Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Iris Zheng April 9, 2025

Speech-Based Detection of Cognitive Impairments in Older Adults: Longitudinal Validity Analysis and Cross-Lingual Generalization

by

Iris Zheng

Hyeokhyen Kwon, Ph.D.
Advisor

Joyce Ho, Ph.D.
Co-advisor

Computer Science

Hyeokhyen Kwon, Ph.D.
Advisor

Joyce Ho, Ph.D.
Co-advisor

Wei Jin, Ph.D.
Committee Member

2025

Speech-Based Detection of Cognitive Impairments in Older Adults: Longitudinal Validity Analysis and Cross-Lingual Generalization

By

Iris Zheng

Hyeokhyen Kwon, Ph.D.
Advisor

Joyce Ho, Ph.D.
Co-advisor

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Art with Honors

Computer Science

2025

Abstract

Speech-Based Detection of Cognitive Impairments in Older Adults: Longitudinal Validity Analysis and Cross-Lingual Generalization
By Iris Zheng

Early detection of Alzheimer's Disease (AD)—the most common form of dementia—and its prodromal stage, Mild Cognitive Impairment (MCI), is crucial for enabling timely interventions and effective care planning. However, current diagnostic practices—relying on clinical assessments, neuropsychological testing, and biomarker analysis—are often costly, time-intensive, and inaccessible, especially in underserved or resource-limited settings. These limitations have driven the development of speech-based screening tools, which offer the advantages of being scalable, non-invasive, and accessible. Most importantly, speech is highly sensitive to early neurodegenerative changes, often reflected in fluency and acoustic patterns. However, key challenges remain regarding the generalizability and stability of such models over time and across diverse linguistic populations.

This dissertation addresses these challenges in two parts. The first part focuses on the longitudinal analysis of speech-based models to assess their ability to track cognitive changes over time. As part of this, it explores the use of both hand-crafted and deep learning-derived speech features—encompassing acoustic and linguistic aspects—for pre-screening MCI. It also examines psychological well-being measures, such as loneliness and neuroticism, as complementary indicators of cognitive status. Results show that speech-based models remain stable over time, underscoring their potential for continuous cognitive monitoring. Additionally, speech features offer moderate utility for MCI pre-screening, and well-being measures have limited predictive value.

The second section turns to the challenge of cross-lingual generalizability, a key barrier in global dementia screening efforts. To address this, the study analyzes speech-based AD detection models across English, Greek, and Slovak datasets, evaluating their robustness and adaptability in multilingual settings. Multiple transfer learning techniques are applied to enhance the transferability of models trained in one language to others. Training approaches that leverage multilingual data, along with fine-tuning, yield strong results; however, transferability varies across language pairs, highlighting the complexity of cross-lingual generalization. These findings highlight both the promise and the limitations of current transfer learning approaches, emphasizing the need for more sophisticated techniques that can bridge linguistic boundaries in speech-based cognitive screening.

Together, these investigations advance the development of speech-based tools for early, accessible, and globally applicable cognitive impairment screening, while also identifying key limitations and future directions for improving their robustness and reach.

Speech-Based Detection of Cognitive Impairments in Older Adults: Longitudinal Validity Analysis and Cross-Lingual Generalization

By

Iris Zheng

Hyeokhyen Kwon, Ph.D.
Advisor

Joyce Ho, Ph.D.
Co-advisor

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Art with Honors

Computer Science

2025

Acknowledgments

I would like to thank my advisor, Dr. Hyeokhyen Kwon, for his invaluable mentorship and support throughout this project. His guidance, encouragement through challenges, and emphasis on key research skills greatly contributed to my growth. I'm especially grateful for his detailed feedback, which helped me improve as both a researcher and communicator.

I also thank my committee chair, Dr. Joyce Ho, and committee member, Dr. Wei Jin, for their time and support. Dr. Ho's machine learning course laid the foundation for many of the techniques I used, and Dr. Jin's data mining course introduced me to concepts that inspired new directions in my work. Their teaching and feedback were essential in shaping this thesis.

I'm grateful to the ViTAL Lab members for creating a collaborative and supportive environment. In particular, thank you to Xiaofan, Merna, Dharini, and the Interview Analysis team for your feedback, teamwork, and encouragement throughout the process.

Thanks also to the BMI department for their support, and to the IT team for providing guidance and access to high-performance computing tools—an incredible resource as an undergraduate.

Finally, I'm deeply thankful to my family and friends for their constant encouragement. Special thanks to Helen, Christina, and Samme for being steady sources of strength and perspective, and to the many others who have cheered me on along the way.

Contents

Chapter 1

Introduction

Alzheimer's Disease (AD) is the most common form of dementia, accounting for approximately two-thirds of all dementia cases [1]. It is characterized by gradual degeneration of neurons within the cerebral cortex and hippocampus [2], leading to symptoms such as memory loss, behavioral changes, and cognitive deterioration. Mild Cognitive Impairment (MCI) is a preceding condition of AD, representing a transitional state between normal aging and dementia [3]. Since older adults with MCI can still function independently in daily activities, early detection is crucial to implement timely interventions to delay cognitive decline [4].

The diagnosis and assessment of MCI often require costly or invasive biomarker evaluations. Neuroimaging techniques such as positron emission tomography (PET) and magnetic resonance imaging (MRI), cerebrospinal fluid analysis by lumbar puncture, and comprehensive neurological assessments are frequently used [5]. However, these methods can be inaccessible due to high costs, limited availability, or patient discomfort, as they require extensive evaluations by dementia specialists. In particular, 20 states in the United States (US) are considered 'dementia neurology deserts', significantly limiting access to screening for older adults at risk [6]. These challenges highlight the need for innovative, accessible, and scalable approaches to prescreen and

monitor disease progression effectively.

As a cost-effective pre-screening to diagnostic methods, speech-based digital biomarkers have emerged as a promising tool in dementia research [5] due to its sensitivity to early cognitive changes. MCI-related speech changes include early lexical-semantic deficits, such as reduced conceptual richness and lower idea density [7]. As disease progresses, individuals with AD may produce less meaningful speech, often marked by increased hesitation and semantic paraphasias [8].

Despite their promise, the generalizability of speech-based biomarkers remains an open challenge. Variability in speech patterns across individuals, time points, and languages can significantly limit model performance. Longitudinally, the subtle and evolving nature of cognitive decline complicates reliable prediction over time. Cross-lingually, differences in linguistic structure and cultural expression hinder the transferability of models trained in one language to another. These challenges raise critical questions about the robustness and scalability of speech-based approaches for dementia detection.

This thesis aim to answer two questions: 1) Can speech-based biomarkers track cognitive impairment consistently over time? 2) Can models trained on one language generalize to others for detecting AD from speech?

To answer these questions, this work employs longitudinal analysis to evaluate the stability of speech biomarkers over time and applies domain-adaptation techniques to test cross-lingual generalizability.

Thesis Statement

Speech-based biomarkers have the potential to support longitudinal and cross-lingual detection of cognitive impairment.

This dissertation is organized as follow: chapter 2 synthesizes prior research on speech biomarkers for cognitive impairment and AD, highlighting gaps in longitudinal tracking and multilingual generalization. Building on this foundation, chapter 3

investigates whether speech features can reliably track cognitive decline over time. chapter 4 then tackles the challenge of cross-lingual adaptation, testing whether models trained on one language generalize to others like Greek or Slovak. Finally, chapter 5 unifies these threads, arguing that while speech biomarkers show promise for both longitudinal and multilingual AD screening, their broader adoption requires overcoming dataset biases and refining domain-adaptation techniques.

Chapter 2

Related Work

2.1 Speech Biomarker

Speech is increasingly recognized as a promising biomarker for the early detection of cognitive impairment (CI) [9]. Researchers have explored two primary types of speech-derived features to effectively harness this potential: acoustic and linguistic features. Acoustic features, such as pitch, intensity, and spectral patterns, capture how something is said. In contrast, linguistic features focus on what is said, examining word choice, syntactic structure, and semantic coherence.

Paralinguistic features play a crucial role in this analysis. These non-verbal aspects of speech—including voice quality, hesitations, laughter, and emotional tone—offer valuable insights into speaker states like emotional well-being, cognitive effort, and fatigue [10]. Prosodic elements, particularly pitch, intensity, speech rate, and pauses, are especially informative. Research has consistently shown that individuals with cognitive decline exhibit more frequent disfluencies and hesitations [11]. Additionally, emotional prosody can serve as an early marker of cognitive impairment [12].

Previous studies have demonstrated the potential of these approaches. Khodabakhsh et al. [13] used prosodic features to detect AD from spontaneous speech

using a private Turkish dataset, achieving over 80% accuracy. The development of standardized acoustic feature sets like The Geneva Minimalistic Acoustic Parameter Set (eGeMAPS [14]) has further advanced this field, providing robust tools for quantifying vocal markers across clinical domains. Additionally, Haider et al. [15] applied paralinguistic feature sets on the DementiaBank Pitt Corpus [16] and demonstrated that purely acoustic features, extracted without transcription, can achieve high classification accuracy.

Researchers have employed various machine learning techniques to classify cognitive status based on these features. Support Vector Machines (SVM), Random Forests (RF), and Extreme Gradient Boosting (XGBoost) have been particularly prominent [5]. For instance, Bhat and Kopparapu [17] employed three binary random forest classifiers to distinguish between healthy controls, MCI, and AD populations.

Other studies have leveraged both acoustic and linguistic features. For instance, Roark et al. [18] incorporated pause-related features and measures of linguistic complexity to classify MCI versus healthy controls using a private English dataset based on a narrative recall task, achieving strong performance.

The emergence of deep learning has significantly transformed cognitive impairment detection research, enabling more sophisticated approaches to feature extraction and classification. Notable advances include:

1) Recurrent Autoencoders: Bertini et al. [19] introduced a classifier using a recurrent autoencoder trained on log-mel spectrograms to learn compact audio representations. This approach, enhanced with data augmentation, outperformed traditional and deep learning baselines on the Pitt Corpus.

2) Hybrid Architectures: Liu et al. [20] proposed a hybrid neural network architecture combining CNNs for local context, BiLSTMs for global temporal modeling, and attention pooling for classification. Using bottleneck features from a pre-trained ASR model and masking-based augmentation, their model achieved state-of-the-art

performance on the DementiaBank Pitt corpus without relying on transcriptions.

3) Transformer-Based Models: Recent research has leveraged self-attention mechanisms, with the Audio Spectrogram Transformer (AST) [21] showing promise for general audio classification task, though specific AD-related predictions were not reported.

Beyond architectural innovations, deep learning models have also been employed as feature extractors. Haulcy and Glass [22] utilized acoustic (i-vectors, x-vectors) and linguistic (e.g., word vectors, BERT embeddings) features with both classical classifiers and neural networks (CNNs, LSTMs) on the ADReSS dataset [23]. Evaluated independently, models trained on BERT embeddings achieved the best performance.

Self-supervised learning (SSL) approaches have also emerged as powerful tools. Models like wav2vec 2.0 [24] and WavLM [25] have enabled rich, pre-trained audio representations that could be used as feature extractor for audios. Chen et al. [26] explored SSL models for AD using the ADReSS dataset, demonstrating that fine-tuning wav2vec 2.0 and HuBERT—enhanced with multi-task learning and data augmentation—significantly improves performance, achieving results comparable to state-of-the-art baselines.

To address the interpretability challenges of deep learning, researchers have begun developing more transparent models. Rodriguez-Salas et al. [27] introduced Forest-Net, which maps decision trees into sparse multilayer perceptrons. Building on this work, Perez-Toro et al. [28] adapted the approach for AD prediction, highlighting interpretable rhythm- and duration-based speech markers.

While the advancements above have significantly improved predictive performance, several critical gaps remain before speech-based models can be deployed in real-world clinical or screening settings.

2.2 Lack of Longitudinal Validity Analysis

First, the majority of existing studies focus on single-timepoint classification, often using baseline data alone [29], which leaves open the question of whether speech biomarkers maintain predictive validity as cognitive decline progresses over time.

Previous studies have employed baseline speech features to differentiate individuals who experienced cognitive decline from those who remained stable [30, 31], but these studies do not examine whether the same features remain predictive as cognition changes. Moreover, cross-validation is often performed without controlling for speaker identity, limiting insight into longitudinal, within-subject performance.

More recent research using voice assistant data collected over 18 months demonstrated that incorporating historical speech data improved MCI detection accuracy [32], suggesting that natural temporal variations in speech offer richer contextual cues. This raises the question of the longitudinal validity of speech-based models for continuous monitoring of cognitive decline.

2.3 Lack of Cross-lingual Generalization Analysis

Second, most frameworks are developed and validated in monolingual or language-specific contexts, raising concerns about their cross-lingual generalizability, particularly in global or multilingual populations.

Previous efforts in the field include the DementiaBank ADReSS-M Signal Processing Grand Challenge [5], which focuses on transferring English-trained models to Greek. Jin et al. [33] addressed the challenge of cross-lingual generalization by developing an ensemble-based framework (CONSEN) that leverages disfluency and pause features—shown to be more robust than standard acoustic embeddings like wav2vec—achieving strong cross-lingual performance. Tamm et al. [34] employed acoustic features (eGeMAPS) alongside demographic covariates in a mixed-batch train-

ing approach using both English and Greek samples, achieving effective cross-lingual adaptation with high accuracy. Chen et al. [35] found that hand-crafted paralinguistic features offered better cross-lingual generalization, outperforming linguistic embeddings when English-trained models were applied to Greek speech.

Outside the Grand Challenge, Gosztolya et al. [36] demonstrated that temporal speech features—such as articulation rate and pause patterns—can be reliably extracted using ASR systems trained in different languages, enabling cross-lingual MCI detection without significant loss in performance.

Collectively, these studies suggest that language-independent features and transfer learning techniques provide a promising direction for cross-lingual speech-based detection of cognitive decline.

Cross-lingual transfer learning in NLP has shown strong potential in leveraging data from high-resource languages to enhance performance in low-resource settings. For example, Chen et al. [37] proposed a model that captures both language-invariant and language-specific features through adversarial training and a mixture-of-experts architecture, enabling effective cross-lingual transfer for text classification tasks. Similarly, Zhang et al. [38] applied language-adversarial training combined with bidirectional language modeling to reduce language-specific biases in part-of-speech tagging, achieving improved results across 14 languages. Recognizing the linguistic impairment related to AD, Guo et al. [39] applied cross-lingual NLP techniques—combining BERT-based contrastive learning with data augmentation—to improve Mandarin AD detection using English data.

While prior studies have shown promising results in cross-lingual AD detection, several gaps remain. First, most existing work focuses on a single language pair, limiting insights into how well models generalize across a broader range of linguistic settings. Inclusion of more diverse languages is essential to build equitable and globally applicable tools. Second, although NLP techniques have proven effective in textual

cross-lingual tasks, their utility for improving speech-based AD detection remains underexplored.

To address these limitations, this thesis investigates the generalizability of speech-based models along two critical and underexplored dimensions: (1) their ability to track cognitive status longitudinally, using speech collected over multiple timepoints from same individuals, and (2) their capacity for cross-lingual adaptation, leveraging transfer learning techniques to extend performance across different languages. By focusing on these dimensions, this work aims to move speech biomarkers closer to practical, scalable, and globally accessible tools for early detection and monitoring of cognitive impairment.

Chapter 3

Longitudinal Validity Analysis for Cognitive Impairment with Speech and Psychological Well-being Patterns

3.1 Introduction

Alzheimer's Disease (AD) and its prodromal stage, Mild Cognitive Impairment (MCI), remain pressing public health concerns, affecting millions worldwide [40]. Early detection is essential for enabling timely intervention and improving patient outcomes [41]. Although clinical evaluations, neuropsychological assessments, and biomarker-based diagnostics remain the gold standard, their high cost, invasiveness, and limited accessibility have motivated the development of scalable, non-invasive screening tools.

Speech analysis has emerged as a promising, accessible, and naturalistic tool for assessing cognitive function, as speech production gradually degrades from healthy aging to MCI and AD [42]. This makes it well-suited for detecting early signs of

cognitive decline. However, its sensitivity raises an question of whether speech-based models reliably generalize as cognitive states change over time. The stability of the relationship between speech features and cognitive function remains unclear, yet understanding this temporal dynamic is key to effectively using speech for long-term dementia screening and monitoring.

Beyond speech characteristics, psychological well-being (PWB) may offer complementary information for cognitive impairment detection. Cognitive impairment severity is linked to self-reported quality of life, with cognitive complaints associated with lower well-being, increased depression, anxiety, and perceived stress [43]. Individual PWB profiles show varying prevalence of MCI, with those experiencing greater social disconnectedness exhibiting higher rates of cognitive decline [44], while strong social networks appear protective [45]. Personality traits further influence cognitive trajectories, with higher conscientiousness and extraversion being protective against dementia, while neuroticism increases susceptibility [46]. Depression also increases dementia risk by 1.28 times [47], though this relationship varies with diagnostic criteria and severity [48].

Given that speech can encode various psychological states [49], integrating PWB measures into speech-based frameworks may enhance cognitive impairment detection sensitivity.

To address these gaps, this work extends previous approaches by: 1) assessing the longitudinal validity of speech biomarkers, specifically examining whether models trained on baseline speech characteristics can generalize to the same participants at their 6-month and 12-month follow-ups, 2) measuring the predictive power of PWB scores for cognitive classification, evaluating whether these scores provide complementary information to speech characteristics. By leveraging speech samples from the same individuals across time, we assess whether models trained on baseline data can generalize to future cognitive states. This evaluation provides insight into

the stability of speech-based predictions and supports more personalized, longitudinal monitoring.

3.2 Methods

3.2.1 Dataset

This study draws on data from the Internet-Based Conversational Engagement Clinical Trial (I-CONECT) (NCT02871921) [50], a longitudinal trial investigating the effects of social engagement on cognitive health in socially isolated older adults at risk of cognitive decline. Participants were aged 75 or older and had either normal cognition or Mild Cognitive Impairment (MCI) at baseline. Recruitment took place in Portland, Oregon (primarily Caucasian participants) and Detroit, Michigan (primarily African American participants).

To qualify as socially isolated, participants met at least one of the following criteria:

- Low social network score ($\leq$ 12 on the 6-item Lubben Social Network Scale (LSNS-6) [51]).
- Limited social engagement ($\geq$ 30-minute conversations no more than twice a week).
- Self-reported loneliness ("often" response on the 3-item UCLA Loneliness Scale [52]).

Exclusion criteria included severe depressive symptoms (Geriatric Depression Scale [53] (GDS) score $\geq$ 7) and a clinical diagnosis of dementia, which were made via a consensus process involving neurologists and neuropsychologists, based on the National Alzheimer's Coordinating Center Uniform Data Set Version 3 (NACC UDS-3 [54]).

Participants completed weekly 10-minute phone check-ins with study coordinators for 12 months. These calls, designed to monitor mental and physical health, also served

as the primary source of longitudinal speech data. Clinical and cognitive assessments were administered at three timepoints: baseline, 6 months, and 12 months.

Demographic characteristics, including age, gender, education, and cognitive status at baseline, are shown in Table 3.1.

Table 3.1: Demographic Characteristics Across Cognitive Assessment Groups. F/M = Female/Male, C/A/O = Caucasians/African Americans/Other. Cognitive classification (Cognitively Normal vs. Cognitively Impaired) in this table is based on **baseline** measures only.

Variable	**Overall (n=103)**	**Cognitively Normal (n=48)**	**Cognitively Impaired (n=55)**
Sex (F/M)	78/25	41/7	37/18
Race (C/A/O)	80/22/1	38/10	42/12/1
Age (Mean ± Std)	80.93 ± 4.67	79.5 ± 3.9	82.2 ± 4.9

Speech Data Collection

Speech was collected during weekly calls, each lasting approximately 10 minutes. Coordinators followed a semi-structured protocol covering:

- Hospital visits or medical concerns during the past week.
- Mood and emotional well-being.
- Social engagement activities (e.g., calls, in-person visits, written communication).
- A brief discussion of a weekly fun fact to encourage spontaneous conversation.

These longitudinal recordings allow for tracking intra-individual changes in speech patterns over time.

3.2.2 Outcomes and Clinical Assessment

Cognitive function was evaluated using three clinical instruments: the Neuropsychological Test Battery (Normcog) [54], the Clinical Dementia Rating Scale (CDR) [55],

and the Montreal Cognitive Assessment (MoCA) [56]. These measures were used to define cognitive status and examine the predictive utility of speech-derived features.

	Baseline (n=94)	6 Months (n=82)	12 Months (n=58)
Normcog	45/49	48/34	36/22
CDR	58/36	61/21	43/15
MoCA	54/40	30/52	15/43

Table 3.2: **Distribution of Cognitive Status Across Time Points**. Each cell represents cognitively normal / cognitively impaired.

Normal Cognition (Normcog) vs. Impaired with distribution shown in Figure 3.1 was determined using clinician ratings from NACC UDS V3 Form D1, which evaluates memory, attention, executive function, language, and visuospatial ability [54]. A score of 1 indicates normal cognition; 0 indicates impairment.

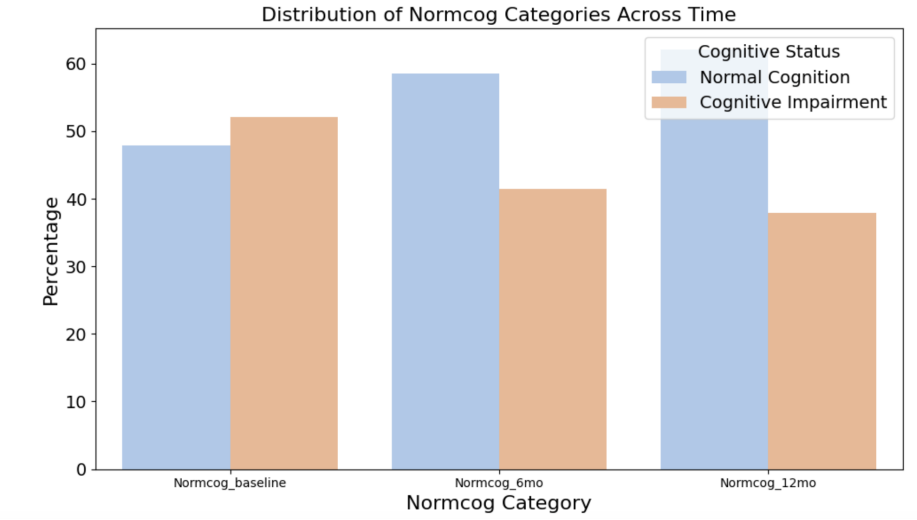

Figure 3.1: Normal Cognition vs. Cognitive Impaired Score Distribution Over Time

The CDR (Figure 3.2) rates six functional domains, including memory, orientation, judgment, community affairs, home and hobbies, and personal care, producing a global score from 0 (no impairment) to 3 (severe impairment). All participants in this dataset had scores of either 0 or 0.5, which were dichotomized into high and low cognitive groups.

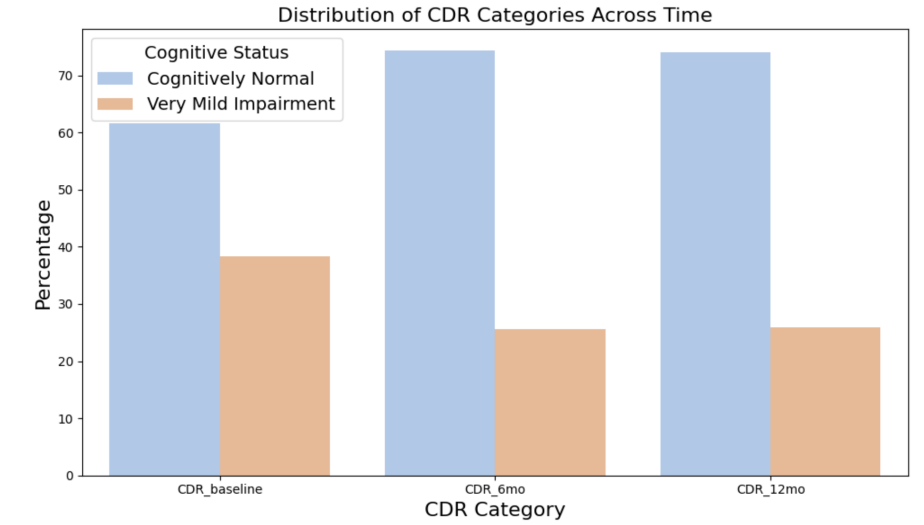

Figure 3.2: CDR Score Distribution Over Time

The MoCA is a sensitive screening tool for MCI in older adults, assessing memory, executive functioning, attention, language, visuospatial skills, and orientation. In this study, the full MoCA was used, excluding visually impaired participants (who were assessed using MoCA-Blind [57] due to different scaling). Participants were grouped into high or low cognitive performance categories based on a median cutoff score of 24. The distribution of MoCA scores across time are shown in Figure 3.3 and the distribution of high MoCA / low MoCA classes shown in Figure 3.4.

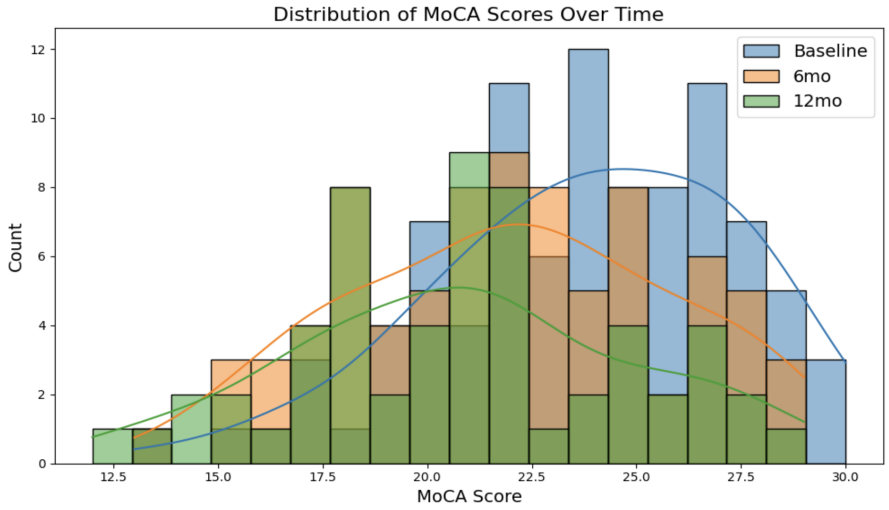

Figure 3.3: MoCA Score Distribution Change Over Time

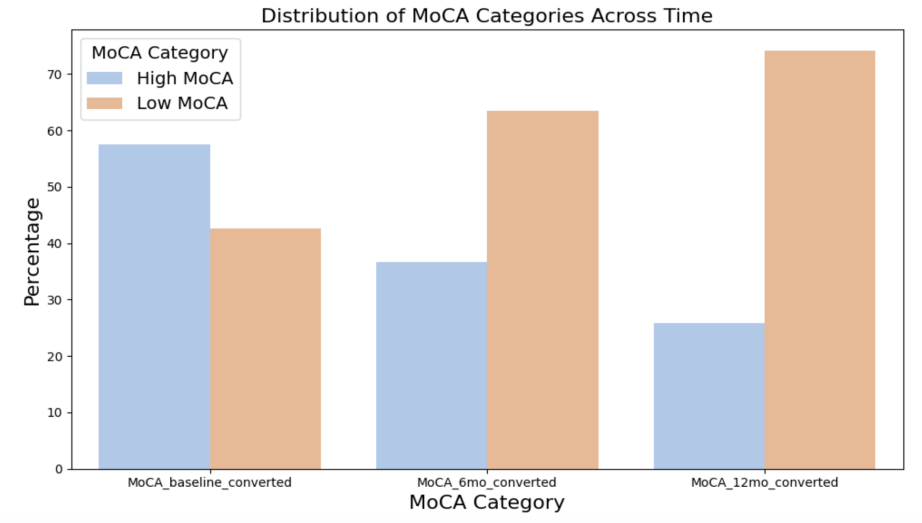

Figure 3.4: MoCA Categories Change Over Time

3.2.3 Audio Processing Pipeline Overview

The proposed pipeline Figure 3.5 quantifies cognitive function by analyzing both acoustic and linguistic speech patterns. Since moderator speech was removed during data collection, the recordings contained extended periods of silence. To address this, the audio was further preprocessed to retain only the participant's voiced segments, followed by feature extraction. The extracted features were then used to train binary classification models.

3.2.4 Preprocessing

To preserve key speech features and minimize the impact of silent segments during mean pooling, a manual voice activity detection (VAD) pipeline was implemented. Speech signals were framed (20 ms, 10 ms shift), and short-term energy (STE) was computed as the squared amplitude sum, then normalized. Frames were classified as voiced if their energy exceeded 5% of the max STE. Consecutive voiced frames were merged, and silent gaps of ≤ 1 second were retained to maintain speech continuity. The energy threshold and merging criteria were optimized after gap-bridging by evaluating

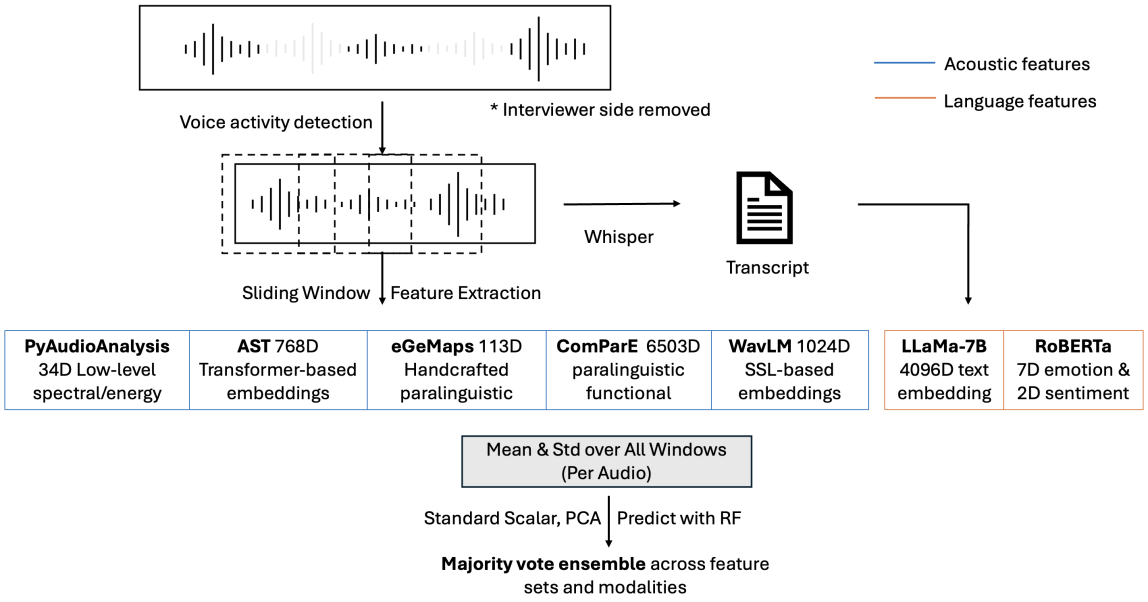

Figure 3.5: Processing Pipeline for Extracting and Analyzing Speech-Based Acoustic and Linguistic Features

10 manually labeled samples, selecting the threshold that maximized the Jaccard Index between the automated pipeline's output and human-annotated speech segments.

Since cognitive and psychological scores were available only at baseline, 6 months, and 12 months, weekly phone call recordings were aggregated to align with these time points. Specifically, the four nearest weekly recordings were concatenated for each assessment: weeks 1–4 for baseline, weeks 22–25 for the 6-month mark, and weeks 45–48 for the 12-month mark.

3.2.5 Feature Extraction

Audio Features

All audio recordings were downsampled to 16 kHz before processing. A diverse set of acoustic features was then extracted, incorporating both deep learning embeddings and handcrafted acoustic descriptors—features that have been widely used to assess mental and cognitive health [5, 34, 58]. By combining deep-learning-based representations

with traditional acoustic and prosodic features, this approach aimed to capture a comprehensive characterization of speech.

Deep Learning Embeddings To leverage pre-trained representations, WavLM-Large embeddings were extracted every 20 ms using the WavLM-Large model [25], which was trained on 94k hours of speech data through masked speech prediction and denoising tasks. Additionally, Audio Spectrogram Transformer (AST) embeddings were extracted from the 9th hidden layer of a pre-trained AST model [21], originally trained on AudioSet [59].

Handcrafted Acoustic Features Complementing deep embeddings, handcrafted features were extracted using OpenSMILE ComParE [60] and eGeMAPS feature sets [14], which include measures of pitch, intensity, voice quality, and spectral descriptors. To capture temporal dynamics, features were computed using a sliding window approach with a frame length of 250 ms and a hop length of 150 ms. Additionally, short-term acoustic features were computed using PyAudioAnalysis [61] with a 100 ms window and 50% overlap, capturing energy, spectral properties, and 13-dimensional Mel-frequency cepstral coefficients (MFCCs).

Language Features

Language features were extracted from transcripts generated using the Whisper-small [62] automatic speech recognition (ASR) model. These transcripts were processed using the LLaMA-7B model [63] to generate text embeddings, with input sequences truncated or segmented to respect the model's maximum context window of 2048 tokens. Additionally, sentiment and emotion scores were computed using deep learning models:

- Emotion classification was processed with DistilRoBERTa-based model [64], which categorized emotions into neutral, happiness, sadness, surprise, fear,

disgust, and anger.

- Sentiment analysis was performed using a RoBERTa-large model [65], which provided scores for positive and negative sentiment.

Psychological Well-being Scores

To assess psychological well-being in relation to cognitive decline or aging, we incorporated the following measures:

- Geriatric Depression Scale (GDS) [53]: Assesses depression symptoms in elderly.
- LSNS-6 [51]: Evaluates social network size and quality.
- Neuroticism from NEO Five-Factor Inventory [66]: Measures emotional stability.
- NIHTB-EB Emotional Well-being Assessment [67]: A composite measure of emotional well-being based on 17 subscales. Following prior methodology [68], three key composite scores were derived—negative affect, social satisfaction, and overall psychological well-being. These composites were computed using standardized factor loadings, averaged across relevant subscales, normalized, and converted to T-scores (T=50, SD=10).

3.2.6 Participant-Level Feature Aggregation

To obtain a fixed-size representation for each speech audio, we applied temporal pooling by computing statistical descriptors—mean and standard deviation—over the extracted feature vectors. While both mean and max pooling were evaluated, mean pooling was ultimately selected due to its superior performance on the validation set across downstream classification tasks.

3.2.7 Experiment Setting

Longitudinal Validity Analysis

To evaluate the predictive power and temporal generalizability of audio-derived features for detecting cognitive decline, models were trained on each participant's baseline data and tested on their corresponding 6- or 12-month follow-ups. Hyperparameters were tuned using data from the remaining participants, following a user-dependent cross-validation strategy. This setup ensured that training and testing occurred within the same individual but across timepoints, allowing us to assess how well speech features captured longitudinal changes in cognitive status. To enhance robustness, the training set order was randomly shuffled 20 times per split.

Psychological Well-being as a Predictor

To investigate the role of PWB in cognitive impairment detection, we employed a distinct evaluation framework. In this case, data from all participants and all timepoints (baseline, 6-month, and 12-month) were pooled together. We evaluated two scenarios: (1) using PWB scores and acoustic features independently, and (2) combining standardized PWB scores with each acoustic or language feature set. Models were trained and evaluated using stratified, user-independent 5-fold nested cross-validation, where the inner loop handled hyperparameter tuning and the outer loop measured generalization to unseen participants. This setup tested whether PWB contributed useful information for distinguishing cognitive status across individuals, independent of specific timepoints.

Multi-Modal Fusion and Classification

To classify cognitive impairment, a late-fusion approach was used, shown to outperform early fusion in prior work [69]. Separate Random Forest classifiers were trained per

modality, and final predictions were obtained via majority voting across modalities.

Evaluation Metrics

Accuracy and Area Under the Curve (AUC) were computed. Final scores were reported as the mean ± standard deviation across CV folds.

3.3 Result

3.3.1 Longitudinal Validity Analysis of Cognitive Impairment Over Time

Feature Set	6mo AUC			12mo AUC		
	CDR	Normcog	MoCA	CDR	Normcog	MoCA
AST	$\mathbf{0.75 \pm 0.03}$	0.74 ± 0.03	0.57 ± 0.04	0.65 ± 0.04	0.58 ± 0.04	0.51 ± 0.04
ComParE	0.74 ± 0.03	0.71 ± 0.03	0.59 ± 0.03	0.63 ± 0.03	0.64 ± 0.04	$\mathbf{0.62 \pm 0.04}$
eGeMAPS	0.73 ± 0.03	0.64 ± 0.04	0.54 ± 0.04	0.71 ± 0.03	0.64 ± 0.03	0.54 ± 0.04
PyAudio	0.64 ± 0.03	0.70 ± 0.03	0.57 ± 0.04	$\mathbf{0.74 \pm 0.03}$	0.66 ± 0.04	0.53 ± 0.04
WavLM	0.74 ± 0.03	$\mathbf{0.76 \pm 0.02}$	0.62 ± 0.04	$\mathbf{0.74 \pm 0.03}$	$\mathbf{0.69 \pm 0.03}$	0.59 ± 0.05
LLaMA	0.57 ± 0.04	0.58 ± 0.04	$\mathbf{0.64 \pm 0.03}$	0.63 ± 0.05	0.58 ± 0.04	0.60 ± 0.05
Sentiment	0.59 ± 0.03	0.50 ± 0.03	0.51 ± 0.03	0.58 ± 0.04	0.48 ± 0.03	0.54 ± 0.03
Majority Vote	0.58 ± 0.03	0.65 ± 0.04	0.56 ± 0.03	0.55 ± 0.05	0.61 ± 0.03	$\mathbf{0.62 \pm 0.04}$

Table 3.3: AUC performance of the RF model for longitudinal prediction of cognitive status at 6- and 12-month follow-ups. Models trained on baseline features; **bold** values denote best performance per column.

Table 3.3 presents the results of the Longitudinal Validity Analysis for models trained on baseline data and evaluated at subsequent time points (6-month and 12-month). CDR classification demonstrated the highest generalizability, with performance remaining stable across time points (Highest AUC = 0.75 for 6-month prediction and 0.74 for 12 month). Normcog also exhibited moderate generalizability, achieving an AUC of 0.76 for 6-month prediction and 0.69 for 12-month prediction. In contrast, MoCA classification showed lower predictive performance, with AUC values of 0.64 for 6-month and 0.62 for 12-month predictions.

3.3.2 User-Independent Audio-Based Classification of Cognitive Impairment

Feature	MoCA		Normcog		CDR	
	Accuracy	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
AST	0.47 ± 0.06	0.48 ± 0.09	0.57 ± 0.04	0.58 ± 0.07	0.69 ± 0.05	0.62 ± 0.08
ComParE	0.50 ± 0.07	0.52 ± 0.10	**0.60 ± 0.06**	**0.65 ± 0.06**	0.69 ± 0.03	0.57 ± 0.09
eGeMAPS	0.54 ± 0.07	0.54 ± 0.07	0.55 ± 0.11	0.62 ± 0.11	0.68 ± 0.07	0.55 ± 0.14
pyAudio	0.55 ± 0.08	0.54 ± 0.06	0.55 ± 0.09	0.54 ± 0.10	0.68 ± 0.06	0.60 ± 0.04
WavLM	0.52 ± 0.08	0.52 ± 0.07	0.60 ± 0.10	0.62 ± 0.14	0.69 ± 0.04	0.57 ± 0.08
LLaMA	**0.57 ± 0.06**	**0.64 ± 0.08**	0.56 ± 0.07	0.59 ± 0.11	0.69 ± 0.05	0.60 ± 0.07
Sentiment	0.51 ± 0.04	0.53 ± 0.05	0.50 ± 0.05	0.48 ± 0.03	0.67 ± 0.03	0.56 ± 0.13
Majority Vote	0.53 ± 0.01	0.6 ± 0.02	0.55 ± 0.03	0.58 ± 0.05	**0.69 ± 0.02**	**0.68 ± 0.02**

Table 3.4: Classification performance using speech-derived acoustic and linguistic features. User-independent cross-validation applied; **bold** indicates best performance for each cognitive assessment.

Table 3.4 summarizes the performance of speech-related features in predicting cognitive assessment categories. Among the feature sets, ComParE achieved the highest AUC for distinguishing Normcog status (0.65), while LLaMA-based embeddings yielded the best AUC for MoCA classification (0.64). For CDR, majority vote result yielded best result, with AUC of 0.68.

3.3.3 User-Independent Classification of Cognitive Impairment via Well-being Scores

Target	**Accuracy**	**F1 Score**	**AUC**
MoCA	0.53 ± 0.06	0.52 ± 0.09	0.56 ± 0.05
Normcog	0.52 ± 0.09	0.51 ± 0.011	0.54 ± 0.08
CDR	0.69 ± 0.11	0.66 ± 0.13	0.62 ± 0.09

Table 3.5: Predictive performance using psychological well-being scores for MoCA, Normcog, and CDR outcomes. User-independent cross-validation; metrics include Accuracy, F1-score, and AUC.

When using psychological well-being alone to predict cognitive impairment (Table 3.5), all scores are approximately random (0.5). CDR demonstrate best predic-

tivabiltiy with AUC of 0.62.

Feature	MoCA		Normcog		CDR	
	Accuracy	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
AST	0.54 ± 0.03	0.43 ± 0.07	0.51 ± 0.06	0.53 ± 0.09	**0.70 ± 0.04**	0.59 ± 0.09
ComParE	0.60 ± 0.07	0.49 ± 0.09	0.55 ± 0.06	0.64 ± 0.05	**0.70 ± 0.04**	0.56 ± 0.10
eGeMAPS	0.52 ± 0.03	0.48 ± 0.05	0.58 ± 0.08	**0.65 ± 0.11**	0.66 ± 0.06	0.52 ± 0.12
PyAudio	0.56 ± 0.08	0.52 ± 0.04	0.58 ± 0.07	0.56 ± 0.10	0.66 ± 0.06	0.60 ± 0.08
WavLM	0.60 ± 0.03	0.61 ± 0.03	**0.60 ± 0.07**	0.59 ± 0.09	0.69 ± 0.04	0.56 ± 0.03
LLaMA	**0.61 ± 0.07**	**0.66 ± 0.06**	0.52 ± 0.03	0.59 ± 0.07	**0.70 ± 0.04**	0.61 ± 0.08
Sentiment	0.54 ± 0.04	0.52 ± 0.07	0.45 ± 0.05	0.43 ± 0.10	0.68 ± 0.06	0.50 ± 0.07
Majority Vote	0.55 ± 0.02	0.57 ± 0.02	0.56 ± 0.02	0.57 ± 0.03	0.69 ± 0.02	**0.67 ± 0.04**

Table 3.6: Performance of models combining speech and psychological features via feature concatenation. User-independent evaluation for predicting cognitive assessments; **bold** values show highest scoring method per assessment.

Table 3.6 presents the performance of models trained on speech-derived features combined with psychological well-being scores. Overall, changes in performance were minimal and remained within the standard deviation range. For MoCA, LLaMA's accuracy increased from 0.57 to 0.61, and its AUC from 0.64 to 0.66, though both remained within the variability of the original results. For Normcog, performance remained largely unchanged. For CDR, accuracy showed a slight improvement (e.g., AST and ComParE increasing from 0.69 to 0.70), while AUC experienced a small decrease (AST: 0.62 to 0.59; ComParE: 0.57 to 0.56).

3.4 Discussion

3.4.1 Longitudinal Validity Analysis of Cognitive Impairment Over Time

Cognitive assessment tools demonstrate markedly different predictive stabilities over time, as evidenced by our analysis (Table 3.3). While CDR and Normcog show moderate predictive consistency with 6-month AUCs of 0.75 and 0.76, respectively, MoCA reveals significant longitudinal variability when models are trained on baseline

data.

This variability stems from multiple interconnected factors. Critically, MoCA's reliance on a rigid cutoff score (≥ 24) amplifies minor score fluctuations into seemingly significant clinical changes.

At baseline, 58.1% of participants were classified as "high MoCA," but this proportion plummeted to 25.8% by 12 months (Figure 3.4)—a dramatic shift attributable to both participant attrition and score volatility. In contrast, Normcog and CDR demonstrate more stable trajectories: Normcog's cognitively normal classifications gradually increased from 48.4% to 62.1%, while CDR showed a similar trend, rising from 61.1% to 74.1% (Figure 3.1, Figure 3.2). The inherent limitations of MoCA's scoring methodology contribute significantly to this longitudinal instability. While its cutoff point is clinically established [70], it remains overly sensitive to minor fluctuations that may not reflect meaningful cognitive change. Prior research demonstrates that MoCA scores tend to decline naturally with age [71], reflecting normal cognitive aging rather than pathological deterioration. Consequently, relying on a binary threshold risks misinterpreting these subtle, expected changes as clinically significant.

In contrast, Normcog and CDR integrate clinician judgment that accounts for broader contextual factors, functional abilities, and compensatory strategies. This holistic approach yields more stable and nuanced cognitive classifications, demonstrating the limitations of mechanistic, threshold-based assessments in capturing the complexity of cognitive function.

The moderate predictive performance (AUCs around 0.75-0.76) underscores the nuanced relationship between speech patterns and cognitive states. While speech provides valuable insights into cognitive function, these metrics suggest that speech alone cannot fully capture the complexity of cognitive changes. This limitation highlights the importance of multi-modal assessment approaches that integrate speech analysis with traditional cognitive screening tools, clinician judgment, and comprehensive

functional evaluations.

3.4.2 Classifying Cognitive Impairment with Audio Features

When predicting CDR, Normcog, and MoCA with speech-based, user-independent CV (Table 3.4), no single feature set demonstrated superior predictive power, and performance remained within a moderate range. Additionally, no single category of speech-derived features (acoustic vs. linguistic) or approach (hand-crafted vs. deep embeddings) uniformly outperformed all others. This outcome contrasts with prior research, where deep neural embeddings frequently outperformed hand-engineered features [72, 73].

A key factor may be the restricted nature and quality of the phone-based audio data. In this study, weekly phone call check-ups were used, where participants engaged in brief conversations but primarily provided yes/no answers, short explanations, or minimal responses (e.g., "3 hours"). Due to the structured nature of these calls, the data offered limited acoustic and linguistic variability, potentially impacting the model's ability to capture nuanced speech patterns. In line with Knopman et al., who noted that while telephone-based assessments (like TICS-m) can distinguish dementia from normal cognition, they do so only moderately well when differentiating MCI [74]. Given the brevity and structure of telephone-based interactions, such speech may lack the depth needed for robust cognitive assessments.

3.4.3 Classifying Cognitive Impairment with Psychological Well-Being Scores

When using psychological well-being scores alone to predict cognitive impairment, performance remained near chance (0.5) (Table 3.5). Moreover, incorporating these scores with speech-derived features did not significantly enhance model performance

(Table 3.6).

One explanation lies in the characteristics of our study sample. One of the exclusion criteria was severe depressive symptoms (GDS-15 ≥ 7), leading to a cohort primarily with no or mild depression. Similarly, 49.0% of participants had LSNS scores between 10 and 15, close to the social isolation cutoff of 12, thereby limiting variability. Such restricted ranges may reduce the predictive power of these psychological measures.

Additionally, the association between cognitive impairment and psychological well-being are shaped by a complex interplay of cognition, mood, memory perception, and quality of life [75]. Relying solely on psychological well-being scores overlooks these multifaceted interactions. As a result, excluding direct measures of functionality can oversimplify the relationship between psychological state and cognitive impairment, thereby restricting predictive performance.

3.4.4 Limitations and Future Directions

Longitudinal Validity Analysis and Label Stability. A key limitation is the fixed cut-off point of 24 used in MoCA, which can amplify minor test fluctuations and resulting in inconsistent classifications over time. Personalized modeling approaches, which have shown promise in improving AD prediction using Electroencephalography (EEG) data [76], merit consideration to account for individual differences in cognitive trajectories. Rather than applying a fixed threshold (e.g., MoCA ≥ 24), models could adapt classification cutoffs based on an individual's baseline cognitive profile and rate of change, thereby improving both sensitivity and specificity.

The direct concatenation of four weeks' data to align with baseline, 6-month, and 12-month assessments presents another limitation. Future research should explore time-aware modeling strategies to better capture the evolving nature of cognitive impairment. Recurrent neural networks such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) are well-suited for tracking temporal dependencies

in speech features [77], enabling models to learn progression patterns that enhance predictive accuracy.

Phone-Based Audio Assessments. A significant limitation stems from the reliance on phone-based calls, which often yield brief, constrained responses and inconsistent data quality. This limited variability in speech can impede both acoustic and linguistic feature extraction. Future studies could expand data collection methods to include more naturalistic speech samples or passive sensing data (e.g., wearable sensors [78]). These approaches may capture richer language use and real-world behavioral patterns over extended periods, providing a more robust foundation for cognitive assessment.

Psychological Well-Being as a Predictor. The exclusion of individuals with more severe depressive symptoms resulted in a restricted range of psychological well-being scores, potentially diminishing their predictive power. Future research should incorporate a larger and more diverse cohort, encompassing a broader spectrum of psychological well-being to more accurately evaluate its relationship with cognitive impairment.

Another constraint lies in the method used to integrate psychological well-being features with speech-derived representations. The straightforward concatenation approach employed may not have effectively merged these modalities, likely due to the high dimensionality of speech features. To enhance integration, future studies should explore more sophisticated fusion techniques, such as attention-based models [79], to better leverage psychological data within multimodal frameworks.

Finally, psychological well-being alone may not comprehensively reflect cognitive and functional changes, as it is influenced by a complex interaction of cognition, mood, memory perception, and overall quality of life [75]. Investigating alternative or complementary constructs—such as perceived cognitive abilities [80] or daily functioning metrics [81]—could provide a more holistic perspective on the relationship between

emotional well-being and cognitive health.

3.5 Conclusion

This study established the feasibility of using speech-based biomarkers to predict cognitive impairment, highlighting both their promise and limitations. Models trained on CDR and Normcog demonstrated moderate longitudinal predictive performance, suggesting stability of speech-based models. However, user-independent prediction accuracy was constrained, likely due to the variability and noise inherent in phone-based audio data. Furthermore, psychological well-being measures yielded near-random predictive power, and combining them with speech features provided only marginal improvements.

These findings underscore the need for more robust and generalizable approaches. Building on this foundation, the next chapter explores transfer learning and domain adaptation for cross-lingual cognitive impairment detection. Unlike the current study's focus on English dataset, cross-lingual modeling introduces additional challenges due to linguistic variability and mismatched data distributions. To address this, chapter 4 investigates methods to enhance the transferability and resilience of speech-derived biomarkers across diverse language contexts.

Chapter 4

Feasibility of Cross-Lingual Audio-Based AD Classification with Domain Adaptation

4.1 Introduction

The rising global prevalence of AD [82] has coincided with increasing worldwide mobility [83], amplifying the linguistic diversity encountered in clinical settings. In many regions, clinicians are routinely faced with assessing patients whose native languages differ from the locally dominant tongue. These shifts in population demographics underscore the pressing need for reliable prescreening biomarkers that can perform robustly across language and cultural boundaries.

Speech-based biomarkers represent a promising tool as non-invasive, cost-effective screening for cognitive assessment. However, most existing speech-based AD detection methods are grounded in monolingual datasets—particularly American English [9]. This narrow focus raises valid concerns regarding ecological validity and cross-cultural generalizability. Cultural and linguistic differences can profoundly influence both

acoustic (e.g., vowel quality, tone, rhythm) and linguistic (e.g., grammatical complexity, lexical choice) features, making it uncertain whether patterns identified in a single language will readily transfer to another [30].

Cross-lingual AD detection presents unique challenges that go beyond traditional domain adaptation, due to a dual domain shift [84]: covariate shifts stemming from language-dependent acoustic features, and concept shifts caused by differences in how AD manifests across languages. These shifts complicate the direct transfer of models trained on one language to another, particularly in speech-based systems.

While several recent studies have explored cross-lingual adaptation—primarily between English and Greek—most frameworks remain limited to monolingual settings or a single language pair, raising concerns about their broader generalizability [5, 33, 34, 35]. Moreover, though NLP-based transfer techniques have shown promise in text-based AD detection [37, 38], their utility for speech remains underexplored.

This work investigates the feasibility of building audio-based AD classification models that are robust across languages. By evaluating various fine-tuning strategies, we assess the transferability of acoustic AD markers across linguistic contexts—toward scalable, cost-effective, and globally inclusive speech-based screening tools.

4.2 Methods

4.2.1 Datasets

Three speech datasets were used in this study. ADReSS-20 [30] consists of American English speech samples from the Cookie Theft picture description task [16, 85]. ADReSS-M [23] provides Greek speech samples elicited by a different picture description task depicting lions in a natural setting. EWA-DB [86, 87] is a Slovak corpus encompassing neurodegenerative conditions such as Parkinson's disease, AD, and MCI. Each Slovak participant performed four tasks—picture description, word

pronunciation, phonation, and pataka syllable repetition—with manual transcription.

To ensure comparability with the other datasets, only the Slovak picture description task was used. This task features multiple scenes (e.g., family gatherings, everyday household activities); audios corresponding to three colored-image tasks were chosen and concatenated into a single audio sample for each subject. To achieve balanced groups in the Slovak dataset, healthy control participants were matched to participants with AD based on age, years of education, and sex. Where exact matches were not possible, controls were selected from a ±2-year range. Detailed demographic information and class distributions for each dataset are presented in Table 4.1.

Dataset	**Language**	**Split** (HC/MCI/AD)	**Sex** (M/F)	**Age**	**Pre-processing**
ADReSS-20	English	78 / 0 / 78 (n = 156)	70 / 86	66.4 ± 6.7	Noise-reduced and volume-normalized
ADReSS-M	Greek	32 / 0 / 30 (n = 62)	17 / 45	69.8 ± 7.5	Unspecified
EWA-DB	Slovak	42 / 3 / 41 (n = 86)	29 / 57	77.9 ± 8.4	Multi-device recording & manual quality assessment.

Table 4.1: Overview of speech datasets used in this study, including language coverage, subject distribution, and demographic characteristics.

4.2.2 Feature Extraction and Evaluation

To assess the feasibility of cross-lingual AD detection, several speech-based features were extracted, similarly from chapter 3, including AST, eGeMAPS, ComParE, WavLM, and PyAudioAnalysis acoustic features, as described previously. Linguistic features including LLaMA and sentiment was also utilized. Similar to chapter 3, statistical features (mean and standard deviation) were computed across the entire audio sequence for each subject's audio.

Initial Evaluation of Feature Sets. Each feature set was evaluated in a standalone manner to gauge its predictive power for AD classification for different languages. Specifically, a Random Forest (RF) classifier was trained for each feature set to provide a baseline performance assessment. ComParE was selected as the primary feature representation for subsequent neural network experiments, based on its consistent performance in baseline analyses and prior research suggesting that paralinguistic features offer superior cross-lingual transferability compared to linguistic and deep acoustic embeddings [33, 34, 35].

4.2.3 Neural Network Model

Given the small dataset sizes, we used a lightweight neural network implemented using PyTorch Lightning [88] for binary AD vs. HC classification, serving as a controlled baseline for testing cross-lingual adaptation strategies.

The architecture comprises:

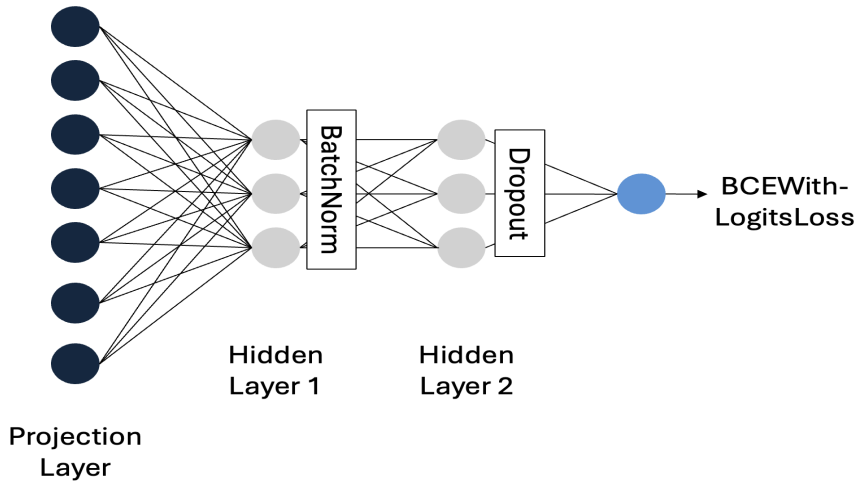

Figure 4.1: Neural Network Architecture for Cross-Lingual Transfer Learning Using Speech-Based Acoustic Features

1. **Projection Layer:** Reduces the high-dimensional ComParE feature (6505D) to a more manageable size.

2. **Fully Connected Layers:** Two hidden layers process the projected features, with batch normalization following the first linear transformation and dropout applied before the final output layer.

3. **Output Layer:** A single neuron with a sigmoid activation for binary classification (AD *vs.* healthy control). Binary Cross-Entropy with Logits (BCEWithLogitsLoss) is employed as the loss function.

4.2.4 Cross-Lingual Adaptation Strategies

Three domain adaptation strategies were investigated to evaluate the model's cross-lingual generalizability. A summary of these approaches is shown in Figure 4.2.

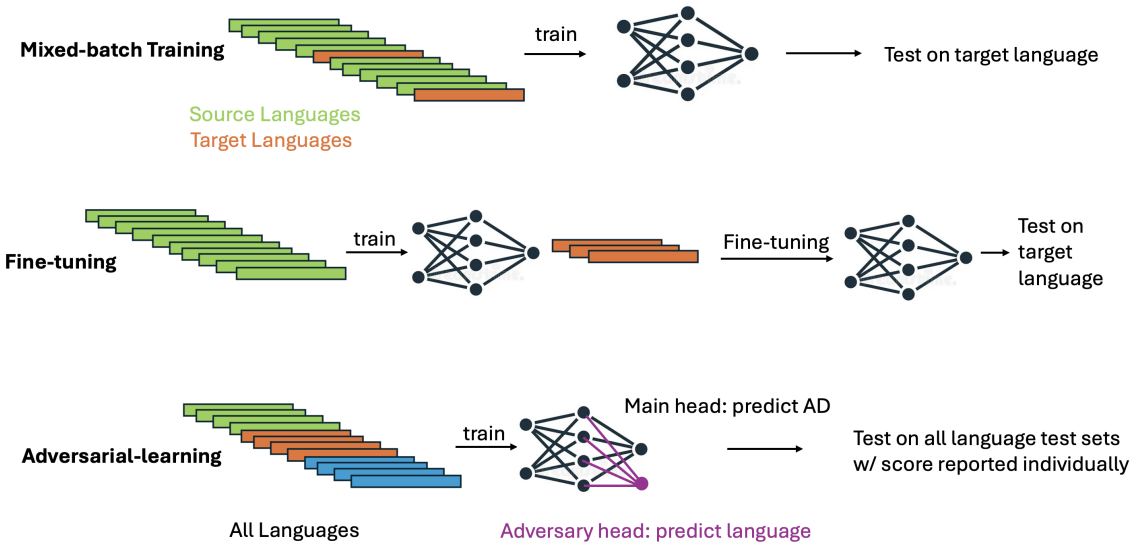

Figure 4.2: Cross-lingual Transfer Learning Techniques

1) Mixed-Batch Training. A mixed-batch strategy that combines source and target language samples within each mini-batch was adopted, following the approach proposed by Tamm et al. [34]. Specifically, the mini-batch ratio is maintained at 5:1 (five source-language samples for every one target-language sample), stratified by

class to preserve the AD-to-healthy ratio. This technique exposes the network to both languages during training while preserving the source-language data's dominance.

2) Fine-Tuning. The second approach involves training the network on a single source language and then fine-tuning it on a target language. To assess how much target data is needed for effective adaptation, we used different subsets of target-language training samples (in increments of three, up to half the dataset). Fine-tuning was performed using a learning rate set to 0.1× the initial training rate. In addition to tracking performance on the target language, we also monitored performance on the source language test set to observe any performance drop—i.e., potential forgetting of previously learned knowledge.

3) Adversarial Learning. The third approach employs an adversarial domain adaptation scheme inspired by [89]. The model is jointly trained on data from all three languages, with two simultaneous objectives: (1) a primary task of predicting AD, and (2) an adversarial task that predicts the language. The core AD-detection gradients and the adversarial gradients are aligned using a projection step:

$$\nabla_{W_P} L_P \; - \; \text{proj}_{\nabla_{W_A}} \nabla_{W_P} L_P \; - \; \alpha \nabla_{W_{LA}} L_{LA}, \tag{4.1}$$

where $\nabla_{W_P} L_P$ is the gradient of the primary AD classification task with respect to the model's parameters, ∇_{W_A} is the gradient for the adversarial task with respect to the adversarial parameters, L_{LA} is the language-adversarial loss, and α is a scaling factor. This technique aims to learn language-invariant features by reducing overlap between the primary-task and adversarial-task gradients.

As a baseline comparison for adversarial learning, samples from all languages were combined to train a single model, and performance scores were reported separately for each language.

4.2.5 Training and Evaluation Setup

Data Splits and Evaluation Protocol Each dataset was stratified into five outer folds, with each fold comprising a training set (80%) and a test set (20%). For each outer fold, the training portion was further split into a single inner validation fold to support hyperparameter tuning. This process resulted in five independently optimized models per language. To evaluate cross-lingual generalization, each model was additionally tested on the held-out test sets of the other languages. Model performance was assessed using Accuracy and Area Under the ROC Curve (AUC).

Input Normalization and Dimensionality Reduction To ensure consistent feature scaling, a `StandardScaler` was fitted on the training data of each dataset and applied to its corresponding test data. For visualization and interpretability, principal component analysis (PCA) was applied after standardization, retaining 95% of the variance. Subsequently, a two-dimensional t-SNE projection (`n_components` $= 2$) was used to explore how samples cluster by class and language in the reduced feature space.

Model Implementation and Hyperparameter Tuning Hyperparameters—including down-projection dimension, hidden layer size, learning rate, dropout rate, and the adversarial regularization coefficient—were tuned using five-fold cross-validation on the training data. Validation AUC served as the selection criterion. To mitigate overfitting, early stopping based on validation loss was applied, with a patience of 5 epochs.

Feature	English		Greek		Slovak	
	Accuracy	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
AST	0.54 ± 0.04	0.56 ± 0.07	0.50 ± 0.15	0.56 ± 0.16	0.69 ± 0.08	0.75 ± 0.11
ComParE	0.65 ± 0.16	0.69 ± 0.16	0.68 ± 0.07	0.77 ± 0.04	0.79 ± 0.08	0.85 ± 0.08
eGeMAPS	0.57 ± 0.10	0.60 ± 0.11	0.65 ± 0.11	0.65 ± 0.06	0.77 ± 0.08	0.83 ± 0.11
PyAudio	0.57 ± 0.09	0.62 ± 0.07	**0.76 ± 0.09**	**0.79 ± 0.09**	0.78 ± 0.11	0.87 ± 0.07
WavLM	0.67 ± 0.12	0.69 ± 0.13	0.53 ± 0.18	0.58 ± 0.18	**0.85 ± 0.09**	**0.92 ± 0.04**
LLaMA	**0.69 ± 0.07**	**0.77 ± 0.03**	0.52 ± 0.09	0.55 ± 0.16	0.66 ± 0.16	0.77 ± 0.13
Sentiment	0.51 ± 0.06	0.54 ± 0.07	0.66 ± 0.04	0.68 ± 0.08	0.66 ± 0.18	0.70 ± 0.20

Table 4.2: Monolingual Classification Results Using RF Across Acoustic and Linguistic Features

4.3 Results

4.3.1 Monolingual Performance with Feature Set Comparison

Table 4.2 presents binary classification results for all feature sets, trained and evaluated within the same language. The results highlight that the best-performing feature varies across languages. For English, LLaMA achieved the highest performance with an accuracy of 0.69 and an AUC of 0.77. In Greek, acoustic features performed best, reaching an accuracy of 0.76 and an AUC of 0.79. For Slovak, WavLM outperformed all other features, achieving 0.85 accuracy and 0.92 AUC. While ComParE did not yield the highest performance, it consistently demonstrated strong results across languages, ranking as the second-best performing feature overall.

4.3.2 Within- and Zero-Shot Cross-lingual Inference

Figure 4.3 presented the t-SNE result of ComParE feature distribution across datasets. From the figure, there are three distinct clusters corresponding to each language. In addition, within each language, the separation between AD (triangles) and HC (circles) is not particularly strong, with noticeable overlap between the two classes.

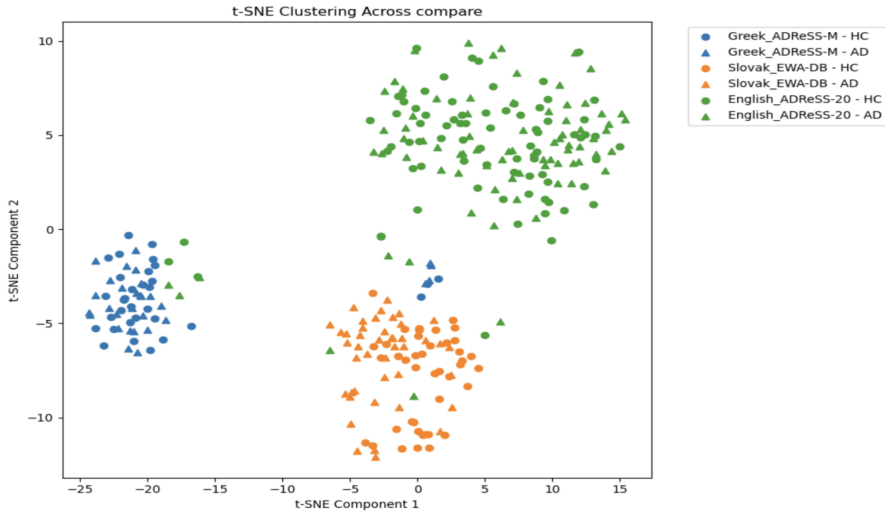

Figure 4.3: ComParE Feature Distribution Across Datasets

Train \ Test	**English**	**Greek**	**Slovak**
English (Accuracy)	**0.61 ± 0.08**	0.54 ± 0.07	0.62 ± 0.06
English (AUC)	**0.60 ± 0.11**	0.57 ± 0.05	0.7 ± 0.07
Greek (Accuracy)	0.54 ± 0.03	**0.61 ± 0.06**	0.58 ± 0.09
Greek (AUC)	0.57 ± 0.03	**0.63 ± 0.08**	0.63 ± 0.14
Slovak (Accuracy)	0.58 ± 0.01	0.55 ± 0.03	**0.74 ± 0.11**
Slovak (AUC)	0.60 ± 0.01	0.61 ± 0.04	**0.84 ± 0.06**

Table 4.3: Binary classification results with NN models trained and tested on self or a different target language. **Bold** text represent training and inference on self.

Table 4.3 presents results of train and inference using a neural network on ComParE feature sets, selected for its consistency across language. The model performs best when trained and tested on Slovak data (0.74 accuracy, 0.84 AUC), while English and Greek show moderate performance when trained and tested in-domain (0.6 and 0.63 AUC respectively). In cross-lingual inferences, the performances generally drop. However, highest cross-lingual performance is observed when applying the English-trained model to Slovak (0.7 AUC).

4.3.3 Mixed-Batch Training

Train \ Test	English	Greek	Slovak
English (Accuracy)	0.60 ± 0.08	**0.67 ± 0.09**	**0.78 ± 0.01**
English (AUC)	0.64 ± 0.10	**0.77 ± 0.06**	**0.83 ± 0.02**
Greek (Accuracy)	0.52 ± 0.03	0.63 ± 0.12	**0.60 ± 0.10**
Greek (AUC)	0.56 ± 0.02	0.69 ± 0.15	**0.79 ± 0.06**
Slovak (Accuracy)	0.58 ± 0.03	**0.62 ± 0.02**	0.72 ± 0.11
Slovak (AUC)	0.60 ± 0.02	**0.68 ± 0.02**	0.84 ± 0.06

Table 4.4: Binary classification results with mixed-batch training. **Bold** values indicate performance improvements over within-language testing

Table 4.4 presents the results of mixed-batch training for cross-lingual generalization. The highest performance is observed when training on English and evaluating on Slovak, achieving an AUC of 0.83. Similarly, using English as the source language with mixed Greek samples improves performance, yielding a higher AUC (0.77). When Greek was used as the source language with a subset of Slovak training samples, the model also demonstrated moderate performance (AUC 0.79). In contrast, models trained on Greek and Slovak with mixed English samples exhibited lower performance when evaluated on English (AUC of 0.56 and 0.6 respectively).

4.3.4 Fine-Tuning Results

Figure 4.4, Figure 4.5, and Figure 4.6 show how AUCs change as the number of fine-tuning samples in the target language increase, while also tracking performance (dashed lines) on the original language.

Slovak Fine-Tuning (target = Slovak) Fine-tuning with English-to-Slovak and Greek-to-Slovak data initially caused a drop in performance. However, after incorporating 6,700 seconds of Slovak audio, AUC scores stabilized at 0.73 for the English-trained

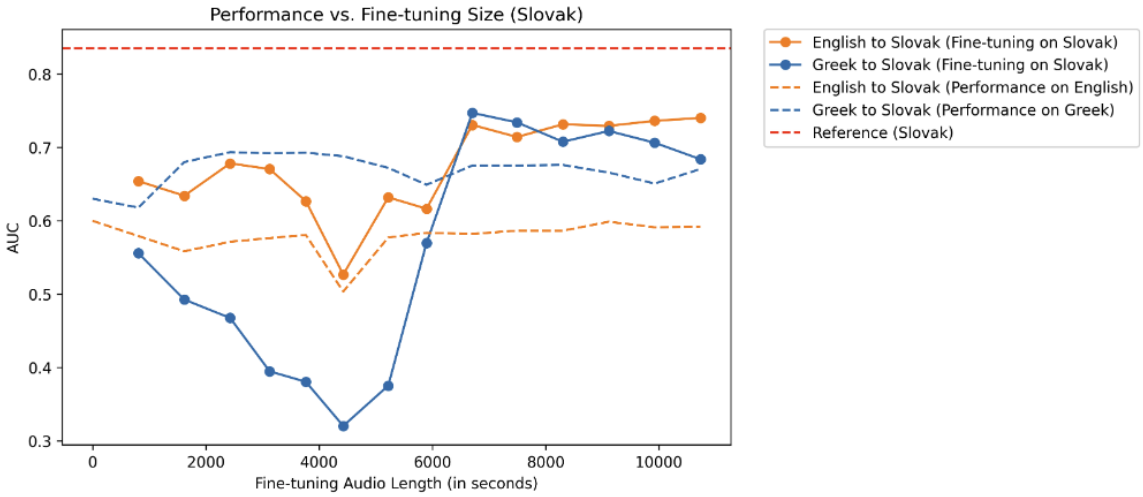

Figure 4.4: Fine-tuning AUC over Fine-tuning Sample Size for Slovak Dataset

model and 0.75 for the Greek-trained model. Performance on the source languages (English and Greek) remained relatively close to baseline levels as more Slovak fine-tuning data was added. The Greek-trained model showed an improvement, with AUC increasing from 0.63 to 0.68, while the English-trained model fluctuated around 0.58.

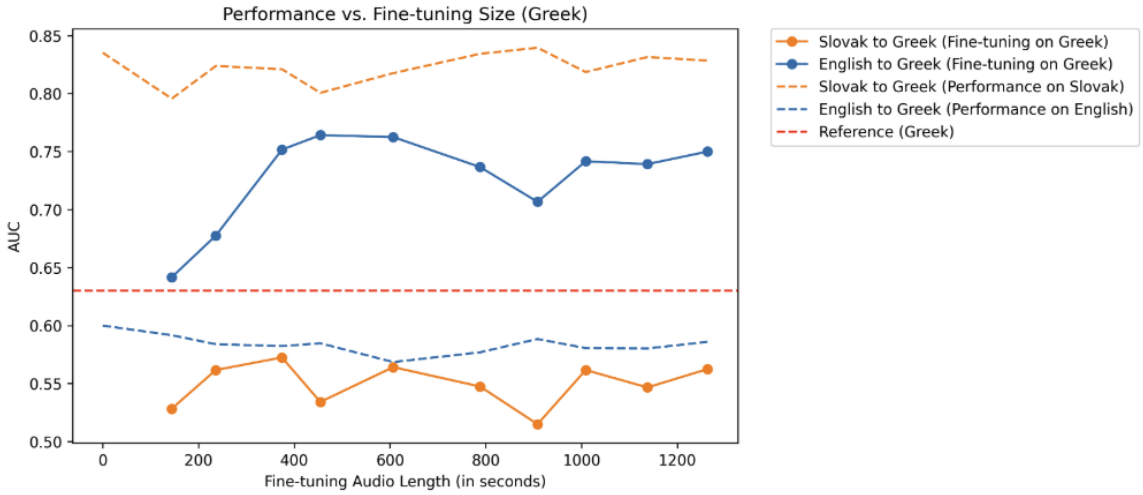

Figure 4.5: Fine-tuning AUC over Fine-tuning Sample Size for Greek Dataset

Greek Fine-Tuning (target = Greek) Fine-tuning with English-to-Greek data began at 0.64 AUC and stabilized at 0.75 AUC after approximately 370 seconds

of target-language fine-tuning. In contrast, Slovak-to-Greek fine-tuning exhibited fluctuations around the baseline AUC of 0.54, with no consistent improvement. Source language performance remained stable throughout, with the Slovak model maintaining 0.84 AUC and the English-trained model fluctuating near 0.59 AUC.

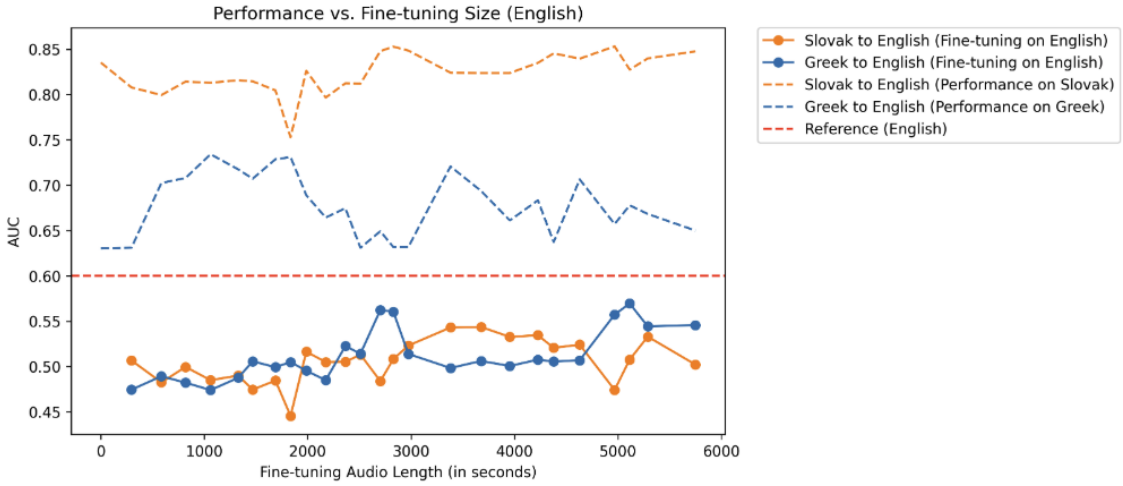

Figure 4.6: Fine-tuning AUC over Fine-tuning Sample Size for English Dataset

English Fine-Tuning (target = English) Fine-tuning from Greek to English yielded gradual performance gains, with AUC improving from 0.47 to 0.56 over 5,000 seconds of English fine-tuning, though remaining below the English-trained baseline. In contrast, Slovak-to-English fine-tuning showed unstable behavior, with AUC fluctuating between 0.45 and 0.54 irrespective of fine-tuning duration. Source language performance varied: the Greek-to-English model maintained Greek AUC between 0.63 and 0.73, while Slovak-to-English performance on Slovak remained stable between 0.75 and 0.85 AUC, comparable to the original baseline.

Figure 4.7 presents the result of comparison between zero-shot inference, mixed-batch training, and fine-tuning across six transfer directions. Overall, mixed-batch training outperformed both zero-shot and fine-tuning across all language pairs.

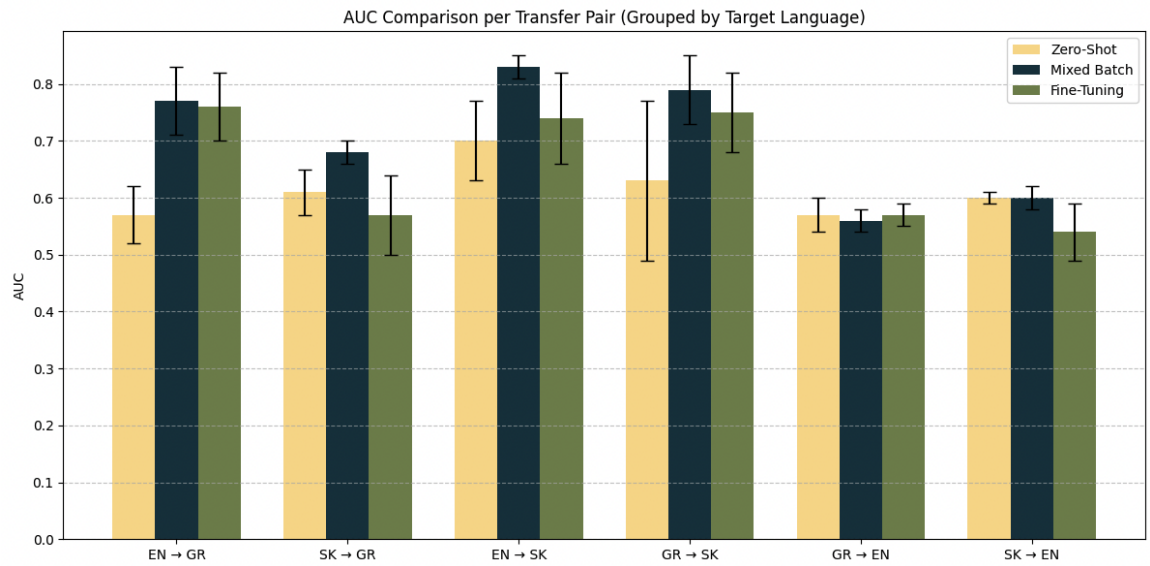

Figure 4.7: AUC Performance Comparison of Transfer Learning Techniques Across Language Pairs

4.3.5 Adversarial Learning

Training Method	Metric	English	Greek	Slovak
Train w/o Adversary	Accuracy	0.61 ± 0.13	0.71 ± 0.14	0.79 ± 0.05
	AUC	0.62 ± 0.13	0.71 ± 0.15	0.79 ± 0.05
Adversarial	Accuracy	0.61 ± 0.09	0.63 ± 0.16	0.76 ± 0.09
	AUC	0.61 ± 0.09	0.63 ± 0.16	0.75 ± 0.09

Table 4.5: Comparison of Adversarial and Non-Adversarial Training Results

Table 4.5 presents the model's performance on English, Greek, and Slovak test sets under two training conditions – with and without adversarial training. The results indicate that adversarial training resulted in a decline in both accuracy and AUC across all languages, with AUC decreasing from 0.62 to 0.61 for English, 0.71 to 0.63 for Greek, and 0.79 to 0.75 for Slovak.

Feature	**Slovak** [90]	**English** [91]	**Greek** [92]
Total Phonemes	42	40–44	31
Vowel Sounds	14	20	5
Consonant Sounds	27	24	31
Syllable Complexity	Allows complex clusters (up to CCCCV onset and CCC coda)	Moderate (allows some complex onsets and codas)	Simple (CV, CVC) but allows complex onsets
Rhythm Type	Mixed	Stress-timed	Syllable-timed
Stress Pattern	Fixed on the first syllable	Variable	Predictable (within last three syllables)
Intonation	Predictable and structured	Highly dynamic and variable	Moderately complex and varied
Vowel Reduction	No	Yes	No
Tonal Language?	No	No	No

Table 4.6: Comparison of Slovak, English, and Greek in Phonetic and Prosodic Features

4.4 Discussion

Language-specific phonetic and prosodic characteristics can influence the effectiveness of speech-based AD detection. Differences in rhythm type, syllable complexity, vowel reduction, and stress patterns may affect the way linguistic and acoustic biomarkers manifest across languages [93]. To better contextualize the observed variations in feature performance, Table 4.6 summarizes key phonetic and prosodic distinctions among Slovak, English, and Greek. These differences provide insight into why certain feature sets may perform better in some languages than others.

4.4.1 Monolingual Performance with Feature Set Comparison

Table 4.2 shows that the best-performing feature sets for AD classification varied across languages. For English, LLaMa led performance, with ComParE and WavLM also performing well (0.69 AUC). In Greek, PyAudioAnalysis and ComParE achieved the highest AUCs (0.79 and 0.77). For Slovak, Acoustic, ComParE, and WavLM features yielded the best results (0.87, 0.85, and 0.92 AUC, respectively).

LLaMa performed the best on the English dataset, which is explained by previous research indicating vocabulary richness, syntactic complexity, and semantic coherence are critical in detecting cognitive impairments [94]. LLaMa's capacity to process these intricate patterns likely contributes to its superior performance in this context. However, since the majority of LLaMa's training data is in English [63], its ability to generate meaningful representation in Greek and Slovak could be limited.

In addition, ComParE performed consistently well across languages, which can be attributed to its comprehensive range of acoustic-prosodic and paralinguistic parameters[60], which was shown to be effective in detecting cognitive impairments across different languages [5].

4.4.2 Within- and Zero-Shot Cross-lingual Inference

Table 4.3 presents within- and zero-shot cross-lingual inference results using the ComParE feature set. The model achieved its highest performance on Slovak data (AUC = 0.84), substantially outperforming English (0.60) and Greek (0.63) under the same architecture. These differences likely reflect the non-standardized nature of data collection and preprocessing across datasets (Table 4.1). Additionally, it is possible that language-specific manifestations of AD—such as more pronounced acoustic changes in Slovak—contributed to the improved classification performance, though direct empirical evidence is limited.

Cross-lingual inference generally resulted in performance degradation. For example,

transferring from English to Greek led to an AUC drop from 0.60 to 0.57, while transferring from Greek to English reduced AUC from 0.63 to 0.57. However, an exception was observed when transferring from English to Slovak, where the model retained relatively high performance (AUC = 0.70). This result may stem from the larger and more diverse English dataset (n=156), which offers richer, more generalizable representations of AD-related speech patterns, aiding cross-lingual generalization through shared acoustic and prosodic features.

4.4.3 Mixed-batch Training

Comparing results from zero-shot inference (Table 4.3), mixed-batch results (Table 4.4) showed a few patterns: firstly, mixed-batch training improved performance when the target language was Slovak or Greek but not English. Specifically, AUC increased for English to Greek (0.57 $\rightarrow$ 0.77), English to Slovak (0.70 $\rightarrow$ 0.83), Greek to Slovak (0.63 $\rightarrow$ 0.79), and Slovak to Greek (0.61 $\rightarrow$ 0.68). On the other hand, mixing English training samples to models primarily trained on Slovak or Greek did not notably enhance performance on English. The AUC remained nearly unchanged for Greek to English (0.57 $\rightarrow$ 0.56) and Slovak to English (0.60 $\rightarrow$ 0.60) under mixed-batch training.

The observed asymmetric transferability of linguistic features underscores the complexity of selecting source and target languages for transfer learning. This phenomenon can be attributed to two primary factors: 1) Dataset size and feature distribution and 2) Linguistic distance and transferability.

Specifically, the English dataset comprises a substantially larger number of samples (n=156), providing a more diverse and extensive speech corpus with a heterogeneous feature distribution, as illustrated in the t-SNE plot (Figure 4.3). In contrast, the Greek (n=62) and Slovak (n=86) datasets contain fewer examples, resulting in a sparser and less discriminative feature space. When training with same number of

mixed-batch samples, this data imbalance may have constrained the model's ability to effectively represent English dataset.

Additionally, prior research in NLP has established that cross-lingual transfer success is influenced by multiple factors, including phylogenetic similarity, typological properties, lexical overlap, and data availability [95]. The observed asymmetry in transferability may be explained by the linguistic divergence between Slovak and English. As members of distinct language families—Slovak being a Slavic language with rich inflectional morphology and flexible word order, and English being a Germanic language with relatively simpler morphology and fixed word order—the structural dissimilarities between the two may hinder effective transfer. However, a more in-depth feature importance analysis is necessary to provide concrete evidence supporting this hypothesis.

Additionally, when testing on the source language, mixed-batch training yielded comparable performance or even better performance to when no target language insertion. Models trained on English, with mixed Greek and Slovak sample, increased performance from 0.6 to 0.64 AUC. For Greek, the performance enhanced from 0.63 to 0.69, while for Slovak, the performance remained unchanged (0.84 AUC). This suggests that exposure to diverse acoustic patterns might help the model learn more generalizable cognitive-related speech patterns and reduce overfitting.

These results align with findings from Lim et al., who show that multi-source language training (MSLT)—akin to our mixed-batch setup—enhances the learning of language-agnostic features by exposing the model to diverse linguistic inputs during training [96].

4.4.4 Fine-tuning

Fine-tuning results (Figure 4.6, Figure 4.5, Figure 4.4) demonstrate that target-language fine-tuning generally leads to performance gains, though the magnitude

of improvement varies across source-target language pairs. Fine-tuning for Slovak inference yielded the most stable and consistent gains, (AUC between 0.73 and 0.75). Greek also showed marked improvement when fine-tuned from an English-trained model (AUC of 0.73). In contrast, fine-tuning toward English resulted in minimal performance gains, with AUC values remaining relatively close to baseline.

Source language performance exhibited initial fluctuations during fine-tuning but tended to recover or stabilize near baseline levels as more target-language data was introduced. This pattern suggests that the model retains core discriminative features with sufficient fine-tuning.

An initial drop in target-language AUC, consistent with known fine-tuning instability [97], was observed but recovered as training progressed, reflecting early-stage adaptation to small or unrepresentative target-language data rather than true model instability. However, the continued improvement in target-language AUC, suggests that the model may be leveraging AD-relevant acoustic or paralinguistic features that are transferable across languages, provided the target dataset is sufficiently rich and representative.

This observation is consistent with prior work demonstrating cross-lingual transferability in AD detection—specifically, the effective use of English-trained models for Greek data [35]. Nevertheless, further investigation into feature attribution and cross-lingual representation learning is required to validate this hypothesis and elucidate the mechanisms underpinning transfer performance.

Comparing mixed-batch training to fine-tuning (Figure 4.7), the improved performance of mixed-batch training may be attributed to its enhanced generalization, potentially resulting from simultaneous exposure to both source and target language distributions during training. While preliminary results are promising, there appears to be limited empirical work directly comparing these two approaches in this context, underscoring the need for more systematic investigation.

4.4.5 Adversarial Learning

Table 4.5 shows that adversarial training did not improve performance and, in some cases, even led to a decline. This suggests that adversarial learning, intended to reduce language-specific biases, may have inadvertently removed essential features needed for AD classification while overemphasizing language-agnostic features.

One possible reason for this decline lies in the nature of cross-lingual AD detection. AD-related speech changes manifest in both linguistic (e.g., lexical access) and paralinguistic (e.g., cognitive processing speed) domains, yet adversarial training aims to neutralize language differences [98]. This process may suppress language-specific cues that are also critical for detecting AD. Research has shown that while grammatical structures often remain intact in AD speech, deficits in fluency, word retrieval, and informativeness are prominent [99]. By minimizing language distinctions, adversarial learning may have inadvertently weakened the model's ability to capture these impairments.

Additionally, adversarial networks are prone to mode collapse [100], where the model overfits to a limited subset of representations. In seeking language-invariant features, the adversarial objective may inadvertently suppress subtle, disease-relevant cues that vary across languages. This loss of discriminative information can hinder cross-lingual generalization, ultimately degrading performance.

4.4.6 Limitation and Future Direction

This study aimed to assess whether speech-based AD detection models trained in one language can generalize to others. The findings indicate that mixed-batch training and fine-tuning enhance cross-lingual transfer, though the extent of transferability was asymmetric across different language pairs. Several limitations remain, requiring further investigation.

Adversarial learning, initially explored as a method for improving language-

invariant feature learning, did not yield the expected benefits and, in some cases, led to decreased performance. A key limitation of this approach is its susceptibility to mode collapse [100], where the model overfits to a limited subset of learned patterns and fails to capture the full diversity of AD-related speech characteristics. To mitigate this issue, various strategies in adversarial training have been proposed. One approach is to use Wasserstein-based loss functions, such as those employed in Wasserstein GANs (WGANs) [101], which replace traditional loss functions with the Wasserstein distance to ensure smoother convergence and better preservation of diverse feature representations. Another promising technique is Unrolled Adversarial Training [102], where the adversarial component is optimized over multiple future iterations, preventing the model from overfitting to a narrow distribution. Future research should explore whether these advanced adversarial techniques improve generalization across languages.

Another key limitation of this study is the variability in datasets. The datasets used differed in preprocessing techniques, clinical diagnostic criteria, recording conditions, sample sizes (English: 156 / Greek: 62 / Slovak: 86), and quality control standards, introducing potential confounding factors. These inconsistencies make it difficult to distinguish true language-specific speech patterns from dataset-related artifacts. Some observed trends may reflect variations in dataset quality rather than intrinsic linguistic differences. Establishing standardized data collection and preprocessing protocols across languages would be essential for ensuring more reliable cross-lingual comparisons. Furthermore, exploring data augmentation techniques tailored for low-resource languages—such as synthetic speech generation, noise injection, and cross-lingual transfer learning—could help mitigate data scarcity and improve model generalizability in multilingual settings [103].

A further limitation is the restricted number of languages examined (n = 3), which limits broader insights into language transferability in speech-based AD detec-

tion. Prior research in NLP suggests that cross-lingual transferability is influenced by multiple factors, including linguistic similarity, lexical overlap, and pre-training configurations [104]. This is particularly relevant for speech-based models, where certain source languages may inherently provide better transferability for specific target languages. The small number of languages in this study prevents a comprehensive analysis of these factors, leaving open the question of how language properties influence the effectiveness of cross-lingual transfer in speech-based AD detection.

Given this, another important direction for future work is the use of feature attribution methods, such as SHAP [105], to analyze model behavior before and after transfer learning. By comparing feature importance scores across languages and transfer stages, such analysis could shed light on the asymmetric transferability observed in our results. Specifically, it may help identify which features are shared across languages and consistently associated with AD, versus those that are language-specific. This would provide more interpretable insights into what the model is actually leveraging for prediction, and guide the design of more robust, cross-lingual systems.

Beyond cross-lingual transfer, the generalizability of speech-based AD detection within a single language remains a challenge due to regional accent diversity and the presence of non-native speakers. Variations in phonetic and prosodic patterns across dialects may impact model performance, while models trained primarily on native speakers risk biases and reduced accuracy for non-native speakers [106]. These challenges are not limited to English but are relevant across many languages with diverse regional and non-native speakers. Future research should explore adaptation techniques, such as accent-invariant feature learning or accent-conditioned modeling, to improve robustness across linguistic and demographic variations.

4.5 Conclusion

This study explored the cross-lingual generalizability of speech-based AD detection models, demonstrating that mixed-batch training and fine-tuning enhance transferability across languages without compromising source-language performance. Additionally, it investigated the feasibility of adversarial learning for cross-lingual adaptation. While this approach did not achieve successful transfer, it provided valuable insights into the challenges of aligning speech representations across languages, guiding future research in robust model adaptation. Lastly, the asymmetric transferability observed suggests that some language pairs adapt more effectively than others, leaving open the question of which languages transfer better and why, an important direction for further exploration.

Developing more robust cross-lingual speech-based models and uncovering hidden linguistic patterns will be crucial for advancing multilingual AD detection systems. Future research should focus on refining adversarial learning techniques, standardizing data collection across languages, and expanding to a broader range of languages and speech tasks. Additionally, addressing intra-language variability, including regional accent differences, will be essential for improving model adaptability and ensuring clinically useful applications. Ultimately, a comprehensive approach that integrates linguistic diversity, methodological rigor, and advanced adaptation techniques will drive the development of more effective and inclusive speech-based AD detection systems.

Chapter 5

Conclusion

This dissertation advances the field of speech-based cognitive impairment detection by demonstrating that speech-derived biomarkers can track cognitive changes over time and transfer across languages, offering a promising tool for early detection. While our findings highlight the potential of speech-based models, they also expose critical challenges and limitations that must be addressed for broader adoption.

In chapter 3, we established that speech-derived features hold longitudinal validity as cognitive assessment markers, showing that models trained at baseline retained moderate predictive power at future time points (6-month and 12-month follow-ups). Despite promising results, predictive performance remained limited, reflecting a complex relationship between speech and cognition—where speech may not always mirror cognitive change. Dataset-specific constraints, such as brief and structured phone-based conversations, and the poor performance of PWB scores, further underscore the challenges of using short, spontaneous speech for longitudinal monitoring. Future work should explore longer, more naturalistic recordings, temporal modeling of speech patterns, and multimodal approaches to improve the reliability of speech-based biomarkers over time.

In chapter 4, we extended this work to a cross-lingual setting, demonstrating that

transfer-learning techniques such as mixed-batch training and fine-tuning improve adaptation across languages without hindering source-language performance. However, the observed asymmetric transferability leaves open questions about which languages transfer best and why. Moreover, our exploration of adversarial learning for cross-lingual transfer, while unsuccessful, provided valuable insights into the challenges of aligning speech representations across languages. These results highlight the need for more advanced adaptation techniques that can better balance language-invariant and language-specific features, ensuring models remain effective across diverse linguistic and demographic contexts.

Despite important progress, major gaps remain in our ability to track AD progression through speech across diverse populations and real-world settings. The field still lacks large-scale, standardized, and longitudinal datasets that capture linguistic diversity, span disease stages, and include consistent cognitive assessments [107]. Challenges such as inconsistent annotation practices, limited labeled data, and variable recording conditions continue to undermine model generalizability and reliability [108].

Cross-lingual modeling offers promise, but disparities in linguistic resources lead to uneven performance across languages and dialects [109]. Data scarcity in minority variants and regional dialects, combined with high variability in speech patterns and recording environments [110], further complicates model adaptation and transfer. Together, these limitations highlight the need for more inclusive, well-curated, and harmonized speech datasets to support robust, generalizable detection systems.

Current models also fail to capture individualized speech characteristics, such as personal conversation styles, emotional tone, and speaker variability. These factors could serve as early subclinical markers of cognitive decline, yet their role in predictive modeling remains underexplored—especially in everyday, non-laboratory settings.

Moreover, there is limited insight into how the observed speech patterns align with clinically and linguistically meaningful measures of cognitive decline. Addressing

this gap calls for interpretable models that connect extracted features to established linguistic and neurological markers, ensuring that speech-based AD detection is grounded in clinical frameworks.

Advancing toward scalable, accessible, ethical, and inclusive screening for cognitive impairments calls for multiple strategic steps. First, standardizing speech collection protocols and annotation practices can foster cross-study comparisons and model replication. Second, privacy-preserving machine learning approaches [111]—such as federated or encrypted methods—can facilitate broader data-sharing while safeguarding patient confidentiality. Third, integrating multimodal data, from speech signals to motor activity, will help capture a more holistic view of early AD manifestations. Ultimately, the goal is a future in which speech-based tools serve as widely accessible, non-invasive pre-screening methods, seamlessly running on everyday devices to continuously monitor cognitive status. Such solutions have the potential to democratize early detection, enable timely interventions, and improve health outcomes across diverse populations worldwide.

Bibliography

[1] Simon Lovestone. *Alzheimer's Disease and Other Dementias (Including Pseudodementias)*, chapter 9, pages 543–615. John Wiley & Sons, Ltd, 2009.

[2] Martin Prince, Anders Wimo, Maelenn Guerchet, Gemma-Claire Ali, Yu-Tzu Wu, Matthew Prina, et al. The global impact of dementia: an analysis of prevalence, incidence, cost and trends. *World Alzheimer Report*, 2015:84, 2015.

[3] Ronald C Petersen. Mild cognitive impairment. *CONTINUUM: lifelong Learning in Neurology*, 22(2):404–418, 2016.

[4] Siegfried Kasper, Christian Bancher, Anne Eckert, Hans Förstl, Lutz Frölich, Jakub Hort, Amos D Korczyn, Reto W Kressig, Oleg Levin, and María Sagrario Manzano Palomo. Management of mild cognitive impairment (mci): the need for national and international guidelines. *The World Journal of Biological Psychiatry*, 21(8):579–594, 2020.

[5] Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and Brian MacWhinney. An overview of the adress-m signal processing grand challenge on multilingual alzheimer's dementia recognition through spontaneous speech. *IEEE Open Journal of Signal Processing*, 5:738–749, 2024.

[6] Nicole D Anderson. State of the science on mild cognitive impairment (mci). *CNS spectrums*, 24(1):78–87, 2019.

[7] Hyunjoo Choi. Performances in a picture description task in japanese patients with alzheimer's disease and with mild cognitive impairment. *Communication Sciences & Disorders*, 14(3):326–337, 2009.

[8] Seyed Ahmad Sajjadi, Karalyn Patterson, Michal Tomek, and Peter J Nestor. Abnormalities of connected speech in semantic dementia vs alzheimer's disease. *Aphasiology*, 26(6):847–866, 2012.

[9] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. *arXiv preprint arXiv:2004.06833*, 2020. To appear in the Proceedings of INTERSPEECH 2020, Oct 2020, Shanghai, China.

[10] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39, 2013.

[11] László Tóth, Ildikó Hoffmann, Gábor Gosztolya, Veronika Vincze, Gréta Szatlóczki, Zoltán Bánreti, Magdolna Pákáski, and János Kálmán. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research*, 15(2):130–138, 2018.

[12] Karmele Lopez-de Ipiña, Jesús B Alonso, Jordi Solé-Casals, Nora Barroso, Patricia Henriquez, Marcos Faundez-Zanuy, Carlos M Travieso, Miriam Ecay-Torres, Pablo Martinez-Lage, and Harkaitz Eguiraun. On automatic diagnosis of alzheimer's disease based on spontaneous speech analysis and emotional temperature. *Cognitive Computation*, 7:44–55, 2015.

[13] Ali Khodabakhsh, Fatih Yesil, Ekrem Guner, and Cenk Demiroglu. Evaluation of linguistic and prosodic features for detection of alzheimer's disease in turkish conversational speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(9), 2015.

[14] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2015.

[15] Fasih Haider, Sofia De La Fuente, and Saturnino Luz. An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):272–281, 2019.

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

[17] Chitralekha Bhat and Sunil Kumar Kopparapu. Identification of alzheimer's disease using non-linguistic audio descriptors. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.

[18] Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. Spoken language derived measures for detecting mild cognitive impairment. *IEEE transactions on audio, speech, and language processing*, 19(7):2081–2090, 2011.

[19] Flavio Bertini, Davide Allevi, Gianluca Lutero, Laura Calza, and Danilo Montesi. An automatic alzheimer's disease classifier based on spontaneous spoken english. *Computer Speech & Language*, 72:101298, 2022.

[20] Zhaoci Liu, Zhiqiang Guo, Zhenhua Ling, and Yunxia Li. Detecting alzheimer's disease from speech using neural networks with bottleneck features and data augmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7323–7327. IEEE, 2021.

[21] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[22] R'Mani M. Haulcy and James Glass. Classifying alzheimer's disease using audio and text-based representations of speech. *Frontiers in Psychology*, 11:624137, 2021.

[23] Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and Brian MacWhinney. Multilingual alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE, 2023.

[24] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.

[25] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[26] Minchuan Chen, Chenfeng Miao, Jun Ma, Shaojun Wang, and Jing Xiao. Exploring multi-task learning and data augmentation in dementia detection with self-supervised pretrained models. In *Interspeech 2023*, pages 5037–5041, 2023.

[27] Dalia Rodríguez-Salas, Nishant Ravikumar, Mathias Seuret, and Andreas Maier. Forestnet–automatic design of sparse multilayer perceptron network architectures using ensembles of randomized trees. In *Asian Conference on Pattern Recognition*, pages 32–45, Cham, 2019. Springer International Publishing.

[28] Paula Andrea Pérez-Toro, Dalia Rodríguez-Salas, Tomás Arias-Vergara, Philipp Klumpp, Maria Schuster, Elmar Nöth, Juan Rafael Orozco-Arroyave, and Andreas K. Maier. Interpreting acoustic features for the assessment of alzheimer's disease using forestnet. *Smart Health*, 26:100347, 2022.

[29] Kewen Ding, Madhu Chetty, Azadeh Noori Hoshyar, Tanusri Bhattacharya, and Britt Klein. Speech based detection of alzheimer's disease: a survey of ai techniques, datasets and challenges. *Artificial Intelligence Review*, 57(12):325, 2024.

[30] Saturnino Luz, Fasih Haider, Sofia de la Fuente Garcia, Davida Fromm, and Brian MacWhinney. Editorial: Alzheimer's dementia recognition through spontaneous speech. *Frontiers in Computer Science*, 3, 2021.

[31] Muet Zhu, Anran Li, and Xuefeng Liu. Tackling the adresso challenge 2021: The muet-rmit system for alzheimer's dementia recognition from spontaneous speech. In *Proceedings of Interspeech 2021*, pages 3870–3874, 2021.

[32] Kristin Qi, Jiatong Shi, Caroline Summerour, John A Batsis, and Xiaohui Liang. Exploiting longitudinal speech sessions via voice assistant systems for early detection of cognitive decline. In *2024 IEEE International Conference on E-health Networking, Application & Services (HealthCom)*, pages 1–6. IEEE, 2024.

[33] Longbin Jin, Yealim Oh, Hyunseo Kim, Hyuntaek Jung, Hyo Jin Jon, Jung Eun Shin, and Eun Yi Kim. Consen: Complementary and simultaneous ensemble for alzheimer's disease detection and mmse score prediction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE, 2023.

[34] Bastiaan Tamm, Rik Vandenberghe, and Hugo Van hamme. Cross-lingual transfer learning for alzheimer's detection from spontaneous speech. *arXiv preprint*, arXiv:2303.03049, 2023.

[35] Xuchu Chen, Yu Pu, Jinpeng Li, and Wei-Qiang Zhang. Cross-lingual alzheimer's disease detection based on paralinguistic and pre-trained features. *arXiv preprint*, arXiv:2303.07650, 2023.

[36] Gábor Gosztolya, Réka Balogh, Nóra Imre, José Vicente Egas-López, Ildikó Hoffmann, Veronika Vincze, László Tóth, Davangere P. Devanand, Magdolna Pákáski, and János Kálmán. Cross-lingual detection of mild cognitive impairment based on temporal parameters of spontaneous speech. *Computer Speech & Language*, 69:101215, 2021.

[37] Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. Multi-source cross-lingual model transfer: Learning what to share. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy, July 2019. Association for Computational Linguistics.

[38] Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2832–2838, 2017.

[39] Zhiqiang Guo, Zhaoci Liu, Zhenhua Ling, Shijin Wang, Lingjing Jin, and Yunxia Li. Text classification by contrastive learning and cross-lingual data augmentation for Alzheimer‘s disease detection. pages 6161–6171, Barcelona,

Spain (Online), December 2020. International Committee on Computational Linguistics.

[40] Kumar B Rajan, Jennifer Weuve, Lisa L Barnes, Emily A McAninch, Robert S Wilson, and Denis A Evans. Population estimate of people with clinical alzheimer's disease and mild cognitive impairment in the united states (2020–2060). *Alzheimer's & Dementia*, 17(12):1966–1975, 2021.

[41] Jill Rasmussen and Haya Langerman. Alzheimer's disease–why we need early diagnosis. *Degenerative neurological and neuromuscular disease*, pages 123–130, 2019.

[42] Guy McKhann, David Knopman, Howard Chertkow, Bradley Hyman, Clifford Jr. Jack, Claudia Kawas, and Creighton Phelps. The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging – alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 7(3):263–269, 2011.

[43] Shana D. Stites, Kristin Harkins, Jonathan D. Rubright, and Jason Karlawish. Relationships between cognitive complaints and quality of life in older adults with mild cognitive impairment, mild alzheimer's disease dementia, and normal cognition. *Alzheimer Disease & Associated Disorders*, 32(4):276–283, 2018.

[44] Kyle Masato Ishikawa, James Davis, John J. Chen, and Eunjung Lim. The prevalence of mild cognitive impairment by aspects of social isolation. *PLoS ONE*, 17(6):e0269795, 2022.

[45] Isobel EM Evans, Anthony Martyr, Rachel Collins, Carol Brayne, and Linda Clare. Social isolation and cognitive function in later life: a systematic review and meta-analysis. *Journal of Alzheimer's disease*, 70(s1):S119–S144, 2019.

[46] Archana Singh-Manoux, Manasa S Yerramalla, Severine Sabia, Mika Kivimäki, Aurore Fayosse, Aline Dugravot, and Julien Dumurgier. Association of big-5 personality traits with cognitive impairment and dementia: a longitudinal study. *J Epidemiol Community Health*, 74(10):799–805, 2020.

[47] Raimundo J Mourao, Guilherme Mansur, Leandro F Malloy-Diniz, Erico Castro Costa, and Breno S Diniz. Depressive symptoms increase the risk of progression to dementia in subjects with mild cognitive impairment: systematic review and meta-analysis. *International journal of geriatric psychiatry*, 31(8):905–911, 2016.

[48] Zahinoor Ismail, Heba Elbayoumi, Corinne E Fischer, David B Hogan, Colleen P Millikin, Tom Schweizer, Moyra E Mortby, Eric E Smith, Scott B Patten, and Kirsten M Fiest. Prevalence of depression in patients with mild cognitive impairment: a systematic review and meta-analysis. *JAMA psychiatry*, 74(1):58–67, 2017.

[49] Laurin Plank and Armin Zlomuzica. Reduced speech coherence in psychosis-related social media forum posts. *Schizophrenia*, 10(1):60, 2024.

[50] Kexin Yu, Katherine Wild, Kathleen Potempa, Benjamin M Hampstead, Peter A Lichtenberg, Laura M Struble, Patrick Pruitt, Elena L Alfaro, Jacob Lindsley, Mattie MacDonald, et al. The internet-based conversational engagement clinical trial (i-conect) in socially isolated adults 75+ years old: randomized controlled trial protocol and covid-19 related study modifications. *Frontiers in digital health*, 3:714813, 2021.

[51] James Lubben, Eva Blozik, Gerhard Gillmann, Steve Iliffe, Wolfgang von Renteln Kruse, John C Beck, and Andreas E Stuck. Performance of an abbreviated version of the lubben social network scale among three european community-dwelling older adult populations. *The Gerontologist*, 46(4):503–513, 2006.

[52] Mary Elizabeth Hughes, Linda J Waite, Louise C Hawkley, and John T Cacioppo. A short scale for measuring loneliness in large surveys: Results from two population-based studies. *Research on aging*, 26(6):655–672, 2004.

[53] Jerome A Yesavage, Terence L Brink, Terence L Rose, Owen Lum, Virginia Huang, Michael Adey, and Von Otto Leirer. Development and validation of a geriatric depression screening scale: a preliminary report. *Journal of psychiatric research*, 17(1):37–49, 1982.

[54] Sandra Weintraub, Lilah Besser, Hiroko H Dodge, Merilee Teylan, Steven Ferris, Felicia C Goldstein, Bruno Giordani, Joel Kramer, David Loewenstein, Dan Marson, et al. Version 3 of the alzheimer disease centers' neuropsychological test battery in the uniform data set (uds). *Alzheimer Disease & Associated Disorders*, 32(1):10–17, 2018.

[55] John C Morris. The clinical dementia rating (cdr) current version and scoring rules. *Neurology*, 43(11):2412–2412, 1993.

[56] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005.

[57] Ziad S. Nasreddine. Montreal cognitive assessment (moca) – blind, version august 18, 2010. `https://www.mocatest.org/wp-content/uploads/2020/10/MoCA-Blind-English-August-2017.pdf`, 2010. Accessed: 2025-04-20.

[58] Xiaofan Mu, Salman Seyedi, Iris Zheng, Zifan Jiang, Liu Chen, Bolaji Omofojoye, Rachel Hershenberg, Allan I. Levey, Gari D. Clifford, Hiroko H. Dodge, and Hyeokhyen Kwon. Detecting cognitive impairment and psychological well-being among older adults using facial, acoustic, linguistic, and cardiovascular patterns derived from remote conversations. *arXiv preprint arXiv:2412.14194*, 2024.

[59] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[60] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[61] Theodoros Giannakopoulos. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12):e0144610, 2015.

[62] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[63] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[64] Jochen Hartmann. Emotion english distilroberta-base. `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/`, 2022.

[65] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87, 2023.

[66] Robert R McCrae, Paul T Costa, Jr, and Thomas A Martin. The neo–pi–3: A more readable revised neo personality inventory. *Journal of personality assessment*, 84(3):261–270, 2005.

[67] NIH Toolbox. Nih toolbox for the assessment of neurological and behavioral function. `https://www.healthmeasures.net/explore-measurement-systems/nih-toolbox`, 2023. Accessed: 2025-03-01.

[68] Ida Babakhanyan, Benjamin S McKenna, Kaitlin B Casaletto, Cindy J Nowinski, and Robert K Heaton. National institutes of health toolbox emotion battery for english-and spanish-speaking adults: normative data and factor-based summary scores. *Patient related outcome measures*, pages 115–127, 2018.

[69] Zifan Jiang, Salman Seyedi, Emily Griner, Ahmed Abbasi, Ali Bahrami Rad, Hyeokhyen Kwon, Robert O Cotes, and Gari D Clifford. Multimodal mental health digital biomarker analysis from remote interviews using facial, vocal, linguistic, and cardiovascular patterns. *IEEE journal of biomedical and health informatics*, 28(3):1680–1691, 2024.

[70] Ciro Rosario Ilardi, Alina Menichelli, Marco Michelutti, Tatiana Cattaruzza, and Paolo Manganotti. Optimal moca cutoffs for detecting biologically-defined patients with mci and early dementia. *Neurological Sciences*, 44(1):159–170, 2023.

[71] Michael Malek-Ahmadi, Kathy O'Connor, Sharon Schofield, David W Coon, and Edward Zamrini. Trajectory and variability characterization of the montreal cognitive assessment in older adults. *Aging Clinical and Experimental Research*, 30:993–998, 2018.

[72] Felix Agbavor and Hualou Liang. Predicting dementia from spontaneous speech using large language models. *PLOS digital health*, 1(12):e0000168, 2022.

[73] Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection. *arXiv preprint arXiv:2008.01551*, 2020.

[74] David S Knopman, Rosebud O Roberts, Yonas E Geda, V Shane Pankratz, Teresa JH Christianson, Ronald C Petersen, and Walter A Rocca. Validation of the telephone interview for cognitive status-modified in subjects with normal cognition, mild cognitive impairment, or dementia. *Neuroepidemiology*, 34(1):34–42, 2010.

[75] Nicola Gates, Michael Valenzuela, Perminder S Sachdev, and Maria A Fiatarone Singh. Psychological well-being in individuals with mild cognitive impairment. *Clinical interventions in aging*, pages 779–792, 2014.

[76] Lorenzo Gaetano Amato, Alberto Arturo Vergani, Michael Lassi, Carlo Fabbiani, Salvatore Mazzeo, Rachele Burali, Benedetta Nacmias, Sandro Sorbi, Riccardo Mannella, Antonello Grippo, et al. Personalized modeling of alzheimer's disease progression estimates neurodegeneration severity from eeg recordings. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 16(1):e12526, 2024.

[77] Farhad Mortezapour Shiri, Thinagaran Perumal, Norwati Mustapha, and Raihani Mohamed. A comprehensive overview and comparative analysis on deep learning models: Cnn, rnn, lstm, gru. *arXiv preprint arXiv:2305.17473*, 2023.

[78] Mohit Shah, Brian Mears, Chaitali Chakrabarti, and Andreas Spanias. Lifelogging: Archival and retrieval of continuously recorded audio using wearable devices. In *2012 IEEE International Conference on Emerging Signal Processing Applications*, pages 99–102. IEEE, 2012.

[79] Jamie Vo, Naeha Sharif, and Ghulam Mubashar Hassan. Multimodal neuroimaging attention-based architecture for cognitive decline prediction. *arXiv preprint arXiv:2401.06777*, 2023.

[80] Chatchawan Rattanabannakit, Shannon L Risacher, Sujuan Gao, Kathleen A Lane, Steven A Brown, Brenna C McDonald, Frederick W Unverzagt, Liana G Apostolova, Andrew J Saykin, and Martin R Farlow. The cognitive change index as a measure of self and informant perception of cognitive decline: relation to neuropsychological tests. *Journal of Alzheimer's Disease*, 51(4):1145–1155, 2016.

[81] Sarah T Farias, Dan Mungas, Bruce R Reed, Danielle Harvey, Deborah Cahn-Weiner, and Charles DeCarli. Mci is associated with deficits in everyday functioning. *Alzheimer Disease & Associated Disorders*, 20(4):217–223, 2006.

[82] World Health Organization. Dementia, 2023. Accessed: 1 March 2025.

[83] International Organization for Migration. World migration report 2024. `https://worldmigrationreport.iom.int/wmr-2024-interactive/`, 2024. Accessed: 2025-03-20.

[84] Wouter M Kouw and Marco Loog. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018.

[85] Mandy WM Fong, Ryan Van Patten, and Robert P Fucetola. The factor structure of the boston diagnostic aphasia examination. *Journal of the International Neuropsychological Society*, 25(7):772–776, 2019.

[86] Milan Rusko, Róbert Sabo, Marián Trnka, Alfréd Zimmermann, Richard Malaschitz, Eugen Ružickỳ, Petra Brandoburová, Viktória Kevická, and Matej Škorvánek. Slovak database of speech affected by neurodegenerative diseases. *Scientific Data*, 11(1):1–16, 2024.

[87] M. Rusko, R. Sabo, M. Trnka, A. Zimmermann, R. Malaschitz, E. Ružický, P. Brandoburová, V. Kevická, and M. Škorvánek. Ewa-db – early warning of alzheimer speech database. `https://catalogue.elra.info/en-us/repository/browse/ELRA-S0489/`, 2023. Accessed via ELRA Catalogue (ID: ELRA-S0489).

[88] William Falcon and The PyTorch Lightning team. Pytorch lightning. `https://github.com/PyTorchLightning/pytorch-lightning`, March 2019. The lightweight PyTorch wrapper for high-performance AI research. Scale your models, not the boilerplate.

[89] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[90] Renata Gregova. Comparative phonetics and phonology of the english and the slovak language. *Kosice: Vydavatel'stvo SafarikPress UPJS*, 2022.

[91] Peter Ladefoged, Keith Johnson, and Peter Ladefoged. *A course in phonetics*, volume 3. Thomson Wadsworth Boston, 2006.

[92] Angeliki Malikouti-Drachman. Greek phonology: A contemporary perspective. *Journal of Greek Linguistics*, 2(1):187–243, 2002.

[93] Juan JG Meilán, Francisco Martínez-Sánchez, Israel Martínez-Nicolás, Thide E Llorente, and Juan Carro. Changes in the rhythm of speech difference between people with nondegenerative mild cognitive impairment and with preclinical dementia. *Behavioural neurology*, 2020(1):4683573, 2020.

[94] Ravi Shankar, Anjali Bundele, and Amartya Mukhopadhyay. A systematic review of natural language processing techniques for early detection of cognitive impairment. *Mayo Clinic Proceedings: Digital Health*, page 100205, 2025.

[95] Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. Choosing transfer languages for cross-lingual learning. *arXiv preprint arXiv:1905.12688*, 2019.

[96] Seong Hoon Lim, Taejun Yun, Jinhyeon Kim, Jihun Choi, and Taeuk Kim. Analysis of multi-source language training in cross-lingual transfer. *arXiv preprint arXiv:2402.13562*, 2024.

[97] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*, 2020.

[98] Daniel Kempler and Mira Goral. Language and dementia: Neuropsychological aspects. *Annual review of applied linguistics*, 28:73–90, 2008.

[99] Greta Szatloczki, Ildiko Hoffmann, Veronika Vincze, Janos Kalman, and Magdolna Pakaski. Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease. *Frontiers in aging neuroscience*, 7:195, 2015.

[100] Youssef Kossale, Mohammed Airaj, and Aziz Darouichi. Mode collapse in generative adversarial networks: An overview. In *2022 8th International Conference on Optimization and Applications (ICOA)*, pages 1–6. IEEE, 2022.

[101] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[102] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

[103] Zolzaya Byambadorj, Ryota Nishimura, Altangerel Ayush, Kengo Ohta, and Norihide Kitaoka. Text-to-speech system for low-resource language using cross-lingual transfer learning and data augmentation. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1):42, 2021.

[104] Fred Philippy, Siwen Guo, and Shohreh Haddadan. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. *arXiv preprint arXiv:2305.16768*, 2023.

[105] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[106] Ingy Farouk Emara and Nabil Hamdy Shaker. The impact of non-native english speakers' phonological and prosodic features on automatic speech recognition accuracy. *Speech Communication*, 157:103038, 2024.

[107] James W Schwoebel, Joel Schwartz, Lindsay A Warrenburg, Roland Brown, Ashi Awasthi, Austin New, Monroe Butler, Mark Moss, and Eleftheria K Pissadaki. A longitudinal normative dataset and protocol for speech and language biomarker research. *medrxiv*, pages 2021–08, 2021.

[108] Lara Gauder, Pablo Riera, Andrea Slachevsky, Gonzalo Forno, Adolfo M Garcia, and Luciana Ferrer. The unreliability of acoustic systems in alzheimer's speech datasets with heterogeneous recording conditions. *arXiv preprint arXiv:2409.12170*, 2024.

[109] Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362, 2025.

[110] Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. Quantifying the dialect gap and its correlates across languages. *arXiv preprint arXiv:2310.15135*, 2023.

[111] Vankamamidi S Naresh and Muthusamy Thamarai. Privacy-preserving data mining and machine learning in healthcare: Applications, challenges, and solutions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2):e1490, 2023.