs, contrastive learning, and interpretable marker discovery to revolutionize Late Onset AD (LOAD) detection. Our key innovations include: (1) T-S Data Imputation: Leveraging LLMs to generate synthetic speech transcripts across different LOAD stages (NC, Normal Control; *e*MCI, early Mild Cognitive Impairment; *l*MCI, late Mild Cognitive Impairment; AD) and languages (Chinese, English, Spanish), addressing data scarcity while preserving clinical relevance (expert validation: 86% agreement with LLM-generated labels). (2) T-S Transformer with Contrastive Learning: A multilingual model that disentangles stage-specific (temporal) and language-specific (spatial) patterns, achieving a notable improvement of 10.9–24.7% in F1-score over existing baselines. (3) Cross-Linguistic Marker Discovery: Identifying language-agnostic markers and language-specific patterns to enhance interpretability for clinical adoption. By unifying temporal LOAD stages and spatial diversity, our framework achieves state-of-the-art performance in early LOAD detection while enabling cross-linguistic diagnostics. This study bridges NLP and clinical neuroscience, demonstrating LLMs' potential to amplify limited biomedical data and advance equitable healthcare AI.

1 Introduction

Alzheimer's Disease (AD) is the 7th leading cause of death worldwide. 95% of AD cases occur after age 65, i.e., Late Onset Alzheimer's Disease (LOAD). In this paper, LOAD refers to the disease category that includes four diagnostic stages: Normal Control (NC), early Mild Cognitive Impairment (*e*MCI), late Mild Cognitive Impairment (*l*MCI), and AD.

Low-cost and non-invasive approaches to LOAD detection and monitoring, especially via speech tests, offer promising solutions for population-based screening (Whelan et al., 2022). Previous studies have demonstrated the utility of speech-based neurodegenerative detection (Henderson et al., 2023; Patel et al., 2022), with cognitive-linguistic markers emerging as potentially effective markers to detect LOAD in an early stage (Eyigoz et al., 2020). However, despite this promise, current methods face three major limitations that arise from both the temporal (T) dimension (different LOAD stages) and the spatial (S) dimension (different language populations with distinct linguistic, cultural, and demographic factors). First, T–S resource constraints: the available speech datasets (e.g., DementiaBank (Lanzi et al., 2023)) remain small, fragmented, and predominantly in English, with insufficient samples for the early LOAD stages, especially *e*MCI and *l*MCI (Mueller et al., 2018). This data scarcity reduces the robustness of AI-driven approaches and constrains cross-linguistic generalizability. Second, T-S diagnostic gaps: while AI methods achieve high accuracy in distinguishing AD from healthy controls, performance drops significantly for earlier stages where subtle cognitive–linguistic changes are most critical for timely intervention (Petti et al., 2020). Moreover, most transformer-based diagnostic models remain language-specific, limiting applicability to diverse populations (Yang et al., 2022). Third, T–S speech marker gaps: although linguistic markers have been studied (Fraser et al., 2015), how speech markers differ across LOAD stages and whether they are shared or language-specific remain to be explored.

*Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14977–14990

To overcome these challenges, this study proposes an innovative Large Language Model (LLM)-driven T-S fusion approach to massively increase and fuse the T-S dimensions of a dataset comprising speech transcripts[1] from individuals with LOAD, and to identify the salient speech markers that characterize fine-grained LOAD stages, including NC, *e*MCI, *l*MCI, and AD, especially in Mandarin Chinese[2] and English populations. The key innovations of this study include: (1) leveraging multilingual LLMs for large-scale T-S LOAD data imputation across various LOAD stages and languages, forming a unified T-S imputed dataset; (2) training a T-S transformer model enhanced with contrastive learning to better distinguish LOAD stages and language groups in latent space; and (3) extracting speech markers that highlight cross-linguistic and stage-specific characteristics of LOAD development.

2 Related Work

2.1 T-S LOAD Speech Datasets

Several datasets have been created to support speech-based LOAD research. These datasets include connected speech data, demographics (e.g., age, gender, and education), and cognitive assessments (e.g., Mini-Mental Status Exam (MMSE)) (de la Fuente Garcia et al., 2020). One of the most widely used public database in speech-based LOAD research is DementiaBank (Lanzi et al., 2023). It contains audio recordings from the Pitt Corpus (Becker et al., 1994), derived from a longitudinal LOAD study with connected speech samples collected from subjects labelled AD, MCI, and NC. In Pitt Corpus, participants were tasked with the Cookie Theft description task, a standard task designed to elicit spontaneous speech. More recently, the TAUKADIAL Challenge is available in DementiaBank, making English and Chinese samples of connected speech for MCI detection publicly available (Luz et al., 2024).

Despite the increasing availability of speech-based datasets and applications, LOAD speech datasets remain T-S resource-constrained. Also, limited focus has been paid to early-stage diagnostics due to the lack of longitudinal/temporal LOAD speech data and labels, particularly those data samples that earmark the early stages of LOAD development (de la Fuente Garcia et al., 2020). Even worse, most LOAD speech samples were collected from the English-speaking cohorts, limiting their applicability to other language populations, especially the Chinese-speaking group (Qi et al., 2023). To facilitate low-resource language learning tasks, existing data alignment and fusion often involve integrating diverse data sources that span across multiple languages (Ranathunga et al., 2023). A large and integrated T-S dataset covering the entire spectrum of LOAD speeches across both the Chinese- and English-speaking populations and different stages is yet to be available.

The advancement of LLMs offers new opportunities to address the T-S resource-constrained challenge. LLMs can be used to encode biomedical knowledge and understand medical records (Singhal et al., 2023; Thirunavukarasu et al., 2023) across different languages (Wang et al., 2024). Recent works have leveraged the prior knowledge encoded in LLMs pre-trained on massive amounts of data for data generation in low-resource language scenarios (Lorandi and Belz, 2024; Ma et al., 2024; Nasution and Onan, 2024; Mo et al., 2024, 2025; Han et al., 2025; Li et al., 2025b). These studies provide new insights into how to utilize and enrich existing LOAD speech datasets in T-S resource-constrained settings. However, existing research has yet to utilize LLMs to generate cross-linguistic data across different stages of LOAD development. How to utilize resource-constrained datasets (e.g. DementiaBank) to guide the generation and fusion of T-S LOAD samples, especially in low-resource languages in the context of LOAD research such as Chinese, has not been investigated.

2.2 T-S LOAD Speech Diagnostics

Recent studies have highlighted the application of AI technologies in speech-based LOAD diagnostics using audio and text data (Li et al., 2021; de la Fuente Garcia et al., 2020; Petti et al., 2020; Yang et al., 2022). For example, text embeddings can discriminate AD patients from healthy controls (Agbavor and Liang, 2022), and when combined with audio features, they can further improve the accuracy of AD classification (Ilias and Askounis, 2022; Wang et al., 2021). Although these AI-driven speech-based LOAD studies have achieved outstanding accuracy (around 90% on average) in detecting AD, the temporal characteristic of LOAD speech across the four LOAD stages (NC, *e*MCI,

[1]This study focuses exclusively on the analysis of speech transcripts. No audio recordings were used. Throughout this paper, the term speech refers specifically to these transcripts.

[2]In this paper, Chinese refers to Mandarin Chinese.

*l*MCI, AD) has not been fully explored and understood. The accuracy of detecting the early stages of LOAD, i.e., *e*MCI and *l*MCI, based on speech samples alone, is comparatively less satisfactory and studied (Petti et al., 2020; Han et al., 2025). MCI stages are usually undetected, but their detection is critical for early intervention and treatment (Shankle et al., 2005). Unlocking the potential of AI for accurate detection of LOAD in its early stages remains a challenging task. On the one hand, subtle cognitive changes may not be as obvious in speech; on the other hand, training data from *e*MCI/*l*MCI is disproportionally much smaller than that of the AD stage.

Moreover, the spatial characteristic (i.e., multilingual nature) of LOAD speech has presented significant challenges for robust LOAD diagnostics in cross-linguistic contexts, especially for low-resource languages in LOAD research, such as Chinese. The characteristics of LOAD speeches vary across language and cultural populations due to linguistic and cultural variability. However, existing AI-driven speech-based models, such as transformer-based models, are typically language-specific (Yang et al., 2022). An AI-driven model that accounts for spatial variability, such as the differences in language structure and cultural variation (e.g., English vs. Chinese), is crucial for accurately detecting LOAD across different language populations. However, only a limited number of models are capable of capturing cross-linguistic similarities and differences for LOAD detection. For example, some work has been done using English-speaking LOAD corpus data to facilitate LOAD diagnostics, capitalizing on low-resource data obtained from other languages, such as Spanish and Chinese, via transfer learning and contrastive learning (Guo et al., 2020; Pérez-Toro et al., 2022). Nonetheless, the application of transformer models to speech-based LOAD diagnostics across both disease stages and languages remains underexplored.

2.3 T-S LOAD Speech Markers

Several studies have identified speech markers for LOAD detection. By analyzing a broad range of linguistic features, such as grammatical complexity, semantic content, and discourse coherence, in language samples from AD patients and healthy individuals, previous research showed that a combination of these features can effectively distinguish AD from normal aging (Fraser et al., 2015). When using NLP techniques to extract syntactic, lexical, and semantic features from speech transcripts, promising results in AD classification were achieved (Orimaye et al., 2014). Other studies focused more on specific linguistic deficiencies, including speech fluency (Campbell et al., 2021), word-finding difficulties (Georgiou et al., 2023), and semantic fluency (Olmos-Villaseñor et al., 2023). Verbal fluency tests were explored as a screening tool for neurodegenerative conditions, demonstrating effectiveness in detecting cognitive impairment (Pakhomov et al., 2010). Additionally, investigations into semantic and phonemic fluency deficits highlighted their potential as linguistic markers for AD (Kavé and Goral, 2016).

Despite these advancements, there is still limited understanding of the salient speech markers (a) characterizing different stages of LOAD and (b) distinguishing the Chinese-speaking population from the English-speaking one. Moreover, the extent to which these salient speech markers are shared across languages and specific to particular languages remains largely unknown.

3 Methodology

Our proposed methodology consists of three components: (1) LLM-based T–S speech data generation and fusion. We leverage multilingual LLMs to impute missing speech samples in both the temporal and spatial dimensions. Temporally, the imputation spans discrete LOAD stages (NC, *e*MCI, *l*MCI, AD), while spatially it extends across language populations (English: United States; Chinese: Taiwan; Spanish: Europe). The fused dataset integrates these temporally and spatially enriched samples into a unified corpus for downstream model training. (2) T–S transformer model with contrastive learning. We construct a transformer-based model, following a BERT-style architecture with multi-head self-attention layers, that projects all samples into a shared embedding space. Contrastive learning is applied at the embedding level, encouraging proximity between samples of the same LOAD stage and language while enforcing separation across different stages and languages. (3) Salient speech marker identification. We identify salient speech markers by applying SHapley Additive exPlanations (SHAP) to the trained T–S transformer. SHAP quantifies the contribution of individual features to model predictions, allowing us to interpret which speech markers most strongly influence the classification of LOAD stages across

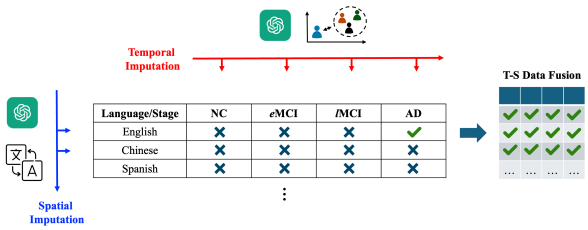

Figure 1: LLM-assisted T-S LOAD Speech Data Generation and Fusion

language populations.

3.1 LLM-based T-S LOAD Speech Data Generation and Fusion

Given a limited set of speech transcript samples, we perform T-S imputation, utilizing multilingual LLMs (e.g., GPT-4o, LLama, etc.) to generate new T-S speech transcripts of the same cognitive-linguistic task (i.e., Cookie Theft picture description) to impute missing transcripts across LOAD development stages (i.e., NC, *e*MCI, *l*MCI, and AD) and languages (Chinese, English, and Spanish) (see Figure 1). Our LLM-driven imputation leverages the extensive linguistic and clinical knowledge embedded in pre-trained LLMs, which have been trained on large and diverse language corpora. This approach enhances data diversity and clinical relevance, mitigating the constraints posed by limited real-world datasets.

First, we generate speech data in the temporal dimension by performing speech transcript imputation at each LOAD stage (four stages in total). Specifically, LOAD speech data are often cross-sectional, exhibiting high sparsity in the temporal dimension, e.g., one subject may only have speech samples at the AD stage but not MCI stages. First, we use LLM-based imputation to fill in the missing tabular values (e.g., MMSE) for each subject at each stage using the LLM-based imputation techniques developed by Li et al. (2024). For missing speech data, we randomly select a set of available transcripts from other subjects with similar characteristics. Specifically, subjects with similar demographic and cognitive characteristics (age, gender, and MMSE) are matched using a K-nearest neighbors (KNN) approach with K=3, where Euclidean distance is computed across these normalized variables to select the most similar samples for each target subject at each stage. For each case, the LLM is instructed to generate a new speech transcript rather than simply paraphrasing the exemplars. This approach promotes both demographic/clinical plausibility and linguistic diversity in the augmented dataset. We use the following LLM prompting strategy: *Based on a list of transcripts collected from other subjects of similar characteristics: [selected transcripts], generate a new speech transcript.*

Second, we generate speech data in the spatial dimension by performing speech transcript imputation for each linguistic population (three languages in total). The multilingual capabilities of LLMs have been utilized for translation-based text data generation across different language populations (Cahyawijaya et al., 2024). While our approach utilizes direct translation for data generation, we emphasize that LLMs, being trained on multilingual data, can inherently preserve language-specific patterns during translation. Specifically, by exploiting the multilingual capabilities of LLMs in processing different languages underlying the same semantic space, temporally imputed LOAD speech samples will be further translated into various languages, including Chinese, English, and Spanish. We use the following LLM prompting strategy accounting for the different language populations covering linguistic, cultural, and demographic factors: *Based on a subject's speech transcript: [transcript], translate it into a new speech transcript in [language], while considering the corresponding demographic and cultural factors (Age: [age], Gender: [gender], Region: [region]).*

After the temporal and spatial speech data imputation, a diverse set of new T-S LOAD speech transcripts is generated to cover different LOAD stages and language populations, significantly increasing the sample size and facilitating cross-stage and cross-lingual learning for the downstream task, i.e., LOAD diagnostics. The newly generated T-S LOAD speech samples are combined into one homogeneous dataset. Each speech sample is denoted as a triple: (T: LOAD stage, S: language, speech transcript).

3.2 T-S Transformer-based Diagnostic Model Development

We develop a transformer-based model to predict the diagnosis label (i.e., LOAD stage). The backbone of the transformer model is based on a pre-trained language model, e.g., BERT (Devlin et al., 2019). The speech transcripts are fed into the transformer model to output unified text embeddings. By projecting speech samples into a shared high-dimensional space, the model enables quantitative comparison between LOAD speech samples, mak-

ing it possible to compute distances between samples, facilitating cross-stage (temporal) and cross-lingual (spatial) analysis by representing the similarity and difference across speech samples of different LOAD stages and populations (see Figure 2).

We utilize contrastive learning to further align samples from different LOAD stages and language populations within the embedding space. Contrastive learning aims to guide the model to minimize the distance between similar samples and maximize the distance between dissimilar ones, making it particularly effective when data are scarce, imbalanced, or noisy. For instance, speech samples from earlier LOAD stages within the same language will be closer together in the embedding space than those from later stages. Specifically, let each transcript x_i be mapped by the T-S transformer encoder into an embedding $\mathbf{z}_i \in R^d$, normalized such that $\|\mathbf{z}_i\| = 1$. Similarity between two embeddings is measured via cosine similarity: $\text{sim}(\mathbf{z}_i, \mathbf{z}_j) = \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\|\mathbf{z}_i\| \, \|\mathbf{z}_j\|}$. We define positive pairs as transcripts from the same LOAD stage and language, while negative pairs are transcripts from different LOAD stages or languages. For an embedding $\mathbf{z}_i$ with a set of positives $P(i)$, the contrastive loss is calculated as:

$$-\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_p)/\tau)}{\sum_{a \in A(i)} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_a)/\tau)}$$

where $\tau > 0$ is a temperature parameter, and $A(i)$ denotes the set of all samples in the batch excluding i. The denominator incorporates both positive and negative samples, so minimization encourages embeddings of positives to cluster while simultaneously pushing negatives apart.

The transformer model is trained on the unified T-S imputed dataset, optimized to generate embeddings that align similar samples and distinguish dissimilar ones across LOAD stages and languages. Specifically, a classification head is added to the transformer model. The model is fine-tuned for LOAD stage classification with four labels, including NC, *e*MCI, *l*MCI, and AD. The primary objective is to classify LOAD stages, while the contrastive loss is incorporated, enhancing distinctions between LOAD stages and reinforcing cross-lingual generalizability. This process helps the model learn salient speech markers that are both specific and shared across languages, improving its performance in LOAD stage classification.

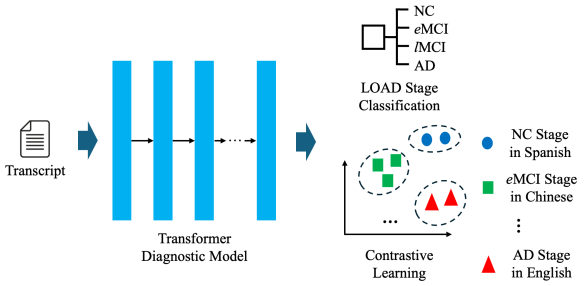

Figure 2: T-S Transformer Diagnostic Model Development with Contrastive Learning

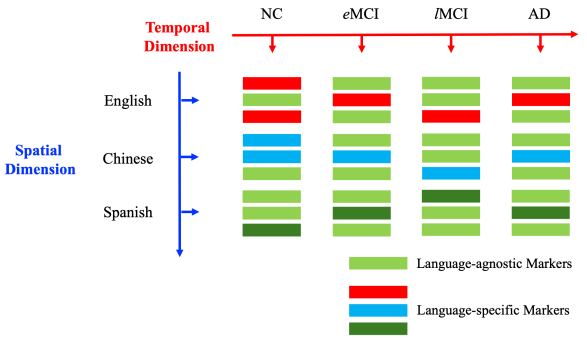

Figure 3: Salient Speech Marker Identification Across LOAD Stages and Linguistic Populations

3.3 Salient Speech Marker Identification Across LOAD Stages and Language Populations

Feature importance analysis is performed to uncover the most salient speech markers associated with different LOAD stages. SHAP analysis developed by Lundberg and Lee (2017) is used to assess the contribution of each speech transcript feature j to the model's output $f(x)$, i.e., the classification of each LOAD stage, such as *e*MCI, *l*MCI, and AD (see Figure 3). For a given input x, the SHAP value ϕ_j of feature j is defined as:

$$\sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!} \left[f_{S \cup \{j\}}(x_{S \cup \{j\}}) - f_S(x_S) \right]$$

where F is the full feature set and S is a subset of features excluding j. Intuitively, ϕ_j measures the average change in model prediction when j is added to subsets S of other features.

The salient speech markers across different LOAD stages for the same language are identified. Based on the salient marker identification results, the temporal similarity and difference across different LOAD stages for the same language are analyzed and interpreted. In addition, the spatial similarity and difference between Chinese-, English-, ans Spanish-speaking cohorts for the same LOAD stage are analyzed and interpreted.

4 Experimental Setup

4.1 Datasets

Connected speech samples in English, Chinese, and Spanish were collected from DementiaBank[3] (Lanzi et al., 2023), including LOAD speech samples across different stages. Moreover, age, gender, education (in years), and Mini-Mental State Examination (MMSE) scores were collected. Specifically, the Pitt Corpus (English only) (Becker et al., 1994) and the TAUKADIAL Challenge (Chinese and English) (Luz et al., 2024) datasets from the DementiaBank were included. We also included a Spanish LOAD speech dataset (the Ivanova Corpus) from DementiaBank (Ivanova et al., 2022) to increase spatial diversity and the sample size for training. After data collection, each speech recording sample was associated with the corresponding subject information and diagnosis label (NC/MCI/AD). The total number of speech samples was 1282. 261 were Chinese-speaking samples, 663 were English-speaking samples, and 358 were Spanish-speaking samples. The number of NC, MCI, and AD samples were 595, 401, 286, respectively.

4.2 Data Preprocessing

The transcripts of speech recordings were extracted using OpenAI's Whisper model (Radford et al., 2023). The Whisper model, pre-trained on internet data such as YouTube, is known to occasionally hallucinate content, particularly in non-speech segments (e.g., inserting phrases such as "thank you for watching") (Koenecke et al., 2024). To mitigate this risk, we used the recommended decoding parameters (log probability threshold: –1; no speech threshold: 0.6) (Radford et al., 2023) for more reliable transcription. Furthermore, we manually inspected a randomly selected subset of transcripts and did not observe hallucinated outputs. Speaker diarization (Bredin, 2023) was used to remove the examiner's speech. These preprocessed speech transcripts were used as inputs for LOAD classification model training and evaluation.

4.3 MCI Staging

Fine-grained MCI stages, i.e., early MCI (*e*MCI) and late MCI (*l*MCI), were defined based on the Alzheimer's Disease Neuroimaging Initiative (ADNI) study (Edmonds et al., 2019). Based on the matched ADNI subject characteristics, MCI labels from DementiaBank were further categorized into *e*MCI and *l*MCI, with reference to their corresponding MMSE scores after controlling for age, gender, and education. After MCI staging, the number of NC, *e*MCI, *l*MCI, and AD samples were 595, 129, 272, 286, respectively.

4.4 Baseline Selection and Abalation Study

For baseline comparison, we included the multilingual BERT model (about 117M parameters) (Devlin et al., 2019), along with two domain-specific BERT variants: BioBERT (Lee et al., 2020) and ClinicalBERT (Huang et al., 2019), which are pretrained on biomedical and clinical text corpora. For the ablation study, we used the best-performing model in the baseline comparison, as the base architecture. We first assessed LOAD diagnostic performance using different training language configurations (Chinese-only, English-only, and combined multilingual data) to evaluate the effect of linguistic diversity on model performance. Then, we introduced T-S fusion, followed by contrastive learning, to quantify their individual contributions to LOAD diagnostic performance.

4.5 Evaluation Metrics

Using a stratified dataset split, we held out 20% of the original speech samples to obtain the testing set. The remaining speech samples, including both original and new speech samples generated by T-S data fusion, were used for model training and selection. All evaluations were conducted exclusively on the held-out testing set composed only of real speech samples from the original dataset, with no synthetic or imputed data in the testing set. The purpose of T-S data fusion was to enhance model learning, but the test of diagnostic performance was performed on real data.

We selected the best LOAD classification model using a validation set derived from the training data and evaluated the fine-tuned model on the testing set. All baseline models were fine-tuned using the same data splits and hyperparameters. The best model for each baseline was selected based on validation performance. This was repeated three times, and we reported the average performance metrics. We used accuracy and F1 score for model performance evaluation. They are two commonly used evaluation metrics in classification tasks. Accuracy, ranging from 0 to 1 (the higher, the better), measures the percentage of correctly detected cases. F1 score, also ranging from 0 to 1 (the higher, the better), combines the precision (positive predictive

[3] https://dementia.talkbank.org/

value) and recall (sensitivity) scores and provides a more comprehensive evaluation of detection accuracy. We used weighted F1 score, which accounts for class imbalances by averaging the F1 scores of all classes while weighting them by their respective support (i.e., the number of true instances for each class).

4.6 Experimental Settings

For T-S data imputation, we generated new speech transcript samples based on the latest GPT-4o model via OpenAI's Chat Completion API. A fixed temperature of 1 was used for text generation.

We fine-tuned the pre-trained BERT models with a learning rate of 1e-5 and a weight decay of 0.01 after performing trials with a range of hyperparameter values (learning rate of 1e-3, 1e-4, and 1e-5, and weight decay of 0.1 and 0.01). For contrastive learning, the temperature parameter was set to 0.1. The training process consisted of 10 epochs with a batch size of 16. The best proposed and baseline models were selected based on the accuracy of the LOAD classification task on the validation set.

All experiments were carried out using a Nvidia A100 40GB GPU on a Linux system through Google Colab, with Python (version 3.11.11), and deep learning packages, including PyTorch (version 2.5.1+cu124) and Transformers (version 4.48.3). The total computational budget was approximately 5 GPU hours. The pre-trained multilingual BERT model was obtained from HuggingFace (Apache 2.0 license).

5 Results and Discussion

5.1 Performance Comparison

Table 1 compares the performance of three pretrained transformer-based language models, including BioBERT, ClinicalBERT, and Multilingual BERT, on the original dataset for classifying subjects into four LOAD stages. Among these, Multilingual BERT achieves the best performance, with an accuracy of 56.6% and an F1 score of 49.4%. While the differences across models are relatively small, the results suggest that broad cross-lingual pretraining enables better generalization compared to domain-specific models such as BioBERT and ClinicalBERT. Despite their specialization in biomedical and clinical language, these models may lack the flexibility needed to capture the linguistic nuances associated with the varying stages of cognitive decline in speech-based data.

To further analyze the effect of training language, we evaluate Multilingual BERT using language-specific subsets of the original data (Table 2). Training on Chinese-only data yields the lowest performance (47.3% accuracy, 35.6% F1 score), while training on English-only data results in moderate improvement (54.2% accuracy, 40.9% F1 score). Notably, combining all language sets in a multilingual training configuration leads to further performance gains (56.6% accuracy, 49.4% F1 score). This demonstrates that cross-lingual training enhances the model's capacity to generalize, likely due to the increased variability in linguistic and cognitive expression. These findings reinforce the advantage of using Multilingual BERT, which is better equipped to process multilingual input in a unified representation space.

Furthermore, building on the multilingual baseline, we assess the impact of T-S fusion and contrastive learning (Table 2). Adding T-S fusion (covering English, Chinese, and Spanish languages) to the model increases accuracy from 56.6% to 61.7%, and F1 score from 49.4% to 57.3%. This indicates that incorporating T-S fusion data helps the model better distinguish between stages. When contrastive learning is added to the T-S Fusion model, accuracy further improves to 62.9%, with a notable F1 score increase to 60.3%. This suggests that contrastive learning enables the model to learn more discriminative representations, reinforcing inter-class separability and improving stage-level classification performance.

While the accuracy improvements are modest, the F1 score increases are substantial (10.9%-24.7%), highlighting improved classification balance across all LOAD stages. Compared to the Chinese-only baseline, the best model achieves a 24.7% gain in F1 score. Compared to the English-only baseline, the best model achieves a 19.4% gain in F1 score. Compared to the strongest baseline using multilingual data without T-S fusion and contrastive learning, the best model improves the F1 score by 10.9%. These results underscore the cumulative benefit of multilingual training, T-S fusion, and contrastive learning.

Furthermore, Table 5 shows the detailed breakdown for each ablation step. Since our model was trained as a multi-class classifier, per-stage F1 scores were computed in a standard one-vs-rest manner (i.e., treating each stage as positive and all others as negative) to provide a granular view of performance for each LOAD stage. These re-

sults consistently show that both T-S fusion and contrastive learning contribute positively across stages. Specifically, the contributions of T-S fusion in F1 gains are 3.9%, 5.7%, 16.4%, 9.1% for NC, *e*MCI, *l*MCI, and AD stages, respectively. The contributions of contrastive learning are 1.5%, 11.5%, 6.7%, and -1.6%, for NC, *e*MCI, *l*MCI, and AD stages, respectively. Notably, T-S fusion is especially beneficial for the more challenging MCI stages, and contrastive learning shows its greatest effect in early-stage classification (*e*MCI), with a slight trade-off in the latest stage (AD).

Base Model	Avg. Accuracy (%)	Avg. F1 (%)
BioBERT	55.6	46.9
ClinicalBERT	56.2	48.2
Multilingual BERT	**56.6**	**49.4**

Table 1: Backbone Comparision Using Original Data for LOAD Classification Across Different Stages (NC/*e*MCI/*l*MCI/AD)

Configuration	Avg. Accuracy (%)	Avg. F1 (%)
Original Data (Chinese)	47.3	35.6
Original Data (English)	54.2	40.9
Original Data (All)	56.6	49.4
+ T-S Fusion	61.7	57.3
+ T-S Fusion + CL	**62.9**	**60.3**

Table 2: Ablation Study Using Multilingual BERT as Base Model. CL: Contrastive Learning.

*e*MCI	Score	*l*MCI	Score	AD	Score
有(having)	0.06	有(having)	0.08	有(having)	0.12
場景(scene)	0.05	廚房(kitchen)	0.07	場景(scene)	0.12
廚房(kitchen)	0.05	場景(scene)	0.07	瞬間(sudden)	0.11
水池(sink)	0.04	情景(scene)	0.06	情景(scene)	0.10
男孩(boy)	0.03	水池(sink)	0.06	水池(sink)	0.09
情景(scene)	0.03	瞬間(sudden)	0.06	廚房(kitchen)	0.08
場景(scene)	0.03	吧(ba)	0.06	男孩(boy)	0.08
瞬間(sudden)	0.03	男孩(boy)	0.05	等等(etc.)	0.07
情況(situation)	0.02	情況(situation)	0.04	情況(situation)	0.07
等等(etc.)	0.02	等等(etc.)	0.04	正試圖(trying)	0.07

Table 3: The Most Salient Speech Markers Among the Chinese-speaking Cohorts Using SHAP Analysis (English Translations in Parentheses)

*e*MCI	Score	*l*MCI	Score	AD	Score
outside	0.15	outside	0.22	outside	0.41
faucet	0.10	faucet	0.16	faucet	0.24
cupboard	0.05	cupboard	0.08	happening	0.18
see	0.04	happening	0.07	cupboard	0.18
happening	0.04	trouble	0.06	see	0.18
trouble	0.03	see	0.04	okay	0.12
kitchen	0.03	quiet	0.03	bushes	0.10
woman	0.02	kitchen	0.03	better	0.09
quiet	0.02	woman	0.02	trouble	0.09
okay	0.01	okay	0.02	ah	0.09

Table 4: The Most Salient Speech Markers Among the English-speaking Cohorts Using SHAP Analysis

5.2 Salient Speech Markers

Tables 3, 4, and 6 show the most salient speech markers identified by SHAP analysis across LOAD stages and linguistic populations. Section C in the Appendix provides more detailed discussions.

For the Chinese population, across all LOAD stages, common words such as "having", "scene", "sink", and "boy" remain consistent, reflecting reliance on familiar, concrete vocabulary, a well-documented feature of dementia discourse (Boschi et al., 2017). In particular, the persistence of "etc." across all stages suggests a compensatory strategy for lexical retrieval failures, consistent with semantic generalization patterns observed in MCI and AD (Kavé and Goral, 2016). For the *l*MCI stage, the emergence of "ba", a discourse particle, suggests increased reliance on filler words to compensate for word-finding difficulties (Chou et al., 2024). For the AD stage, general verbs such as "trying" emerge, indicating an increased struggle with verbal planning and execution and consistent with prior reports of syntactic simplification and verb generalization in advanced cognitive decline (Williams et al., 2023).

For the English population, across all LOAD stages, common words, such as "outside", "faucet", and "cupboard", persist, indicating preserved use of familiar and concrete terms, a pattern well documented in Alzheimer's discourse studies (Boschi et al., 2017). In particular, the use of "okay" is also notable, as it may also serve as a filler or conversational placeholder, similar to how other fillers (e.g. "um", "uh") are used more frequently by individuals with cognitive decline (Fraser et al., 2015). The speech markers are similar across the *e*MCI and *l*MCI stages, although their rankings differ. In the AD stage, new speech markers emerge, such as "ah", signaling increased disfluency and hesitation, along with others like "bushes" and "better", reflecting that the use of adjectives and nouns in speech can reveal important aspects of cognitive decline, e.g., unrelated nouns and non-specific adjectives (Fraser et al., 2015).

For the Spanish population, salient markers shift from abstract and emotionally nuanced terms in the *e*MCI stage, such as "sensación" (feeling), "situación" (situation), and "vida" (life), toward more routine and generalized vocabulary in later stages. During *l*MCI, words like "quehaceres" (chores), "cocina" (kitchen), and "niños" (children) emerge, reflecting increased reliance on fa-

miliar and concrete concepts. This trajectory aligns with evidence that in LOAD, semantic network breakdown begins with specific attributes and concrete concepts, while broader categories and emotional associations are preserved until later stages (Martínez-Nicolás et al., 2019). By the AD stage, the frequent use of broad terms such as "cotidiana" (everyday), "todo" (everything), and "actividad" (activity) indicates reduced lexical diversity and growing difficulty with precise word retrieval (Cuetos et al., 2003).

5.3 LLM-generated Data Validation

Furthermore, we have performed a rigorous validation on the LLM-generated text. Specifically, a blinded expert evaluation on randomly selected 100 LLM-generated and 100 real speech transcripts across four LOAD stages was performed. The expert judgments conform with the LOAD stages labeled by LLM 86% of the time. This result has demonstrated that the generated samples can reflect meaningful cognitive markers consistent with different LOAD stage.

Specifically, clinical plausibility was judged by whether the linguistic features (e.g., grammatical complexity, lexical diversity, semantic appropriateness, and disfluency) in each transcript matched the expected patterns for the given LOAD stage. Expert validation was conducted by individuals with multilingual proficiency in both Chinese and English, including one linguist with experience in dementia language analysis, to ensure cross-linguistic appropriateness. The agreement rates between expert judgment and the stage labels of the LLM-generated data were 96% (NC), 92% (*e*MCI), 73% (*l*MCI), and 83% (AD). Most disagreements occurred between *l*MCI and neighboring stages, reflecting the inherent diagnostic ambiguity at these transitions.

6 Discussion and Conclusion

In this study, we introduce an innovative framework that leverages LLMs for the early detection of LOAD across Chinese, English, and Spanish-speaking populations. The primary aim is to address the significant challenges posed by temporal (different LOAD stages) and spatial (different language populations) dimensions in current diagnostic methods, which often lack scalability and precision.

The main challenges include the scarcity of longitudinal datasets that capture subtle linguistic shifts across different stages of LOAD, and the difficulty of adapting models to diverse linguistic and cultural contexts, particularly in low-resource languages such as Chinese. These constraints hinder effective early-stage detection, which is critical for timely intervention and treatment.

To overcome these challenges, our study introduces several key innovations. First, we apply LLMs for T-S data imputation, generating synthetic speech transcripts validated with 86% expert agreement, thereby addressing data scarcity. Second, we develop a T-S transformer model enhanced with contrastive learning, which significantly improves the model's ability to distinguish stage-specific and language-specific patterns, achieving F1-score gains of 10.9–24.7% over existing baselines. Lastly, we identify both cross-linguistic and language-specific speech markers, enhancing the interpretability and clinical utility of our approach.

Our T-S fusion LOAD detection framework has outperformed baseline models such as BioBERT and ClinicalBERT by 10.9-24.7% in F1 score. Salient speech markers include increased use of fillers and action verbs in Chinese, reliance on concrete nouns and disfluencies in English, and a shift from abstract to generic terms in Spanish as cognitive decline progresses.

Future work can integrate complementary signals such as genetic markers (Li et al., 2025a) and explore advanced spatial-temporal imputation strategies from other domains, such as air pollution data reconstruction (Yu et al., 2023, 2025), to further enhance the robustness and generalizability of T–S fusion for LOAD detection.

The impacts of this work are multifaceted. Clinically, it provides a scalable, non-invasive tool for early LOAD detection, which is essential for timely intervention. Methodologically, it demonstrates the potential of LLMs to enrich limited biomedical datasets, supporting more equitable diagnostics across languages and settings. This study underscores the importance of integrating linguistic diversity into AI-driven healthcare solutions, enhancing both diagnostic accuracy and cross-cultural applicability. By bridging natural language processing and clinical neuroscience, our framework offers a promising approach to early Alzheimer's detection, paving the way for personalized medicine and continuous monitoring of neurodegenerative disease stages over time.

Limitations

In this study, T-S fusion shows strong potential for advancing speech-based diagnostics of LOAD by integrating data across various disease stages and linguistic populations. While T-S fusion currently leverages imputed speech data to mitigate challenges in real-world data availability, this approach also highlights an exciting opportunity: the ability to generate high-quality, representative samples from limited datasets. These imputed samples, informed by existing datasets and pre-trained models, serve as a valuable foundation for developing scalable diagnostic tools, especially in underrepresented regions or populations.

The effectiveness of T-S imputation and fusion opens the door to further enhancements as more real-world LOAD speech data becomes available. Rather than viewing the reliance on imputed data as a constraint, it can be seen as a stepping stone toward building a robust framework that continuously improves in accuracy and generalizability. Expanding the collection of diverse, real-world speech samples will directly support the refinement of T-S fusion methods and ensure that the identified speech markers are both reliable and broadly applicable. This presents a clear path for future work: by prioritizing the acquisition of speech data across different languages and stages of disease progression, researchers can unlock new possibilities for inclusive and scalable diagnostic tools.

Looking ahead, the salient speech markers identified through T-S fusion offer a rich resource for iterative model development. These markers can guide adaptive refinement of the model, especially when new data from underrepresented linguistic or demographic groups is introduced. Embedding these markers within a feedback-driven learning framework, where model outputs are continuously validated and updated using real-world clinical data, can further enhance performance and adaptability. Such an iterative, data-driven approach not only boosts diagnostic accuracy but also supports the model's evolution alongside emerging research and clinical practices.

Finally, we recognize that prospective clinical validation is essential for demonstrating real-world utility. In the next stage of our work, we will explicitly pursue this as a key direction by collaborating with clinical partners to validate our findings in independent and prospective cohorts. Such validation is critical to translating T–S fusion into clinically actionable tools for early detection, personalized diagnostics, and longitudinal monitoring of LOAD, paving the way for broader impact in neurodegenerative disease care.

Ethical Statement

One key consideration in the effectiveness of T-S data fusion for LOAD diagnostics is the potential risks due to bias in model training. Since T-S data fusion relies on pre-trained LLMs, the speech data it generates and the speech markers it identifies are inherently shaped by the data these models were trained on. If the training data lack diversity in terms of demographics, language variations, or disease stages, our proposed approach may identify speech markers that do not generalize well across all populations, leading to disparities in LOAD diagnostic accuracy.

Acknowledgments

ChatGPT-4o was used to improve the language of the manuscript and to assist in the writing of the manuscript (only covering literature review and discussion of results). The authors reviewed the content generated and took full responsibility for the content of the manuscript.

This work was supported in part by the United States National Academy of Medicine Healthy Longevity Catalyst Award (Grant No. HLCA/E-705/24), administered by the Research Grants Council of Hong Kong, awarded to V.O.K.L. and J.C.K.L, and by The Hong Kong University Seed Funding for Collaborative Research 2023 (Grant No. 109000447), awarded to V.O.K.L. and J.C.K.L.

References

Felix Agbavor and Hualou Liang. 2022. Predicting dementia from spontaneous speech using large language models. *PLOS Digital Health*, 1(12):e0000168.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Veronica Boschi, Eleonora Catricala, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F Cappa. 2017. Connected speech in neurodegenerative language disorders: a review. *Frontiers in Psychology*, 8:269.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Interspeech*, pages 1983–1987. ISCA.

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2024. High-dimension human value representation in large language models. *Preprint*, arXiv:2404.07900.

Edward L Campbell, Raul Yanez Mesia, Laura Docio-Fernandez, and Carmen Garcia-Mateo. 2021. Paralinguistic and linguistic fluency features for Alzheimer's disease detection. *Computer Speech & Language*, 68:101198.

Chia-Ju Chou, Chih-Ting Chang, Ya-Ning Chang, Chia-Ying Lee, Yi-Fang Chuang, Yen-Ling Chiu, Wan-Lin Liang, Yu-Ming Fan, and Yi-Chien Liu. 2024. Screening for early alzheimer's disease: enhancing diagnosis with linguistic features and biomarkers. *Frontiers in Aging Neuroscience*, 16:1451326.

Fernando Cuetos, Teresa Martinez, Carmen Martinez, Cristina Izura, and Andrew W Ellis. 2003. Lexical processing in spanish patients with probable alzheimer's disease. *Cognitive Brain Research*, 17(3):549–561.

Sofia de la Fuente Garcia, Craig W Ritchie, and Saturnino Luz. 2020. Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. *Journal of Alzheimer's Disease*, 78(4):1547–1574.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Emily C Edmonds, Carrie R McDonald, Anisa Marshall, Kelsey R Thomas, Joel Eppig, Alexandra J Weigand, Lisa Delano-Wood, Douglas R Galasko, David P Salmon, Mark W Bondi, et al. 2019. Early versus late MCI: Improved MCI staging using a neuropsychological approach. *Alzheimer's & Dementia*, 15(5):699–708.

Elif Eyigoz, Sachin Mathur, Mar Santamaria, Guillermo Cecchi, and Melissa Naylor. 2020. Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine*, 28.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2015. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's disease*, 49(2):407–422.

Eleni-Zacharoula Georgiou, Maria Skondra, Marina Charalampopoulou, Panagiotis Felemegkas, Asimina Pachi, Georgia Stafylidou, Dimitrios Papazachariou, Robert Perneczky, Vasileios Thomopoulos, Antonios Politis, et al. 2023. Validation of the test for finding word retrieval deficits (WoFi) in detecting Alzheimer's disease in a naturalistic clinical setting. *European Journal of Ageing*, 20(1):29.

Zhiqiang Guo, Zhaoci Liu, Zhenhua Ling, Shijin Wang, Lingjing Jin, and Yunxia Li. 2020. Text classification by contrastive learning and cross-lingual data augmentation for Alzheimer's disease detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6161–6171.

Yang Han, Jacqueline C. K. Lam, Victor O. K. Li, and Lawrence Y. L. Cheung. 2025. A large language model based data generation framework to improve mild cognitive impairment detection sensitivity. *Data & Policy*, 7:e33.

Shalom K Henderson, Katie A Peterson, Karalyn Patterson, Matthew A Lambon Ralph, and James B Rowe. 2023. Verbal fluency tests assess global cognitive status but have limited diagnostic differentiation: evidence from a large-scale examination of six neurodegenerative diseases. *Brain Communications*, 5(2):fcad042.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Loukas Ilias and Dimitris Askounis. 2022. Multimodal deep learning models for detecting dementia from speech and transcripts. *Frontiers in Aging Neuroscience*, 14:830943.

Olga Ivanova, Juan José G Meilán, Francisco Martínez-Sánchez, Israel Martínez-Nicolás, Thide E Llorente, and Nuria Carcavilla González. 2022. Discriminating speech traits of Alzheimer's disease assessed through a corpus of reading task for Spanish language. *Computer Speech & Language*, 73:101341.

Gitit Kavé and Mira Goral. 2016. Word retrieval in picture descriptions produced by individuals with Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology*, 38(9):958–966.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless Whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681.

Alyssa M Lanzi, Anna K Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, and Matthew L Cohen. 2023. DementiaBank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*, 32(2):426–438.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Victor O. K. Li, Yang Han, and Jacqueline C. K. Lam. 2025a. Unravelling causal genetic biomarkers of Alzheimer's disease via neuron to gene-token backtracking in neural architecture: A groundbreaking

Reverse-Gene-Finder approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18466–18475.

Victor O. K. Li, Yang Han, Jacqueline C. K. Lam, and Lawrence Y. L. Cheung. 2025b. Reverse-Speech-Finder: A neural network backtracking architecture for generating Alzheimer's disease speech samples and improving diagnosis performance. *Preprint*, arXiv:2505.17477.

Victor O. K. Li, Jacqueline C. K. Lam, and Yang Han. 2024. LMP-TX: An AI-driven integrated longitudinal multi-modal platform for early prognosis of late onset Alzheimer's disease. *Preprint*, medRxiv:2024.10.02.24314019.

Victor O. K. Li, Jacqueline C. K. Lam, Yang Han, Lawrence Y. L. Cheung, Jocelyn Downey, Tushar Kaistha, and Illana Gozes. 2021. Designing a protocol adopting an artificial intelligence (AI)–driven approach for early diagnosis of late-onset Alzheimer's disease. *Journal of Molecular Neuroscience*, 71(7):1329–1337.

Michela Lorandi and Anya Belz. 2024. High-quality data-to-text generation for severely under-resourced languages with out-of-the-box large language models. *Preprint*, arXiv:2402.12267.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Preprint*, arXiv:1705.07874.

Saturnino Luz, Sofia De La Fuente Garcia, Fasih Haider, Davida Fromm, Brian MacWhinney, Alyssa Lanzi, Ya-Ning Chang, Chia-Ju Chou, and Yi-Chien Liu. 2024. Connected speech-based cognitive assessment in Chinese and English. *Preprint*, arXiv:2406.10272.

Mingyu Derek Ma, Xiaoxuan Wang, Po-Nien Kung, P Jeffrey Brantingham, Nanyun Peng, and Wei Wang. 2024. STAR: Boosting low-resource information extraction by structure-to-text data generation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18751–18759.

Israel Martínez-Nicolás, Juan Carro, Thide E Llorente, and Juan José García Meilán. 2019. The deterioration of semantic networks in alzheimer's disease. *Exon Publications*, pages 179–191.

Tingyu Mo, Jacqueline C. K. Lam, Victor O. K. Li, and Lawrence Y. L. Cheung. 2024. Leveraging large language models for identifying interpretable linguistic markers and enhancing Alzheimer's disease diagnostics. *Preprint*, medRxiv:2024.08.22.24312463.

Tingyu Mo, Jacqueline C. K. Lam, Victor O. K. Li, and Lawrence Y. L. Cheung. 2025. DECT: Harnessing llm-assisted fine-grained linguistic knowledge and label-switched and label-preserved data generation for diagnosis of Alzheimer's disease. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24885–24892.

Kimberly D Mueller, Bruce Hermann, Jonilda Mecollari, and Lyn S Turkstra. 2018. Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *Journal of Clinical and Experimental Neuropsychology*, 40(9):917–939.

Arbi Haza Nasution and Aytuğ Onan. 2024. ChatGPT label: Comparing the quality of human-generated and LLM-generated annotations in low-resource language NLP tasks. *Ieee Access*, 12:71876–71900.

Rocio Olmos-Villaseñor, Consuelo Sepulveda-Silva, Teresa Julio-Ramos, Eduardo Fuentes-Lopez, David Toloza-Ramirez, Rodrigo A Santibañez, David A Copland, and Carolina Mendez-Orellana. 2023. Phonological and semantic fluency in Alzheimer's disease: a systematic review and meta-analysis. *Journal of Alzheimer's Disease*, 95(1):1–12.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Karen Jennifer Golden. 2014. Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 78–87.

Serguei VS Pakhomov, Glenn E Smith, Dustin Chacon, Yara Feliciano, Neill Graff-Radford, Richard Caselli, and David S Knopman. 2010. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology*, 23(3):165–177.

Nikil Patel, Katie A Peterson, Ruth U Ingram, Ian Storey, Stefano F Cappa, Eleonora Catricala, Ajay Halai, Karalyn E Patterson, Matthew A Lambon Ralph, James B Rowe, et al. 2022. A 'Mini Linguistic State Examination' to classify primary progressive aphasia. *Brain Communications*, 4(2):fcab299.

Paula Andrea Pérez-Toro, Philipp Klumpp, Abner Hernandez, Tomas Arias, Patricia Lillo, Andrea Slachevsky, Adolfo Martín García, Maria Schuster, Andreas K Maier, Elmar Noeth, et al. 2022. Alzheimer's detection from English to Spanish using acoustic and linguistic embeddings. In *Interspeech*, pages 2483–2487.

Ulla Petti, Simon Baker, and Anna Korhonen. 2020. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association*, 27(11):1784–1797.

Xiaoke Qi, Qing Zhou, Jian Dong, and Wei Bao. 2023. Noninvasive automatic detection of Alzheimer's disease from spontaneous speech: a review. *Frontiers in Aging Neuroscience*, 15:1224723.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

William R Shankle, A Kimball Romney, Junko Hara, Dennis Fortier, Malcolm B Dick, James M Chen, Timothy Chan, and Xijiang Sun. 2005. Methods to improve the detection of mild cognitive impairment. *Proceedings of the National Academy of Sciences*, 102(13):4919–4924.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940.

Ning Wang, Yupeng Cao, Shuai Hao, Zongru Shao, and KP Subbalakshmi. 2021. Modular multi-modal attention network for Alzheimer's disease detection using patient audio and language data. In *Interspeech*, pages 3835–3839.

Teng Wang, Zhenqi He, Wing-Yin Yu, Xiaojin Fu, and Xiongwei Han. 2024. Large language models are good multi-lingual learners: When LLMs meet cross-lingual prompts. *Preprint*, arXiv:2409.11056.

Robert Whelan, Florentine M Barbey, Marcia R Cominetti, Claire M Gillan, and Anna M Rosická. 2022. Developments in scalable strategies for detecting early markers of cognitive decline. *Translational Psychiatry*, 12(1):473.

Eric Williams, Catherine Theys, and Megan McAuliffe. 2023. Lexical-semantic properties of verbs and nouns used in conversation by people with Alzheimer's disease. *PLoS One*, 18(8):e0288556.

Qin Yang, Xin Li, Xinyun Ding, Feiyang Xu, and Zhenhua Ling. 2022. Deep learning-based speech analysis for Alzheimer's disease detection: a literature review. *Alzheimer's Research & Therapy*, 14(1):186.

Yangwen Yu, Victor O. K. Li, Jacqueline C. K. Lam, and Kelvin Chan. 2023. GCN-ST-MDIR: Graph convolutional network-based spatial-temporal missing air pollution data pattern identification and recovery. *IEEE Transactions on Big Data*, 9(5):1347–1364.

Yangwen Yu, Victor O. K. Li, Jacqueline C. K. Lam, Kelvin Chan, and Qi Zhang. 2025. CTDI: CNN-Transformer-based spatial-temporal missing air pollution data imputation. *IEEE Transactions on Big Data*, 11(5):2443–2456.

Baseline	**Accuracy (%)**	**F1 (%)**
NC	60.6	67.6
*e*MCI	88.8	0.0
*l*MCI	77.8	22.6
AD	86.1	62.9
TS-Fusion	**Accuracy (%)**	**F1 (%)**
NC	66.7	71.5
*e*MCI	87.8	5.7
*l*MCI	78.8	39.0
AD	90.1	72.0
TS-Fusion + CL	**Accuracy (%)**	**F1 (%)**
NC	70.7	73.0
*e*MCI	87.1	17.2
*l*MCI	79.5	45.7
AD	88.8	70.4

Table 5: Ablation Study with Detailed Breakdown. Baseline: Multilingual BERT Trained on All Original Data. CL: Contrastive Learning.

A Ablation Study with Detailed Breakdown Across LOAD Stages

B Salient Speech Markers

***e*MCI**	**Score**	***l*MCI**	**Score**	**AD**	**Score**
cotidiana (everyday)	0.30	todo (everything)	0.16	cotidiana (everyday)	0.42
brillante (bright)	0.11	cotidiana (everyday)	0.14	todo (everything)	0.27
este (this)	0.08	primer (first)	0.08	primer (first)	0.15
cuadro (picture)	0.08	brillante (bright)	0.05	brillante (bright)	0.15
situación (situation)	0.07	este (this)	0.04	este (this)	0.11
sensación (feeling)	0.06	lugar (place)	0.04	lugar (place)	0.09
alguna (some)	0.06	quehaceres (chores)	0.04	cuadro (picture)	0.08
palabras (words)	0.05	niños (children)	0.03	actividad (activity)	0.07
vida (life)	0.05	cocina (kitchen)	0.03	palabras (words)	0.06
todos (everyone)	0.04	luego (then)	0.03	quehaceres (chores)	0.06

Table 6: The Most Salient Speech Markers Among the Spanish-speaking Cohorts Using SHAP Analysis (English Translations in Parentheses)

C Cross-Linguistic Comparison and Linguistic Implications

Across the three linguistic populations, Chinese, English, and Spanish, consistent patterns emerge that reflect both shared and language-specific trajectories of cognitive-linguistic decline in AD.

C.1 Shared Markers and Patterns

Across all languages, the early stages of cognitive impairment (*e*MCI and *l*MCI) are characterized by the use of relatively rich and specific vocabulary. In contrast, the later stages (LOAD) show a marked shift toward general, familiar, or high-frequency terms. Common markers include general nouns (e.g., *todo* "everything", *outside*, *scene*), concrete objects and locations (e.g., *sink*, *kitchen*, *lugar* "place"), and deictic or referential terms (e.g., *este* "this"). This pattern suggests that as lexical

retrieval becomes more effortful, individuals tend to rely on broader or more accessible vocabulary.

Additionally, fillers and discourse markers (e.g., "ba" in Chinese, "okay" and "ah" in English) become more salient in the later stages, indicating increased disfluency, syntactic disruption, and potential compensatory mechanisms in maintaining conversational flow.

C.2 Language-Specific Features

While the trajectory of decline is broadly shared, specific lexical patterns reflect language-specific linguistic structures:

Chinese: The emergence of the discourse particle *ba* in the *l*MCI stage highlights increased reliance on filler elements to maintain fluency amidst lexical difficulties. In the LOAD stage, verbs such as "trying" appear more frequently, suggesting a shift toward describing cognitive effort or action planning, which may indicate syntactic compensation.

English: Speech remains grounded in perceptually salient and concrete vocabulary (e.g., *outside*, *cupboard*) across stages. In LOAD, the increased use of vague adjectives (e.g., *better*) and discourse fillers (e.g., *ah*, *okay*) reflects semantic generalization and decreased lexical precision.

Spanish: Abstract and emotionally expressive terms (e.g., *vida* "life", *sensación* "feeling") are common in *e*MCI, but gradually give way to routine-related and generic terms (e.g., *quehaceres* "chores", *actividad* "activity", *todo* "everything") in later stages, indicating semantic narrowing and increased dependence on familiar schemas.

C.3 Linguistic Implications

These observations highlight key linguistic consequences of cognitive decline across languages:

Semantic Impairment: A reduction in lexical diversity and specificity, with a shift toward generic, high-frequency vocabulary.

Syntactic Simplification: Increased use of simple sentence structures, fillers, and reduced syntactic variety, reflecting diminished verbal planning and execution.

Discourse Cohesion: Greater reliance on discourse markers (e.g., *ba*, *okay*, *luego*) suggests an attempt to maintain narrative structure despite lexical retrieval challenges.

Pragmatic Compensation: Use of generalization, repetition, and vague terms indicates adaptive strategies to preserve communicative intent.

In summary, while the specific lexical items differ across languages due to structural and cultural differences, the underlying linguistic patterns reveal consistent neurocognitive changes across different stages of LOAD. Cross-linguistic analysis thus provides crucial insight into how AD affects semantic richness, syntactic complexity, and pragmatic function across diverse linguistic contexts.