Identifying Early Markers of Alzheimer's Disease through Longitudinal Analysis of Language and Speech Patterns

Kunxiao Gao

September 2024

Abstract

Alzheimer's disease (AD) often manifests with early speech and language impairments, making these features valuable for detection and monitoring. This study analyzes longitudinal speech data from 80 individuals—40 with AD and 40 neurologically health (CN) controls—over two intervals: 10 and 5 years before the year of diagnosis (YoD). Acoustic features, such as intensity and pitch variability, and linguistic measures, including lexical diversity and syntactic complexity, were examined to identify temporal trends and group differences. Results highlight significant declines in acoustic and linguistic features in the AD group, while CN participants showed relative stability. Non-interpretable feature models (NIFMs) outperformed interpretable models (IFMs), underscoring the value of comprehensive features for early detection. Despite challenges with data noise and sparsity, this study demonstrates the potential of speech analysis as a non-invasive tool for early AD diagnosis and progression tracking, paving the way for enhanced diagnostic methods and personalized interventions.

1 Introduction

Alzheimer's disease (AD) is marked by a gradual decline across various cognitive domains, including language and speech [1]. Frequent manifestations such as aphasia and dysarthria highlight the prevalence of language impairments in Alzheimer's patients. In particular, individuals with AD may exhibit subtle prodromal characteristics for several years before the manifestation of obvious symptoms [2]. These early impairments often include lexical retrieval difficulties, reduced verbal fluency, and initial breakdowns in spoken language comprehension [3, 4]. Distortion is one of the most prevalent speech-related manifestations of AD, affecting over 80% of patients, often accompanied by slower speech rate [5]. These subtle linguistic and phonetic-phonological alterations can emerge years before the onset of more overt symptoms. Even minor behavioral changes may be identified during the presymptomatic phase using sensitive measurement methods. Based on a few prior studies [3], it is plausible to use acoustic and linguistic measures of speech to aid in the development of early diagnostic tools [6].

Identifying AD at an early stage offers significant benefits that go well beyond merely establishing a diagnosis. As new therapeutic options become available, early detection becomes crucial for evaluating the effectiveness of these treatments promptly. Advanced diagnostic tools—such as biomarkers and neuroimaging techniques—can provide sensitive and objective assessments of how patients respond to emerging therapies during the preclinical or mild cognitive impairment phases. This enables healthcare professionals to personalize interventions more effectively, potentially slowing the progression of the disease [2]. Moreover, early identification of individuals at risk or in the initial stages of AD has important implications for research and clinical trials. It allows for strategically recruiting participants who are most likely to benefit from experimental treatments, thereby enhancing the success rates of clinical studies and accelerating the development of effective therapeutic strategies [7].

Despite the significant potential for using speech analysis in early AD detection, research in this area remains limited. There is a particular scarcity of studies examining how speech patterns change during the prodromal phase of AD. Gaining a comprehensive understanding of how speech-related symptoms evolve in relation to disease progression requires extensive longitudinal research. Furthermore, while there is increasing interest in applying speech-based machine learning and deep learning techniques for AD detection, the development of robust models is hindered by the lack of ample and relevant training data in clinical contexts. Existing diagnostic and monitoring methods that utilize speech predominantly rely on post-diagnosis data collected from a small number of individuals in controlled environments, which may not capture the full spectrum of speech alterations associated with the disease.

This work makes the following contributions to overcome these challenges and address the gaps in the existing literature:

1. Collecting a longitudinal speech data set of spontaneous speech, *ADCeleb*, containing speech samples from up to 10 years before the year of diagnosis. This data set comprises recordings from 40 celebrities who publicly claimed that they had been diagnosed with AD and who publicly revealed the year of diagnosis (YoD), and 40 control subjects (CNs) matched in sex, age, and ethnicity. For each subject with AD, this data set contains speech samples (when available) ranging from ten years before YoD to the year before diagnosis.

2. Investigating the effectiveness of different classification techniques at different pre-diagnosis stages of AD using both interpretable feature-based models (IFMs) and non-interpretable feature-based models (NIFMs). The primary objectives are to determine at what point in time it becomes feasible to distinguish the speech patterns of subjects with AD from those of CNs and to establish baseline performance on this data set.

3. Conducting a longitudinal study to monitor possible changes in various speech and language-based features, starting from the 10 years before diagnosis and considering an average observation period of 10 years. The longitudinal study considers both subjects with AD and CNs, acknowledging the potential overlap of AD symptoms with normal aging. The study aims to distinguish speech changes specific to AD from those influenced by other factors, such as age.

In essence, the collection of pre-diagnosis data in the creation of ADCeleb, the investigation of the predictive capabilities of various models developed using data belonging to various stages of the disease, and the design of a longitudinal study spanning the prodromal period of AD contribute to a more nuanced understanding of how speech evolves in time due to AD. This approach not only enriches the literature but also holds promise for early detection and intervention strategies.

2 Related Work

The following subsections will review prior research in AD detection and monitoring, summarizing key developments and findings in the field. A separate subsection will then focus on the common limitations frequently encountered in these studies.

2.1 AD Detection

Early research on AD detection using speech data has focused on capturing both acoustic and linguistic features that reflect the cognitive-linguistic decline associated with the disease. Acoustic measures, such as jitter, shimmer, pitch and intensity variability, and formant frequencies, have been widely used to differentiate between individuals with AD and CN [8–13]. These features aim to capture motor-related changes in speech, often linked to the progression of neurodegenerative conditions [13,14]. Prosodic aspects, such as speech rate, articulation rate, and average syllable duration, have also been explored in AD studies, as speech becomes slower and more effortful in individuals with AD [8]. In addition to acoustic features, linguistic measures reflecting lexical and syntactic complexity have proven valuable in identifying AD-related changes. Studies have shown that individuals with AD exhibit reduced lexical diversity, shorter and simpler syntactic structures, and increased repetition of certain words, indicating cognitive impairment [15–19]. Commonly used linguistic features include the Type-Token Ratio (TTR), Root Type-Token Ratio (RTTR), counts of noun and verb phrases, dependency lengths, and the use of connectives, as these markers reflect deteriorations in vocabulary and grammatical structure.

In recent years, deep learning approaches, including Convolutional Neural Networks and Transformer-based architectures, have been applied to the task of AD detection using speech data, with notable models such as BERT and LSTM demonstrating substantial improvements over traditional classification techniques [20, 21]. These models can capture complex patterns in both acoustic and linguistic data that might be missed by hand-crafted features. Additionally, speaker recognition technologies like x-vectors have been explored for parameterizing distinctive speech traits in AD, contributing to more robust AD detection frameworks [7, 22]. Despite advancements in AD detection using DL, the field faces challenges related to data scarcity, particularly for pre-diagnosis and longitudinal data. The effectiveness of DL models is contingent upon access to large datasets encompassing varied patient information, including demographic and health history details, to allow for comprehensive training and evaluation [20]. To address this, researchers have increasingly relied on transfer learning and domain adaptation to leverage existing datasets and enhance model performance with limited medical

data [23–25] . However, due to the scarcity of prodromal data in AD research, prior studies predominantly use post-diagnosis data, often capturing a single or limited number of disease stages, with few studies investigating the earliest stages of cognitive impairment that precede diagnosis.

2.2 AD Monitoring

Analyzing speech patterns in individuals with AD has emerged as a promising method for detecting early cognitive decline, offering an objective, non-invasive alternative to traditional diagnostic tools. Conventional assessments, which often rely on subjective scales, may lack sensitivity in identifying subtle early-stage symptoms and are influenced by inter-rater variability [1]. In contrast, speech analysis provides insight into both cognitive and motor functions affected by AD, allowing researchers to track linguistic and acoustic deterioration over time. Linguistic markers are particularly indicative of cognitive decline, with studies showing reduced lexical diversity, syntactic simplification, and increased disfluencies—all pointing to impaired language processing and vocabulary access in AD patients [2, 4]. Measures like Type-Token Ratio (TTR) and Root Type-Token Ratio (RTTR) have proven effective in quantifying these deficits, as lower values typically indicate reduced vocabulary range, a common sign of language impairment in AD [3, 26].

Acoustic markers also play a crucial role in AD detection. Research demonstrates that AD patients exhibit diminished pitch and intensity variability, which are associated with a monotonic speech quality that reflects emotional flattening and reduced prosody [5]. Phonetic characteristics, such as jitter (frequency perturbations), shimmer (amplitude perturbations), and changes in fundamental frequency (pitch), help to capture motor speech impairments linked to neurological degradation. These impairments, which are often subtle in the prodromal phase, become more pronounced over time, with pitch variability and intensity fluctuations commonly declining as the disease advances [3, 26]. Additionally, longitudinal studies have shown that as AD progresses, articulation rate and speech rate decline, while average syllable duration increases, suggesting an underlying deterioration in motor planning and execution needed for smooth and fluent speech production [5, 13]. These acoustic features provide quantifiable markers that distinguish AD-related speech changes from those related to normal aging processes. Conversational timing and pause patterns also serve as diagnostic markers for AD. Studies reveal that AD patients frequently produce more prolonged pauses, reflecting both cognitive processing delays and motor planning difficulties [2, 8]. Unlike healthy older adults, AD patients exhibit increased pausing in non-syntactic locations, indicating disruptions in natural speech flow. Timing deviations in conversation are particularly valuable for early detection, as even minor impairments in conversational flow can indicate cognitive decline before overt symptoms emerge [4, 27]. Additionally, AD patients struggle to maintain complex syntactic structures, often producing simpler, more repetitive speech [5].

2.3 Limitations of Previous Studies

1. **Limited longitudinal and prodromal data.** Previous studies have highlighted the scarcity of longitudinal datasets that include the prodromal phase of AD. For instance, research has demonstrated the need for more extensive analysis during these early stages to identify subtle speech and language changes before clinical diagnosis [2, 4, 5]. The lack of such data hampers the ability to design models that can accurately detect AD in its earliest phases.

2. **Data scarcity and diversity challenges.** The limited availability of diverse and comprehensive datasets remains a significant barrier. Studies often rely on relatively small cohorts, which may not adequately represent variations across demographics or disease severities. This lack of diversity reduces the generalizability of models developed for AD detection [13, 20]. Expanding datasets to include broader and more representative populations is essential for future progress.

3. **Methodological and feature limitations.** Methodological challenges, including the reliance on traditional features and the absence of advanced computational approaches, have constrained the field. Recent work has begun to address these limitations using techniques such as domain adaptation and transfer learning, which leverage existing data to enhance model performance even with limited new data [23, 28]. However, these efforts need to be further developed to improve the robustness of AD detection frameworks.

4. **Lack of multimodal integration.** While speech and language features are valuable, the integration of multimodal data—such as neuroimaging or physiological biomarkers—has been limited in AD research. Multimodal approaches can offer more comprehensive insights into disease progression, yet they remain underexplored [1, 27]. Future studies should prioritize the combination of linguistic, acoustic, and multimodal data to enhance diagnostic accuracy and predictive power.

5. **Real-world noise and data quality concerns.** Real-world noise, including background sounds, overlapping speech, and variations in recording quality, significantly impacts the reliability of acoustic and linguistic features. While laboratory settings provide controlled conditions for data collection, most datasets used for AD research include these real-world imperfections, complicating the analysis and interpretation of results [3, 13]. Addressing this issue remains a critical challenge for the field.

3 Method

The following subsections provide a comprehensive overview of the data collection process, followed by an in-depth description of our classification experiments and longitudinal analysis.

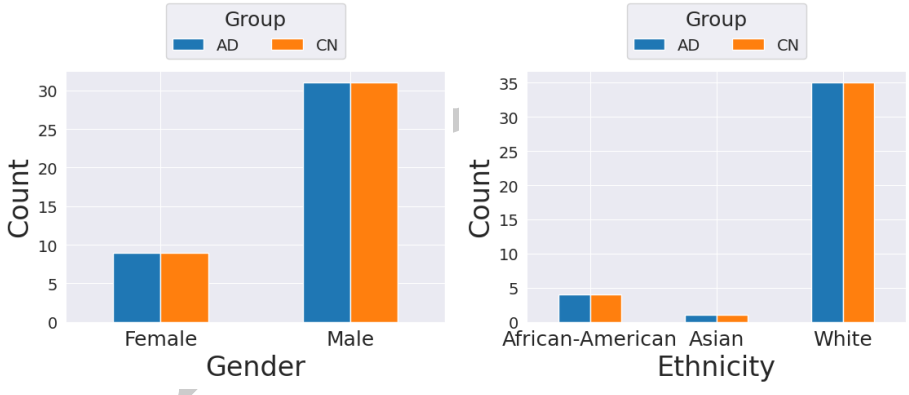
3.1 ADCeleb

In this study, we developed a novel dataset called *ADCeleb*, which contains speech recordings from individuals with AD obtained from publicly accessible sources. ADCeleb encompasses over 5,347 recordings—amounting to approximately 25 h of speech—from 80 celebrities, split evenly between 40 diagnosed with AD and 40 CN counterparts. The speech samples were extracted from videos uploaded to YouTube, with some adhering to Creative Commons licenses and others under the standard YouTube licensing agreements.

The dataset includes a wide range of accents (e.g., British, American, Canadian, Scottish, Finnish), professions, and age groups. The AD and CN groups are matched based on age, sex, and ethnicity. The videos in the dataset feature a limited variety of complex multi-speaker acoustic environments, primarily including settings like calm studio interviews, public conferences, red-carpet interviews, TV show talk delivered to large audiences, and informal talk in daily conversation. Notably, these videos have been exposed to background noise, such as chatter, laughter, overlapping speech, reverberation, and variations in recording equipment quality and channel noise. Table 1 presents the overall statistics of the dataset, grouping the data into two distinct time intervals relative to the Year of Diagnosis (YoD) of the AD subjects. CN subjects are grouped into corresponding intervals based on the YoD of the AD individuals with whom they were matched. The two time intervals (t) used for data division are:

- Time interval -2: from 10 to 6 years before the YoD;
- Time interval -1: from 5 to 1 years before the YoD.

Further information regarding the data collection process and the formation of the experimental groups is outlined in the subsequent section. Figure 1 illustrates the distributions of the two groups according to sex and ethnicity.



(a) *(b)*

Figure 1: Barplot of gender (a) and ethnicity (b) distributions for AD and CN groups.

A larger number of videos featuring male celebrities compared to female celebrities

were found. One possibility is the availability of interview footage. Public figures, like celebrities or politicians, who have openly discussed or shown signs of AD may be mostly male, as men have traditionally had more prominent roles in public life. It is also possible that cultural factors or biases in media coverage lead to more documentation of men with the condition [29]. Additional visualizations related to nationality and profession are available in the Supplementary Materials.

	# of Speakers	# of Recordings (Total)	# of Recordings per Speaker	# of Segments per Speaker	Segment Length (s)
Time Interval -2 (from -10 to -6 years before YoD)					
AD	34	72	1/2.12/4	2/39.88/224	8.01/16.13/109.76
CN	39	73	1/1.87/2	1/30.54/189	8.01/16.01/134.13
Time Interval -1 (from -5 to -1 years before YoD)					
AD	35	83	1/2.37/5	1/34.48/141	8.01/17.67/153.60
CN	38	82	1/2.16/4	1/41.95/197	8.01/16.62/150.0

Table 1: This table presents comprehensive dataset statistics, providing details for both the AD and CN groups during the two considered time intervals: -2 and -1. In fields where three numbers are listed, these correspond to the minimum, average, and maximum values, respectively. The statistical data encompasses the number of speakers, the total number of recordings per speaker, the number of segments per speaker, and the length of segments in seconds. A recording refers to the audio linked to a specific video, while a segment denotes a portion of a recording that contains speech from the target speaker.

3.1.1 Data Collection Pipeline

This section describes the steps involved in the ADCeleb data collection process. Figure 2 presents an overview of the primary stages within the data collection pipeline.

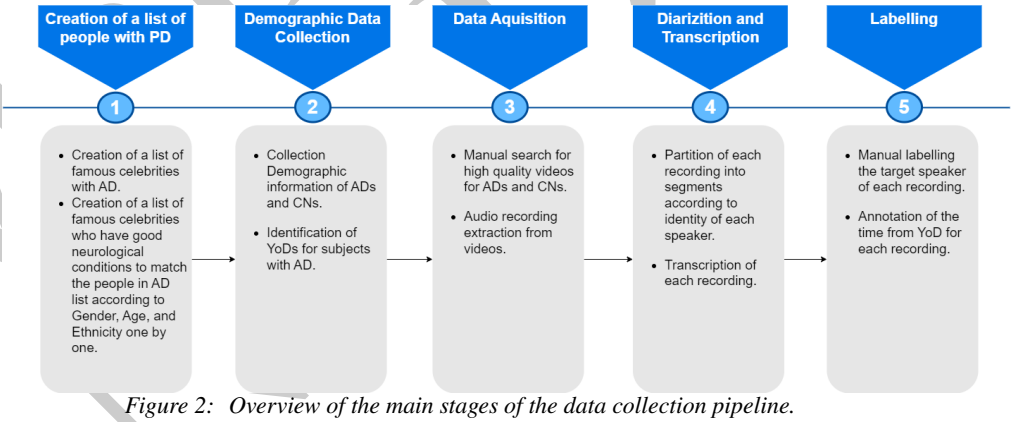


Figure 2: Overview of the main stages of the data collection pipeline.

1. **Creation of a list of people with AD.** We first compiled a comprehensive list of celebrities diagnosed with AD. This was achieved by extracting names from the Wikipedia page titled *Category:People with Alzheimer's disease* [30]. From this compilation, we identified a subset of 40 English-speaking individuals for whom video recordings were available in ten years before YoD. All selected

subjects voluntarily disclosed their AD diagnosis and the year it occurred through interviews on television, in journals, or newspapers. The sources of these publicly available interviews are documented and compiled in the previously mentioned list.

2. **Identification of the years of diagnosis.** In the next phase of our research, we collected fundamental demographic data on the subjects identified previously, which included their sex, age, nationality, ethnicity, professions, and dates of birth and death. We also conducted thorough searches for publicly available interviews or journal articles in which these individuals with AD disclosed their YoD. Subjects were excluded from the study if we could not find any direct or indirect sources confirming the disclosure of their YoD.

3. **Audio download.** At this stage, our primary focus was a manual search for high-quality videos, especially television interviews when accessible. For each subject with AD, we downloaded audio from YouTube videos available up to ten years before their YoD.

4. **Diarization:** WhisperX [31] was used for automatic speech recognition and speaker diarization. Speaker diarization involves partitioning an audio stream containing human speech based on the identity of each speaker.

5. **Annotation of the target speaker and year.** Following the diarization process, we manually categorized each speaker in every recording as either a target or a non-target individual. The target label was assigned to speakers identified in the initial list of subjects with AD from stage one. Additionally, we annotated the temporal interval indicating how long before or after the diagnosis each video was recorded. At this stage, we implemented measures to verify the accuracy of the video's original recording year, acknowledging potential discrepancies between the upload date and the actual recording date. To address this issue, we cross-referenced video metadata, including titles and descriptions.

6. **Creation of a CN group:** After completing data collection for the AD group, we replicated the pipeline to gather data for the CN group. Specifically, we matched 40 celebrities without known neurological or psychological impairments in a 1:1 ratio with AD subjects based on sex, age, and ethnicity. The comprehensive list of subjects with AD and their matched CN counterparts is presented in Tables 1 and 2 of the Supplementary Materials. Subjects assigned the same identification number represent a matched pair (e.g., ad_01 with cn_01; ad_02 with cn_02).

3.2 Classification Experiments

By performing a series of classification experiments, we evaluated and compared the predictive capabilities of machine learning models based on both interpretable and non-interpretable features to detect AD at various stages of its progression using speech analysis. We aim to determine whether speech and language patterns indicative of AD can be identified during the prodromal phase. For a detailed explanation of the models and the classification pipeline utilized, please refer to Section 4. The classification

experiments conducted aimed to rigorously evaluate the predictive capabilities of various machine learning models at distinct stages of AD. The primary goal was to determine whether it is feasible to distinguish between the speech patterns of individuals with AD and those of CN even before a formal diagnosis, thereby offering insights into potential early biomarkers of the disease. To achieve this, we utilized the ADCeleb dataset to train and test multiple models, each focusing on one of the two specified time intervals relative to diagnosis (namely, intervals -2 and -1). Each experiment was independently executed using data exclusively from a single time interval to assess the models' performance at that specific stage.

To rigorously evaluate our models in the experiment, we implemented a Nested Cross-Validation (NCV) strategy. The dataset was initially divided into ten folds for the outer cross-validation loop. In each outer iteration, one fold was designated as the test set, while the remaining nine folds were combined to form the outer training set. This outer training set was then further partitioned into stratified inner folds for the purposes of inner-loop training and validation. Within these inner folds, we conducted an exhaustive grid search to optimize hyperparameters. This procedure yielded ten distinct sets of optimal configurations, each corresponding to the best hyperparameters identified during the inner cross-validation of their respective outer iterations.

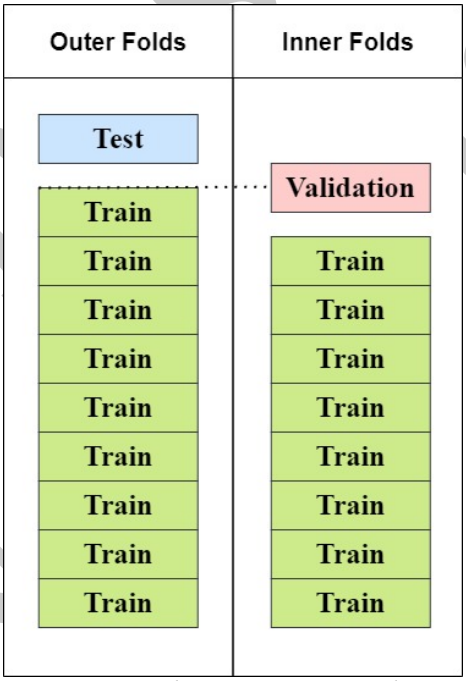


Figure 3: Mono-corpus experiment - data partition on a single time interval of ADCeleb.

In the mono-corpus experiments, we utilized a NCV approach, as illustrated in Figure 3. During each outer iteration, data standardization was performed by subtracting the mean and dividing by the standard deviation, both calculated from the training set. Independent mono-corpus experiments were carried out for the two distinct time intervals

defined within the ADCeleb dataset, ensuring a focused analysis of each temporal stage.

3.3 Longitudinal Analysis

The longitudinal analysis emphasizes exploring interpretable acoustic and linguistic features that are simple to extract from spontaneous speech. By focusing on features that can be derived directly from natural, unscripted interactions, we aimed to capture significant acoustic and linguistic patterns that reflect real-world communication dynamics. This choice supports a more realistic and practical investigation, offering an in-depth perspective on the differences and similarities within and across experimental groups in the context of spontaneous speech.

3.3.1 Features

Acoustics Features The longitudinal study examined the temporal trajectory of acoustics features related to prosody and articulation. Prosody refers to the way elements of speech—such as intonation, stress, and rhythm—are structured to convey meaning and emotion. In people with AD, prosody is significantly affected, and these changes can be noticed even in the early stages. Research shows that pitch, intonation, and stress often decline over time, leading to a speech pattern that sounds flat and lacks expressiveness [5, 11]. This reduced pitch variability and disrupted rhythm make it challenging for people with AD to convey emotions effectively, which affects their overall communication quality. For example, Roark et al. [14] found that individuals with AD showed less variation in pitch and more monotone intonation, which reduces their ability to express emotions and manage different speech elements properly.

Articulation is also affected as AD progresses, although the impact is usually less severe than in other conditions like Parkinson's disease. People with AD may experience slower speech rates, longer pauses, and inconsistent pronunciation [5, 13, 32]. These issues often arise due to difficulties in planning speech movements and retrieving words, which are influenced by cognitive decline. Boschi et al. [33] reported that people with AD tend to have longer segment durations and more frequent pauses, indicating difficulty in accessing words and maintaining smooth speech. Taler and Phillips [16] highlighted that these challenges lead to reduced fluency, with increased hesitations, slower speaking rates, and less precise articulation.

Cognitive decline in AD impacts not just the prosody and articulation but also the overall ability to control speech production. Issues with executive functioning make it difficult for individuals to manage timing in speech, resulting in frequent, often inappropriate, pauses that disrupt the flow. Forbes-McKay and Venneri [15] observed that people with AD have more pauses, both in frequency and length, which breaks up speech and makes it less coherent. These disruptions are compounded by challenges with planning movements of the vocal tract and reduced control over muscles, which affect speech clarity and resonance.

Prosodic and articulatory changes have been studied as possible biomarkers for early detection of AD. According to König et al. [12] , changes in features like reduced pitch range and increased variability in speaking rate can help distinguish individuals with AD from those without the disease. These features, if measured consistently, could

be effective indicators for early diagnosis. These markers could potentially help in identifying the onset of AD before more severe symptoms develop.To characterize the impact of D on prosody, we analyzed features related to pitch and intensity, including the standard deviation of the fundamental frequency and the standard deviation of intensity. The intensity, I (dB), of an input signal $s(t)$ with a duration of T, was calculated as in Equation 1:

$$I = 10 \log_{10} \frac{1}{T} \int_0^T 10^{\frac{s(t)}{10}} dt \tag{1}$$

We also examined several prosodic features related to pauses, such as the time spent speaking versus pausing, the average length, and how often pauses occurred. These features help us understand how AD impacts the natural flow of speech. People with AD often pause more frequently and for longer periods, which makes their speech less fluent and more fragmented [34, 35].

For articulation, we looked at features like speech rate, average syllable duration, and articulation rate. Speech rate was measured as the number of syllables spoken divided by the total length of the audio, including pauses. People with AD tend to speak more slowly, partly due to difficulty finding the right words. The articulation rate was calculated similarly but only during the actual speaking time, without counting pauses, giving us a clearer picture of how quickly words are produced when a person is actively speaking. Average syllable duration was found by dividing the total speaking time by the number of syllables spoken. In AD, this is often longer because of slower and less precise movements needed to produce speech.

We also analyzed how well individuals controlled their vocal tract by looking at the variability in the first two formants (F1 and F2). These formants reflect movements of the tongue and lips, and greater variability suggests difficulty in maintaining consistent speech articulation. Lastly, we measured the vocal tract length (VTL), which can provide insights into how well people can control their articulatory muscles. VTL has been used in voice research to identify speaker traits and verify speaker identity [36, 37]. Studies have found that changes in VTL can distinguish between people with AD and those without, as AD affects the ability to control the tongue and lips, leading to changes in voice resonance and quality [13, 14, 36]. To estimate VTL, we employed the approach based on the principles that vocal tract length and formant frequency dispersion correlate well with body size [36, 38]. More precisely, VTL for each recording was estimated (in cm) from the median of the first four formants, F_i, with the following formula:

$$VTL = \frac{\sum_{n=1}^{N} (2i-1) \frac{c}{4F_i}}{N} \tag{2}$$

where VTL is the vocal tract length, N is the total number of formants measured, and F_i is the median frequency in Hz of the formant i. The constant $c = 33500\ cm/s$ represents the speed of sound in a uniform tube with one end closed.

Linguistics Features. For linguistics features in the longitudinal study, we analyzed various syntactic and lexical features to understand the impact of AD on language production. These features cover aspects of lexical diversity and syntactic complexity,

which are crucial for identifying impairments in language abilities linked to AD. Previous research has demonstrated that linguistic features, such as reduced vocabulary diversity and simpler syntactic constructions, are significant indicators of cognitive decline in individuals with AD.

By examining these features—related to vocabulary richness, sentence structure, and grammatical complexity—our study aimed to characterize how AD affects both the expressive content and structural sophistication of language. Reduced lexical diversity and changes in syntactic complexity are commonly observed in individuals with AD, reflecting difficulties in word retrieval, sentence formulation, and organizing coherent speech. These linguistic features serve as essential markers for identifying and tracking cognitive decline, providing valuable insights for early detection and monitoring of AD.

To characterize the impact of AD on lexical diversity, we analyzed attributes related to word usage and vocabulary richness, such as type-token ratio (TTR), root type-token ratio (RTTR), and moving average TTR (MATTR) [39–42].

$$TTR = \frac{V}{N} \tag{3}$$

$$RTTR = \frac{V}{\sqrt{N}} \tag{4}$$

$$MATTR = \frac{1}{T - W + 1} \sum_{i=1}^{T-W+1} TTR_i \tag{5}$$

We calculated TTR and CTTR by using number of unique words (types) in the text (V), and total number of words (tokens) in the text (N). These measures provide insights into the diversity of words used by individuals with AD, which often diminishes as the disease progresses. We also considered features that capture overall lexical variability, such as Herdan's and Maas's lexical diversity measures, to quantify the richness of language over extended segments of speech [43, 44]. These features allow us to evaluate the extent of vocabulary repetition and the use of unique words, which are critical for understanding changes in expressive language capabilities in individuals with AD.

With respect to the impact of AD on syntactic complexity, we analyzed a range of features related to sentence structure and grammatical construction. These features include counts of noun phrases, verb phrases, and prepositional phrases per sentence, providing insights into the use and diversity of different types of phrases. Individuals with AD often use simpler grammatical constructions and shorter sentences, leading to fewer complex phrases compared to CN individuals. By quantifying these aspects, we aim to assess how the progression of AD affects the structural richness of language. In addition, we analyzed features related to clause usage, such as the number of subordinate clauses per sentence. Subordinate clauses contribute to sentence complexity and cohesion by expanding on the main idea, and a reduction in their use can indicate difficulties in maintaining complex thoughts. We also considered different types of connectives—including temporal, causal, additive, contrastive, and exemplifying connectives—to evaluate how well individuals with AD create logical connections between parts of their speech. Connectives are essential for creating well-formed and coherent discourse, and reduced use can reflect cognitive impairment that affects the

ability to produce connected and logical utterances. To further characterize syntactic complexity, we considered the average dependency length (ADL), which measures the average distance between syntactically related words within a sentence:

$$ADL = \frac{1}{N}\sum_{i=1}^{N}|position(w_i) - position(head(w_i))| \quad (6)$$

where w_i is the i-th word in the sentence; position(w_i) is the position index of the word w_i in the sentence; and head(w_i) is the syntactic head of the word w_i, i.e., the word to which w_i is syntactically related. This metric is based on the principles proposed by Gao and He [45], who demonstrated that dependency length is a robust indicator of syntactic complexity. Longer dependency lengths are typically associated with more complex and sophisticated sentence constructions, while shorter dependencies may reflect simpler, less cognitively demanding structures. Dependency length analysis allows us to quantify the cognitive load associated with sentence construction, which often diminishes as AD progresses, reflecting a move towards simpler sentence structures [14].

3.3.2 Analysis

Within Group Analysis. A primary objective of the study was to determine whether there were significant differences in feature values across different time intervals within both the AD and CN groups. To assess changes over time within each experimental group (AD or CN), the Wilcoxon test, a non-parametric test for repeated measures, was employed to evaluate the significance of each feature across the analyzed time intervals. Following the Wilcoxon test, the Benjamini-Hochberg procedure was used to identify which specific time intervals exhibited significant differences from one another. Given that the Wilcoxon test requires two paired samples of equal size, with a minimum of ten participants per group, the within-group analyses were restricted to a subset of 18 AD and 18 CN participants who had recordings available across both two-time intervals: -2 and -1. The summary statistics of the subset of data used in the within-group analyses are presented in Table 2.

	# of Speakers	# of Recordings (Total)	# of Recordings per Speaker	# of Segments per Speaker	Segment Length (s)
Time Interval -2 (from -10 to -6 years before YoD)					
AD	18	33	1/1.83/4	2/12.61/49	8.01/16.57/109.76
CN	18	34	1/1.89/2	2/12.61/49	8.01/18.24/134.13
Time Interval -1 (from -5 to -1 years before YoD)					
AD	18	40	1/2.22/5	2/12.61/49	8.01/16.00/84.45
CN	18	42	1/2.33/4	2/12.61/49	8.01/15.85/92.94

Table 2: The table presents statistics for the data used in within-group experiments. The reported statistics cover two-time intervals for both the AD and CN groups. When three values are provided in a field, they indicate the minimum, average, and maximum, respectively. The statistics include the number of speakers, total recordings, recordings per speaker, segments per speaker, and segment length in seconds. A recording is defined as the audio associated with a specific video, while a segment represents a portion of a recording that contains the target speaker's speech.

Between Group Analysis. The main aim of this study was to determine if significant differences in feature values existed across AD and CN groups within a certain time interval. To compare the differences between two independent groups, the Mann-Whitney U test, a non-parametric test for repeated measures, was used to assess the significance of each feature. Following this, the Benjamini-Hochberg procedure was used to identify which specific features showed significant differences from each other. The between-group analyses were limited to 18 AD and 18 CN participants who were matched and had recordings for both time intervals: -2 and -1. Table 3 presents the summary statistics of the subset of data used for these between-group analyses.

	# of Speakers	# of Recordings (Total)	# of Recordings per Speaker	# of Segments per Speaker	Segment Length (s)
Time Interval -2 (from -10 to -6 years before YoD)					
AD	18	35	1/1.94/4	3/17.00/52	8.01/16.05/109.76
CN	18	35	1/1.94/2	3/17.00/52	8.01/18.55/134.13
Time Interval -1 (from -5 to -1 years before YoD)					
AD	18	42	1/2.33/5	2/33.00/103	8.01/16.41/84.45
CN	18	43	1/2.39/4	2/33.00/103	8.01/15.22/92.94

Table 3: The table presents statistics for the data used in between-group experiments. The reported statistics cover two time intervals for both the AD and CN groups. When three values are provided in a field, they indicate the minimum, average, and maximum, respectively. The statistics include the number of speakers, total recordings, recordings per speaker, segments per speaker, and segment length in seconds. A recording is defined as the audio associated with a specific video, while a segment represents a portion of a recording that contains the target speaker's speech.

4 Experimental Setup

4.1 Data

ADCeleb The statistics for the ADCeleb dataset utilized in the classification experiments are presented in Table 1. Meanwhile, the statistics for the data subset employed in the within-group and between-group experiments are provided in Table 2 and 3.

4.2 Data Preprocessing

The recordings were initially processed by resampling to a $16\,\mathrm{kHz}$, essential for compatibility with the feature extraction algorithms described in subsections 4.3 and 4.4. This resampling was conducted using the SoX tool. In addition, Root Mean Square loudness normalization was applied to standardize the perceived loudness across recordings. The normalization procedure involved calculating the root mean square value of the signal to determine the perceived loudness, which was then used to derive the gain for adjustment. The normalized audio files were subsequently used to extract intensity-related features, as described in section 4.3. For the ADCeleb dataset, we included only those recordings with durations exceeding $8\,\mathrm{sec}$ in the experiments, as segments shorter than this threshold were more likely to contain interference from other speakers. A detailed discussion of the limitations associated with the diarization process is provided in section 6. Furthermore, throughout all experiments, the two groups were balanced regarding sample size, age, gender, and ethnicity to ensure comparability.

4.3 Interpretable Acoustic Representations

Table 2 and 3 summarizes the features included in the longitudinal analysis along with their Expected behavior (EBs). Each feature's EB is supported by prior longitudinal and/or cross-sectional studies. In the longitudinal analysis, we focused exclusively on features with a direct interpretive value and the potential to serve as biomarkers of AD in clinical settings. The comprehensive set of features used in the classification experiments is detailed in the Supplementary Materials. The following two paragraphs describe the methods for extracting prosodic and articulatory features.

Prosodic The prosodic feature extraction process involved several steps. F0 (std) and intensity (std) were extracted using Parselmouth (`https://parselmouth.readthedocs.io/en/stable/`), a Python library for Praat software. In particular, the following library was used: `https://github.com/uzaymacar/simple-speech-features`. To perform pitch analysis, an autocorrelation method was used. Other prosodic descriptors related to duration, F0, and energy were extracted using Disvoice (`https://github.com/jcvasquezc/DisVoice/tree/master/disvoice/prosody`), a Python library designed to extract phonological, prosodic, articulatory, and glottal features from speech. In each segment, 103 prosody features were calculated from six statistical functionals (mean, standard deviation, minimum, maximum, kurtosis, and skewness) of different prosody parameters. We also considered prosodic features related to pauses, such as speech time, total pause time, percentage pause time, mean pause duration, silence-to-speech ratio, and pause variability. In this respect, we used DigiPsychProsody (`https://github.com/NeuroLexDiagnostics/DigiPsych\_Prosody`), another Python library, to compute pause-based features. This library computes the features using the WebRTC Voice Activity Detector (`https://github.com/wiseman/py-webrtcvad`). $20\,\mathrm{ms}$ was used as frame length for voice activity detection. A total of 18 pause-related features were extracted.

Articulatory. To capture articulation alterations associated with AD, several speech features were analyzed. These included speech rate, average syllable duration, and articulation rate, which provide insights into the fluency and timing aspects of speech production often affected in AD [46]. In addition, formant-related features such as the first and second formant (F1 and F2) standard deviations and vocal tract measures were considered, as they indicate the changes in vowel production linked to neuromuscular decline [13]. These features were extracted using Parselmouth, which serves as a Python interface for Praat software—a standard tool for speech analysis [47]. The formants of each recording were extracted with a maximum formant value of $5.5\,\mathrm{kHz}$, a window length of $25\,\mathrm{ms}$, a time step of $6.25\,\mathrm{ms}$, and a pre-emphasis of $50\,\mathrm{kHz}$. Articulatory features such as formants are particularly relevant for assessing the motor aspects of speech in individuals with AD, as disruptions in motor control can lead to changes in vowel articulation and reduced formant variability, which can serve as early indicators of the disease [48].

4.4 Interpretable Linguistics Representations

The linguistic features examined in this study focus on lexical and syntactic aspects to assess how AD impacts language production. In the longitudinal analysis, we focused exclusively on features that possess direct interpretability and potential utility as biomarkers for AD in clinical settings. A comprehensive set of features used in the classification experiments can be found in the Supplementary Materials. The below two paragraphs describe the methods utilized for extracting the lexical and syntactic features.

Lexical. To analyze lexical features, we employed the LexicalRichness module from the LexicalRichness package to compute various lexical diversity metrics, including type-token ratio (TTR). TTR was utilized to assess vocabulary richness, providing an indication of the balance between unique words (types) and the total number of words (tokens) in a text. Individuals with AD often exhibit a lower TTR, indicating reduced word variety and word retrieval difficulties, which aligns with the findings of Forbes-McKay and Venneri [15] that demonstrated repetitive language patterns in individuals with AD.

Syntactic. The syntactic features analyzed included counts of nouns, verbs, and adjectives per sentence, as well as the number of VPs per sentence. For feature extraction, we utilized the spaCy natural language processing model ($en\_core\_web\_sm$) to parse sentences and extract key grammatical elements. These metrics quantify variations in grammatical usage and provide a detailed understanding of structural changes in language production. Individuals with AD tend to construct simpler grammatical structures, reflected in reduced use of diverse noun and verb phrases, highlighting a decline in syntactic complexity [33]. The use of temporal and causal connectives was also examined to evaluate coherence and logical flow within sentences. Reduced usage of such connectives indicates challenges in maintaining organized and coherent discourse—a hallmark of cognitive decline associated with AD. Moreover, we calculated the average dependency length (ADL) to quantify syntactic complexity. This measure was derived from syntactic parses provided by the spaCy model, with ADL reflecting the average distance between related words in a sentence. Longer dependency lengths typically suggest more sophisticated syntactic structures, whereas shorter dependency lengths, more common in AD patients, reflect simpler constructions that are cognitively less demanding [14, 45]. Combining these analyses of lexical and syntactic features, our study offers a comprehensive understanding of how AD impacts language, which can assist in early diagnosis and monitoring of cognitive decline.

4.5 Non-Interpretable Acoustic Representations

The next paragraphs provide a thorough explanation of each of the five distinct non-interpretable characteristics considered in this study.

X-vectors. This study utilized 512-dimensional x-vector embeddings derived from deep neural networks trained for speaker recognition tasks [49]. X-vectors have demonstrated

effectiveness in distinguishing speech patterns between individuals with AD and CN controls, capturing subtle prosodic and articulatory changes indicative of cognitive decline [13,49]. The embeddings were generated using a pre-trained x-vector model from the SpeechBrain toolkit [50], trained on VoxCeleb1 (English speech) and VoxCeleb2 (multilingual recordings) [51, 52]. These features are particularly relevant for analyzing AD-related vocal changes in pitch, articulation, and fluency.

TRILLsson. Paralinguistic features in this study were derived from TRILLsson, a distilled version of the 600-million-parameter CAP12 Conformer-based model trained via self-supervised learning [53]. CAP12, known for its ability to capture non-lexical speech elements like tone and emotion, has excelled in tasks such as speech emotion recognition and dysarthria classification. TRILLsson, created through knowledge distillation with CAP12 as the teacher model, was trained on datasets like Audioset and Libri-light to produce a compact and efficient model [54]. Using the TRILLsson1 model, 1,024-dimensional embeddings were extracted from $10\,\mathrm{sec}$ speech segments, with shorter segments zero-padded.

Wav2Vec 2.0. Wav2Vec 2.0 is a self-supervised model for automatic speech recognition that extracts meaningful representations from raw audio using a convolutional feature encoder and transformer layers. This structure captures both fine-grained and high-level speech characteristics, making it effective for tasks like speaker recognition and emotion detection [55]. In this study, we used the 95M parameter Wav2Vec 2.0 model from S3PRL, pre-trained on LibriSpeech (LS-960) but not fine-tuned for downstream tasks [56, 57]. Consistent with prior work, including Favaro et al. [32], We focused on the fourth layer's 768-dimensional embeddings, shown to perform well for related analyses [58]. Recordings were segmented into $10\,\mathrm{sec}$ intervals, with shorter segments zero-padded. Embeddings were extracted every $20\,\mathrm{m\,sec}$ and averaged within each segment, then across segments, to create robust representations for each recording.

HuBERT. Hidden-unit BERT (HuBERT) is a leading speech SSL model known for its effectiveness in ASR and versatility across tasks like speaker verification and diarization [56, 59]. In this study, we utilized a base HuBERT model from the S3PRL toolkit, extracting embeddings from the seventh layer, identified as optimal for analyzing Alzheimer's Disease-related speech features [56]. Audio recordings were segmented into $10\,\mathrm{sec}$ intervals, with shorter segments zero-padded. HuBERT generated embeddings every $20\,\mathrm{m\,sec}$, which were averaged to form a 768-dimensional feature vector per segment.

Whisper. This study utilized Whisper, a pre-trained ASR model designed to map speech to text across 60 languages, leveraging 680,000 hours of multilingual speech data [32, 60]. Its encoder-decoder transformer architecture processes 80-channel mel-spectrogram inputs using convolutional layers, sinusoidal positional encoding, and self-attention mechanisms, producing robust speech representations. We employed the Whisper tiny model from S3PRL, extracting 384-dimensional embeddings from its final encoder layer [56].

4.6 Non-Interpretable Linguistics Representations

The next paragraphs provide a thorough explanation of each of the eight distinct non-interpretable characteristics considered in this study.

XLM-RoBERTa. XLM-RoBERTa, a multilingual transformer model trained on 100 languages, was used to extract 768-dimensional embeddings from speech transcripts [61]. By capturing lexical and syntactic features, these embeddings identified linguistic markers like reduced vocabulary richness and simpler syntactic structures indicative of AD [13, 15].

text2vec. The text2vec-base-multilingual model, based on the CoSENT architecture, provided 384-dimensional semantic embeddings optimized for multilingual data [32, 62]. These embeddings captured linguistic nuances, such as vocabulary diversity and sentence complexity, aiding in the analysis of cognitive-linguistic decline in AD [13, 18].

Multilingual-E5-large. The multilingual-e5-large model, a transformer-based architecture trained through contrastive pre-training and supervised fine-tuning on multilingual datasets, was employed to extract linguistic embeddings [28]. This model generates 1024-dimensional embeddings that effectively capture semantic and syntactic features across languages. We processed speech transcripts by cleaning irrelevant text, tokenizing, and applying the model to produce embeddings. Averaging and normalization techniques were used to create consistent representations, enabling the detection of lexical diversity and sentence complexity changes linked to AD [13].

LaBSE. LaBSE (Language-agnostic BERT Sentence Embedding) is a dual-encoder model trained on over 100 languages, producing 768-dimensional embeddings optimized for semantic and syntactic analysis [63]. Using preprocessed transcripts, LaBSE embeddings captured linguistic markers such as lexical diversity and semantic coherence. These features provided valuable insights into language changes associated with cognitive decline, supporting early AD detection through non-invasive linguistic analysis [8, 13].

distilbert-base-multilingual-cased. The distilbert-base-multilingual-cased model is a lightweight version of BERT, trained on Wikipedia data in 104 languages. With 6 transformer layers, 768 dimensions, and 134 million parameters, it is approximately twice as fast as mBERT while retaining strong multilingual capabilities [64]. In this study, 768-dimensional embeddings were extracted to analyze syntactic and semantic changes linked to AD, focusing on lexical diversity, sentence complexity, and semantic coherence [8, 13].

distiluse-base-multilingual-cased-v1. Derived from the Universal Sentence Encoder (USE), this model generates 512-dimensional embeddings optimized for cross-lingual semantic tasks [62]. Its efficient architecture balances speed and contextual representation, making it suitable for analyzing semantic coherence and syntactic structure related to cognitive decline in AD.

cross-en-de-roberta-sentence-transformer. The cross-en-de-roberta, part of the SentenceTransformers framework, is a fine-tuned version of RoBERTa optimized for generating 768-dimensional cross-lingual embeddings in English and German. Built on RoBERTa's 12-layer transformer architecture, it uses self-attention to capture deep contextual relationships within and across languages, essential for tasks requiring semantic alignment [65]. The model outputs embeddings from the last hidden state, which are summarized through mean pooling into a single vector per transcript, effectively capturing both syntactic and semantic nuances [62].

bert-based-multiligual-cased This model, developed by Google, is a multilingual version of BERT pre-trained on over 100 languages. Its architecture includes 12 transformer layers and generates 768-dimensional embeddings, effectively capturing syntactic and semantic relationships across languages [66]. For this study, we cleaned and tokenized transcripts using BERT's cased tokenizer to preserve the original casing, then passed the text through the model to extract pooler output embeddings. These 768-dimensional embeddings summarize the semantic and syntactic features of the text, highlighting markers such as lexical diversity and sentence complexity, which are indicative of cognitive decline in AD [13, 66].

4.7 Classifiers and Evaluation Metrics

As outlined in section 4, our experiments utilized a NCV scheme, a standard method commonly adopted for feature selection and parameter tuning to enhance classification accuracy and prevent an overfitting process; we ensured that the training and testing subsets did not contain the same speakers. This separation is essential, as including the same speakers across these sets could lead models to learn speaker-specific characteristics, resulting in overly optimistic and biased predictions. For feature selection in the inner folds of the NCV, we used an Extremely Randomized Trees Classifier, which allowed us to refine the feature set using only the training data. To classify AD speech patterns, we employed interpretable classifiers, including Support Vector Machine (SVM), Random Forest (RF), Multilayer Perceptron (MLP), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), and Bagging (BG). For classifiers with non-interpretable features, we used Probabilistic Linear Discriminant Analysis (PLDA) following Principal Components Analysis (PCA) for dimensionality reduction . The PCA transformation matrix on the training embeddings was subsequently applied to transform training, validation, and testing embeddings in each iteration. After reducing feature dimensionality with PCA (with tuned hyperparameters), we trained the PLDA model using the PCA-reduced training subset. Enrollment embeddings were created based on the mean of PCA-reduced AD and CN training embeddings, providing a basis for scoring. A log-likelihood ratio was calculated relative to the enrollment embeddings and compared against the equal error rate (EER) threshold for each PCA-reduced testing or validation embedding. If this ratio exceeded the EER threshold, the sample was classified as AD; otherwise, it was classified as CN.

To improve the accuracy of AD detection, a fusion-based prediction framework was designed by integrating outputs from models trained on diverse feature sets. Specifically,

five feature groups were considered: interpretable acoustic features, non-interpretable acoustic features, linguistic features, interpretable linguistic features, and a comprehensive combination of all features. Separate models were trained for each group, and the top three models within each category were selected based on accuracy metrics. For each speaker, the binary predictions (0 or 1) generated by these selected models were averaged to produce a continuous prediction score. A threshold of 0.5 was then applied to classify speakers, with scores equal to or above the threshold indicating AD (1) and scores below the threshold categorized as CN (0). This fusion methodology leverages the predictive strengths and complementarities of multiple models across varied feature domains, enabling more robust and generalizable detection of AD. The fused model's performance was compared against baseline models, demonstrating its efficacy in enhancing detection accuracy. The fusion model reflects an effort to bridge the gaps in existing single-domain approaches by integrating acoustic and linguistic feature sets to improve the classification accuracy and generalizability of AD models. Inspired by findings that acoustic and linguistic markers provide complementary information about cognitive and motor impairments in AD, the fusion model is designed to exploit the synergies across these domains. Prior studies in dementia research have shown the idea and opportunities of applying feature fusion in enhancing predictive performance, particularly in noisy or incomplete datasets [20, 67, 68]. This multimodal approach not only improves the robustness of the model but also offers a deeper understanding of the interplay between speech production and cognitive decline, serving as a foundation for the development of advanced, real-world AD diagnostic tools.

The performance of our models was primarily evaluated using metrics such as accuracy (ACC), F1-score (F1), specificity (SPE), sensitivity (SEN), and the area under the ROC curve (AUC). These metrics provide a comprehensive assessment of the classifier's ability to detect AD-specific linguistic features, crucial for reliable diagnostic support.

4.8 Statistical Analysis

The normality of the extracted linguistic features was evaluated using the Shapiro-Wilk test [69], implemented via `scipy.stats.shapiro`. This test examines the null hypothesis that a dataset follows a normal distribution, allowing us to determine the distribution properties of features related to AD analysis. Since not all feature distributions met the normality assumption, we used the Wilcoxon Signed Ranks test—a non-parametric alternative suitable for analyzing within-group changes over time without assuming normal distribution. This test was executed with `scipy.stats.wilcoxon`. The Wilcoxon Signed Ranks test requires paired observations without normal distribution assumptions. To meet this requirement, we both balanced the data between AD and CN groups and paired the segment's data according to the speakers' names in different time intervals. For posthoc analysis, the False Discovery Rate (FDR) correction was employed to manage the inflated risk of Type I errors (false positives) that can arise from multiple comparisons, implemented using `statsmodels.stats.multitest`. This combination of statistical tests allowed us to monitor and interpret subtle within-group changes in language use, which may signal early cognitive decline associated with AD. For between group experiment, we used the Mann-Whitney U test (also known as the Wilcoxon rank-sum test), a non-parametric statistical test used to compare differences between

two independent groups. This test was executed with `scipy.stats.mannwhitneyu`. We also balance the data between AD and CN groups in this experiment with FDR correction employed for post-hoc analysis. This combination of statistical tests allowed us to monitor and interpret subtle between-group changes in language use, which may signal early cognitive decline associated with AD.

5 Results and Discussion

The results from the classification experiments and longitudinal analysis are reported and commented on in the following two subsections.

5.1 Classification Experiments

The classification outcomes for IFMs and NIFMs are shown in TablesTable 4, 5, 6 and 7, respectively, with results calculated a per-speaker basis. Besides, the outcomes for fusion-based feature models is shown in Table 8. To obtain speaker-level results for IFMs, feature values were averaged across each speaker's data. For NIFMs, speaker-level embeddings were calculated by applying length normalization to each embedding, averaging these normalized embeddings per speaker, and applying a second round of length normalization on the averaged result. For model-specific layer details with acoustic features, we focused on embeddings derived from the 4th layer of Wav2Vec 2.0 and the 7^{th} l layer of HuBERT, as these layers have been shown to provide effective representations for the features being studied. For linguistics features, we specifically focused on embeddings derived from the pooler output refer to the final hidden state of bert-base-multilingual-cased, since it provides a more condensed, sentence-level representation. In the following subsections, we examine the results of IFMs and NIFMs across different experimental setups which allow for an understanding of how these models perform across varying dataset configurations.

5.2 Acoustic Modelling

Table 4 and 5 present the outcomes of the monocorpus experiments, highlighting the performance of IFMs and NIFMs, respectively. For IFMs, the highest accuracy scores were 0.59 and 0.71 in time intervals -2 and -1, respectively. In contrast, NIFMs achieved peak accuracy values of 0.67 and 0.73 for these same intervals. As anticipated, both models yielded improved classification results in time interval -1, where accuracy was notably higher than in time interval -2. This trend aligns with expectations, as language and neurological impairments are more pronounced in later stages of AD, making it easier to distinguish AD from CN individuals. Interestingly, while both IFMs and NIFMs demonstrated the capacity to separate CN and AD groups in pre-diagnosis speech data, NIFMs outperformed IFMs when using pre-diagnosis data. This finding implies that NIFMs may capture subtle indicators of dysarthria even in the early or prodromal phase of the disease. The superior accuracy of NIFMs in detecting these early linguistic changes suggests that NIFMs are particularly adept at identifying complex and nuanced

speech patterns that traditional IFMs might overlook, underscoring their potential value in early-stage detection.

Model	F1	Acc	AUC	Sens	Spec
Time interval -2					
SVM	**0.59**	**0.59**	**0.56**	0.62	0.55
RF	0.50	0.50	0.40	0.45	0.55
GB	0.48	0.48	0.52	0.55	0.41
BG	0.38	0.38	0.31	0.31	0.45
MLP	0.50	0.50	0.41	0.48	0.52
XGB	0.45	0.45	0.39	0.52	0.38
Time interval -1					
SVM	0.60	0.60	0.65	0.66	0.55
RF	**0.71**	**0.71**	0.76	0.76	0.66
GB	0.69	0.69	**0.79**	0.66	0.72
BG	0.59	0.60	0.65	0.76	0.45
MLP	0.67	0.67	0.78	0.66	0.69
XGB	**0.71**	**0.71**	0.77	0.76	0.66

Table 4: The per-speaker results for Interpretable Feature Models (IFMs) are presented with acoustics features, organized into the two designated time intervals. For each type of feature, performance metrics include F1-score, accuracy (ACC), Area Under the Curve (AUC), sensitivity (SENS), and specificity (SPEC). In each row, the highest F1 and AUC values are highlighted in bold to mark top performance.

Feature Name	F1	Acc	AUC	Sens	Spec
Time interval -2					
xvector	0.53	0.53	0.53	0.57	0.50
trillsson	0.52	0.52	0.56	0.57	0.47
hubert	0.50	0.50	0.52	0.53	0.47
wav2vec	**0.67**	**0.67**	**0.68**	0.70	0.63
whisper	0.53	0.53	0.56	0.50	0.57
Time interval -1					
xvector	0.60	0.60	0.64	0.53	0.67
trillsson	0.56	0.57	0.68	0.47	0.67
hubert	**0.73**	**0.73**	**0.76**	0.80	0.67
wav2vec	0.70	0.70	0.75	0.73	0.67
whisper	**0.73**	**0.73**	**0.76**	0.73	0.73

Table 5: The per-speaker results for Non-Interpretable Feature Models (NIFMs) are presented with acoustics features, organized into the two designated time intervals. For each type of feature, performance metrics include F1-score, accuracy (ACC), area under the curve (AUC), sensitivity (SENS), and specificity (SPEC). In each row, the highest F1 and AUC values are highlighted in bold to mark top performance.

5.3 Linguistic Modelling

Tables 6 and 7 present the outcomes of the monocorpus experiments, highlighting the performance of (IFMs) and Non-Interpretable Feature Models (NIFMs), respectively. For IFMs, the best accuracy values achieved were 0.62 and 0.63 in time intervals -2 and -1, respectively, whereas NIFMs reached higher accuracy peaks of 0.73 and 0.75 for the same intervals. Although the performance improvement from interval -2 to -1 was modest for both models, this pattern aligns with expectations, as linguistic impairments tend to intensify in the later stages of AD, making it easier to differentiate AD from CN individuals. Notably, while both IFMs and NIFMs demonstrated an ability to separate AD and CN groups in post-diagnosis data, NIFMs consistently outperformed IFMs across both intervals. In time interval -2, for instance, IFMs reached a maximum F1 and ACC of only 0.62, whereas NIFMs achieved a significantly higher peak of 0.73. This substantial improvement suggests that NIFMs are more effective at detecting subtle linguistic shifts indicative of early cognitive decline, even in the prodromal AD stage. The enhanced performance of NIFMs in identifying these initial linguistic alterations implies that they are particularly skilled at capturing nuanced language patterns that may go unnoticed by traditional IFMs, highlighting their potential value for early intervention and diagnosis.

Model	F1	Acc	AUC	Sens	Spec
Time Interval -2					
SVC	0.48	0.5	0.57	0.31	0.69
RF	0.59	0.59	0.59	0.63	0.56
GB	**0.62**	**0.62**	**0.72**	0.69	0.56
BG	0.53	0.53	0.50	0.50	0.56
MLP	0.59	0.59	0.62	0.56	0.62
XGB	0.50	0.50	0.59	0.50	0.50
Time Interval -1					
SVC	0.48	0.52	0.58	0.27	0.77
RF	0.58	0.58	0.64	0.57	0.60
GB	**0.63**	**0.63**	**0.68**	0.63	0.63
BG	0.51	0.52	0.58	0.60	0.43
MLP	0.55	0.55	0.62	0.57	0.53
XGB	0.57	0.57	0.57	0.57	0.57

Table 6: The per-speaker results for Interpretable Feature Models (IFMs) are presented with linguistics features, organized into the two designated time intervals. For each type of feature, performance metrics include F1-score, accuracy (ACC), area under the curve (AUC), sensitivity (SENS), and specificity (SPEC). In each time interval, the highest F1, ACC, and AUC values are highlighted in bold to mark top performance.

Feature Name	F1	Acc	AUC	Sens	Spec
Time Interval -2					
bert-based-multilingual-cased	0.6	0.6	0.71	0.53	0.67
cross-en-de-roberta-sentence-transformer	0.7	0.7	0.71	0.67	0.73
distilbert-base-multilingual-cased	0.63	0.63	0.67	0.63	0.63
distiluse-base-multilingual-cased-v1	**0.73**	**0.73**	0.73	0.77	0.7
Multilingual-e5-large	**0.73**	**0.73**	**0.76**	0.7	0.77
LaBSE	0.7	0.7	0.72	0.7	0.7
text2vec	0.68	0.68	0.71	0.7	0.67
xlm-RoBERTa	0.68	0.68	0.66	0.73	0.63
Time Interval -1					
bert-based-multilingual-cased	**0.75**	**0.75**	**0.74**	0.73	0.77
cross-en-de-roberta-sentence-transformer	0.7	0.7	0.73	0.73	0.67
distilbert-base-multilingual-cased	0.67	0.67	0.67	0.7	0.63
distiluse-base-multilingual-cased-v1	0.68	0.68	0.74	0.63	0.73
Multilingual-e5-large	0.67	0.67	0.74	0.67	0.67
LaBSE	0.63	0.63	0.67	0.67	0.6
text2vec	0.65	0.65	0.65	0.7	0.6
xlm-RoBERTa	0.58	0.58	0.63	0.5	0.67

Table 7: The per-speaker results for Non-Interpretable Feature Models (NIFMs) are presented with linguistics features, organized into the two designated time intervals. For each type of feature, performance metrics include F1-score, accuracy (ACC), area under the curve (AUC), sensitivity (SENS), and specificity (SPEC). In each time interval, the highest F1, Acc and AUC values are highlighted in bold to mark top performance.

5.4 Fusion-Based Modelling

Tables 8 summarize the results of the fusion-based models combining different feature sets. For time interval -2, the highest accuracy was 0.70, achieved by the Linguistics-only fusion, which integrated predictions from the top three models based on interpretable and non-interpretable linguistic features. In contrast, for time interval -1, the best performance was observed in the Non-Interpretable-only fusion, achieving an accuracy of 0.80 by combining predictions from non-interpretable linguistic and acoustic features. When compared to the best accuracy values reported for individual feature sets in earlier experiments, the fusion approach demonstrated notable changes. At time interval -1, accuracy improved significantly from 0.75 to 0.80, indicating enhanced model stability and performance through feature fusion. However, at time interval -2, the highest accuracy declined from 0.73 to 0.70, suggesting that fusion may not consistently improve performance in earlier time intervals where feature variability is higher. These findings highlight the potential of fusion-based modeling in leveraging complementary information from diverse feature sets to enhance classification performance, particularly in time intervals closer to the year of diagnosis. The results underscore the robustness of fusion models at later stages and their capacity to outperform single-feature models under certain conditions.

Feature Group	F1 Score	Accuracy	AUC	Sensitivity	Specificity
Time Interval -2					
Acoustics Only	0.58	0.5	0.51	0.7	0.3
Linguistics Only	**0.73**	**0.70**	**0.75**	0.81	0.59
Interpretable Only	0.61	0.56	0.54	0.69	0.44
Non-Interpretable Only	0.70	0.68	**0.77**	0.77	0.6
All Features Combined	0.65	0.63	0.70	0.69	0.56
Time Interval -1					
Acoustics Only	0.65	0.6	0.67	0.73	0.47
Linguistics Only	0.67	0.67	0.74	0.67	0.67
Interpretable Only	0.59	0.53	0.53	0.67	0.4
Non-Interpretable Only	**0.81**	**0.8**	**0.86**	0.9	0.7
All Features Combined	0.76	0.75	0.78	0.80	0.70

Table 8: The per-speaker results for fusion-based features are presented with top three features with highest ACC in each subgroup, organized into the two designated time intervals. For each type of feature, performance metrics include F1-score, accuracy (ACC), area under the curve (AUC), sensitivity (SENS), and specificity (SPEC). In each time interval, the highest F1, ACC, and AUC values are highlighted in bold to mark top performance.

5.5 Discussion

When examining the results from NIFMs, which consistently outperform IFMs, a distinct pattern emerges between acoustic and linguistic features. In the acoustic feature experiments, there is a significant improvement of approximately 6% from time interval -2 to time interval -1. However, in the linguistic feature experiments, the improvement is comparatively modest, at only around 2%, since peak performance could reach 0.73 even in time interval -2. The observed differences in improvement and stability between acoustic and linguistic features in detecting AD reflect how these feature types capture different aspects of cognitive decline. Acoustic features, such as pitch variation, intensity, and speech rate, often correlate with motor and speech coordination issues that emerge as dysarthria becomes more pronounced in later stages of AD [8, 33]. This progression aligns with the notable increase in accuracy and F1 scores from time interval -2 to time interval -1, as acoustic impairments grow more detectable over time. However, because these features depend largely on involuntary motor processes, they may not capture the more subtle cognitive signs present in earlier stages, hence the lower performance at time interval -2. In contrast, linguistic features reflect higher-order cognitive functions, such as lexical retrieval, syntactic complexity, and coherence, which are affected by cognitive decline even before motor symptoms manifest [13, 15]. The high and stable performance of linguistic features, achieving 0.73 at time interval -2 and increasing only slightly to 0.75 at time interval -1, suggests that these features are sensitive to early cognitive changes and can detect subtle language alterations that appear in the prodromal stages of AD. Although both acoustic and linguistic models show similar results by time

interval -1, linguistic models outperform acoustic models at time interval -2, suggesting that linguistic markers may capture early cognitive decline more effectively [32, 70]. This combination of findings indicates that linguistic features may serve as reliable indicators of early-stage cognitive impairment, while acoustic features are more effective in capturing the pronounced symptoms associated with later-stage AD. Together, these observations underscore the complementary roles of linguistic and acoustic features in tracking the progression of AD, with linguistic markers providing early detection potential and acoustic markers reflecting later-stage severity [13, 33].

It is also noteworthy that IFMs in the acoustic experiment achieve a more substantial improvement in F1 and accuracy scores, rising from 0.59 to 0.71 from time interval -2 to time interval -1, compared to the more modest increase from 0.62 to 0.63 observed in the linguistic IFMs. This indicates that, while NIFMs in the linguistic experiment outperform those in the acoustic domain, IFMs yield consistently better results in the acoustic experiment than in the linguistic one, particularly at time interval -1. This discrepancy is likely influenced by the difference in the number of interpretable features used in each experiment. In the acoustic analysis, 161 features were used for classification, of which 103 were prosodic. These prosodic features capture a wide array of acoustic patterns, enabling a more comprehensive analysis of speech characteristics affected by AD [8]. In contrast, only 29 interpretable features were available for the linguistic experiment, limiting the model's capacity to capture a broad spectrum of language-related changes associated with cognitive decline. The greater number of acoustic features likely contributes to the higher overall performance of the IFMs in the acoustic domain. Interestingly, however, the linguistic IFMs still outperform their acoustic counterparts in time interval -2, despite the smaller feature set. This aligns with previous findings that linguistic markers are particularly sensitive to early cognitive changes, capturing subtle alterations in language use even with a limited number of features [13, 15]. This suggests that while the number of features contributes to model performance, linguistic features may inherently provide valuable early indicators of cognitive impairment, as we discussed in the preceding paragraph.

The analysis of fusion-based models highlights the strength of integrating diverse feature sets to enhance the detection of AD. By combining acoustic and linguistic features, these models capitalize on the unique strengths of each domain, offering a comprehensive approach to capturing the multifaceted changes associated with the disease. Notably, the best-performing fusion model achieves an accuracy of 0.80 at time interval -1, a significant improvement over single-feature models. This underscores the value of combining linguistic and acoustic cues to detect the more pronounced and multimodal manifestations of AD closer to the YoD. The enhanced performance at time interval -1 reflects the ability of fusion models to synergistically incorporate motor-related impairments, such as changes in pitch, intensity, and speech rate, alongside higher-order cognitive impairments, such as reduced lexical diversity and simpler syntax. Acoustic features effectively capture later-stage motor and prosodic impairments, while linguistic features remain sensitive to earlier cognitive declines [13, 33]. This dual capability allows fusion models to provide robust classification during advanced stages of AD progression. However, at time interval -2, linguistic-only models slightly outperform fusion models, suggesting that linguistic features alone are better suited to identifying the subtle, early signs of cognitive decline. The inclusion of acoustic features, while

valuable in later stages, may introduce noise at this stage when motor impairments are less apparent [13, 15]. Despite this, the consistent performance of fusion models across both time intervals demonstrates their adaptability and highlights the advantage of combining complementary feature sets for robust and balanced detection.

Ultimately, the fusion-based approach underscores the importance of multimodal analysis in AD research. By leveraging the early sensitivity of linguistic features and the specificity of acoustic markers in later stages, these models not only enhance detection accuracy but also offer a more holistic understanding of the disease's progression. This integrative framework holds promise for improving early diagnosis and tracking AD, paving the way for more targeted and effective interventions [8, 32].

5.6 Longitudinal Analysis

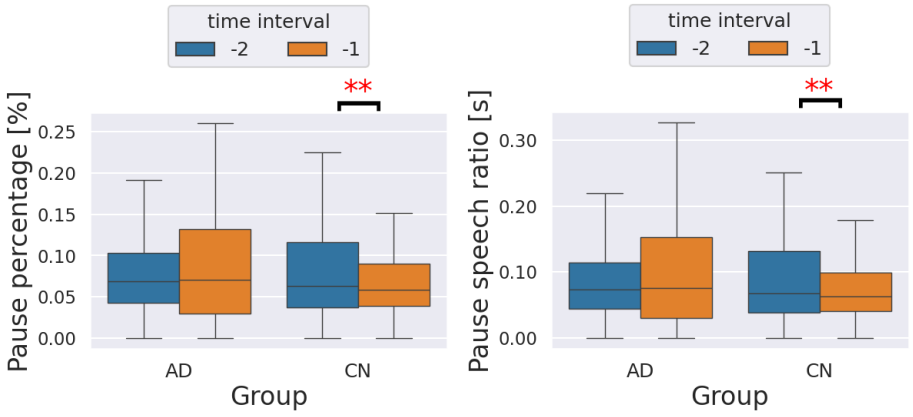
This study focused on tracking the progression of multiple speech measures over the prodromal period of AD. By comparing speech performance in individuals with AD to a CN group matched for age, sex, and nationality, we could distinguish changes associated with AD from those potentially due to normal aging. In the following sections, we will conduct a detailed analysis of both within-group and between-group experiments, examining linguistic and acoustic features independently.

5.6.1 Acoustic

Within-group The objective of this study was to investigate the trajectory of various speech measures from the prodromal phase of AD. Speech performance in individuals with AD was evaluated in comparison to a control group matched by age, sex, and nationality, enabling the identification of changes specific to AD as well as those potentially associated with normal aging.

Feature Name	**Group**	**p-value Adjusted**	**OB**
Pause percentage [%]	CN	0.004	$\downarrow$
Pause speech ratio [s]	CN	0.004	$\downarrow$
Speech rate	CN	0.003	$\downarrow$
Articulation rate	CN	< 0.001	$\downarrow$
Average syllable duration [(sec)]	CN	< 0.001	$\uparrow$
Intensity (std) [(dB)]	AD	< 0.001	$\downarrow$
Pitch (std) [Hz]	AD	0.01	$\uparrow$

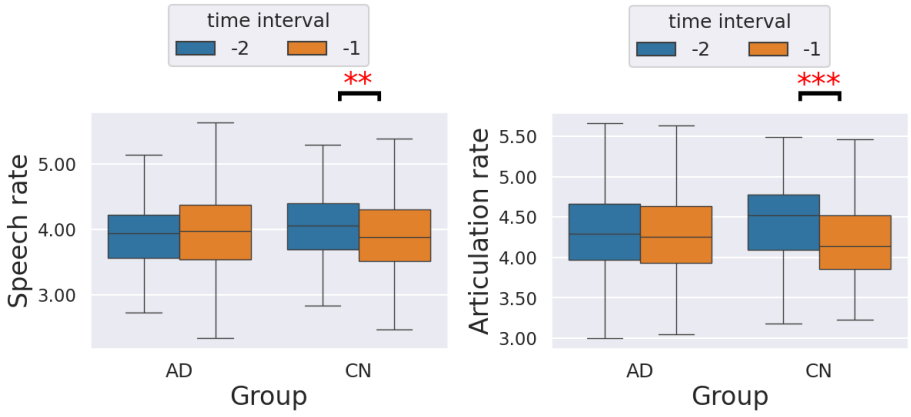
Table 9: The results of the within-group analysis for acoustics features are presented, where statistical significance was determined using Wilcoxon Signed Ranks test, with False Discovery Rate (FDR) correction for pairwise comparisons. For each feature showing significant differences, the table provides the relevant experimental group, p-value, and observed behavior (OB). Abbreviations: OB, Observed Behavior.



(a) *(b)*

Figure 4: Boxplots of pause percentage [%] (a) and pause speech ratio [s] (b) for the AD group and the CN group in the two time intervals considered in the within-group analysis.

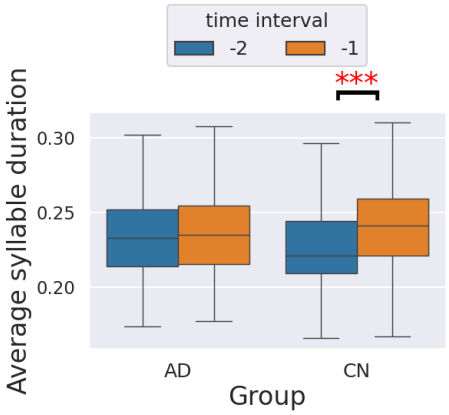
In AD group, the lack of significant change in pause percentage between time intervals -2 and -1 may suggest that the motor aspects of speech associated with pause frequency and duration are not as severely affected in the prodromal phase of AD. This aligns with findings that cognitive impairments in AD tend to primarily impact higher-level linguistic and semantic processing rather than motor functions [13, 32]. The slight upward trend in pause percentage, though not statistically significant, could reflect subtle early cognitive challenges affecting speech fluency, potentially linked to increased cognitive load in organizing and retrieving information [15]. On the other hand, the significant decrease in pause percentage observed in the CN group ($p = 0.004$) over the same period might reflect a typical adaptation pattern in normal aging. Healthy older adults often develop strategies to maintain or even improve their fluency, possibly by becoming more efficient in verbal expression or by adapting to cognitive changes with experience [71]. This contrast between AD and CN groups may suggest that while CNs can adjust their speech patterns over time, AD individuals begin to show early indications of cognitive decline that might later manifest as more pronounced linguistic impairments. Moreover, pause percentage and pause-speech ratio displayed comparable statistical results and trends across time intervals. Both features exhibited a non-significant upward trend in the AD group from 10 to 5 years before the YoD, while the CN group showed a statistically significant decrease in both measures ($p = 0.004$). This pattern aligns with previous studies suggesting that motor aspects of speech. However, the slight upward trend in AD pause measures could indicate emerging cognitive challenges, where the ability to organize fluent speech is subtly compromised. In summary, the significant decrease in pause percentage in the CN group versus the non significant upward trend in the AD group highlights a divergence in aging patterns. While controls may demonstrate compensatory adaptations, AD patients seem to follow an early trajectory of subtle cognitive-linguistic decline, albeit without marked motor impairment affecting pauses.



(a) *(b)*

Figure 5: Boxplots of speech rate (a) and articulation rate (b) for AD and CN groups in the two time intervals considered in the within-group analysis.

Moreover, the speech rate remains relatively stable between the two intervals, with minimal change in the median and the spread of values. In contrast, the CN group shows a significant decrease in speech rate over time ($p = 0.003$), possibly as a natural, age-related adjustment to maintain communicative effectiveness. This age-related slowing in healthy individuals may be a compensatory strategy, whereas the AD group's decline likely reflects an underlying pathological change rather than an adaptive one. The AD group's speech rate at time interval -2 already appears slightly lower than the CN group's speech rate at time interval -1 (see boxplots 5) . This could suggest that individuals with AD may experience an early decline in speech rate, which then stabilizes over time as the disease progresses. Similarly, the boxplots for articulation rate (shown above) demonstrate a trend that closely mirrors the patterns observed in speech rate, where both AD and CN groups experience a significant decline over time ($p < 0.001$), with a more pronounced decrease in the CN group. Notably, the articulation rate in the AD group appears low as early as time interval -2, while in the CN group, the significant decrease mainly occurs from time interval -2 to -1. This observation aligns with the hypothesis that cognitive-linguistic decline in AD can begin subtly even before formal diagnosis. The initial decline in speech and articulation rates may occur as early cognitive processes—such as processing speed, memory retrieval, and linguistic planning—become compromised [13] [72]. Once this initial decline occurs, it is possible that the speech rate and articulation rate stabilize, reflecting a *floor effect* where the disease does not cause further reductions in speech rate but instead impacts more complex linguistic and cognitive abilities as it progresses. Thus, the pattern observed in the AD group -an already reduced and stable speech rate—suggests that speech rate and articulation rate might be an early indicator of cognitive decline in AD. This early decrease might decrease as the disease progresses, with further cognitive-linguistic impairments manifesting in other areas. The significant reduction in speech rate for the CN group, however, is more characteristic of typical aging and reflects adaptive changes rather than pathological decline.



(a)

Figure 6: Boxplot of syllable duration for AD and CN groups in the two time intervals considered in the within-group analysis.

The boxplots for average syllable duration reveal an increase over time for both the AD and CN groups experience. However, the AD group shows a relatively high average syllable duration at time interval -2, which remains stable and slightly increases by time interval -1. This suggests that individuals with AD may exhibit subtle motor speech impairments early on, even before reaching the significant levels seen in the CN group at time point -1. Previous studies indicate that AD often involves subtle motor control deficits that can slow speech production, leading to longer syllable durations [13]. The relatively high starting level of syllable duration in AD compared to CN may reflect these early deficits, which are indicative of cognitive-motor impairments that are typical in early AD stages. For the CN group, there is a statistically significant increase in syllable duration from time interval -2 to time interval -1. This increase could reflect age-related changes in motor speech control, where syllable duration tends to increase as a natural consequence of aging [71]. Healthy older adults may exhibit slower articulation as part of an adaptive strategy, compensating for reduced motor flexibility or processing speed by speaking more slowly to maintain clarity. The significant increase in the CN group suggests that, unlike in AD, this change is part of a gradual, age-related adjustment rather than a pathological process. The lack of significant increase in the AD group could imply that their average syllable duration has already plateaued due to early cognitive-motor impairments. While the AD group's syllable duration is slightly higher than the CN group's initial duration, it does not increase as substantially over time. This could be due to a *ceiling effect*, where AD patients already exhibit speech-motor deficits that do not progress in the same way as age-related changes seen in healthy controls. This stability in AD may indicate that the disease affects speech motor patterns early, with less room for further decline in this particular feature.

It is worth noting that the lack of obvious change in the AD group between time intervals -2 and -1, despite the boxplot showing a generally worse baseline situation of the three features above (Speech rate, articulation rate, and average syllable durations) in -2, may be influenced by the segmentation and data exclusion criteria applied in our

analysis. Given that AD speech is often characterized by slower articulation rates, longer pauses, and extended syllable durations, the segmentation process may disproportionately impact the AD group. For instance, WhisperX relies on pauses to segment speech, and in cases where AD speech is particularly slow or drawn-out, this segmentation can split even a single sentence into multiple smaller segments. With a minimum segment length threshold set at 8 seconds, many of these shorter segments are excluded from our analysis, effectively filtering out the slowest segments from the AD group. This data loss could mean that the most characteristic speech patterns of the AD group—such as elongated syllable duration and reduced articulation rate—are underrepresented in our dataset. As a result, the remaining data may reflect a slightly faster subset of AD speech, making it appear that there is less change over time. Furthermore, since AD symptoms often progress gradually in the early stages, the lack of significant change between time intervals -2 and -1 could also be due to an early ceiling effect. This refers to the possibility that the most salient motor-linguistic impairments in AD, such as slowed speech rate, may already be present at -2 and reach a plateau, resulting in minimal further measurable decline by -1. Together, these factors—segmentation-related data loss and the potential for early onset and stabilization of speech impairments—contribute to the appearance of a stable trend in the AD group over time. This highlights the importance of selecting appropriate segmentation parameters and the need to interpret results in the context of these methodological limitations.

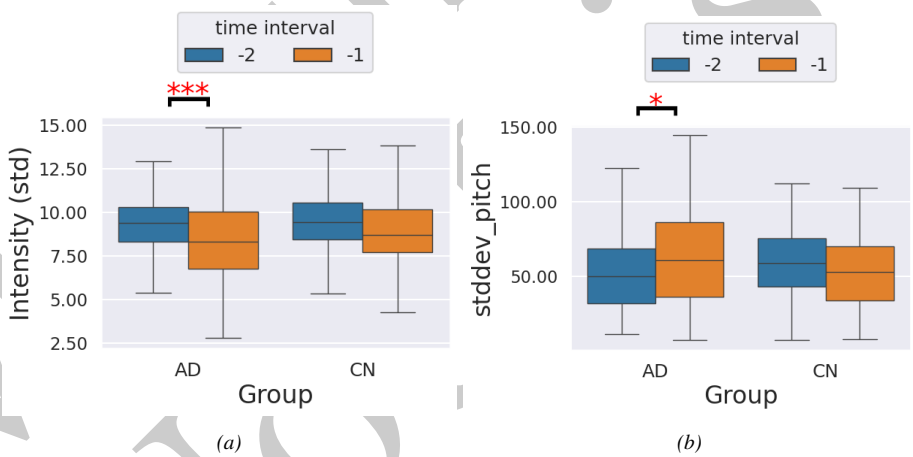


Figure 7: Boxplot of Intensity std (a) and stddev pitch (b) for AD and CN groups in the two time intervals considered in the within-group analysis.

The significant decrease in the standard deviation of intensity for the AD group ($p < 0.001$), along with the absence of a significant change for the CN group, suggests that individuals with AD experience a notable reduction in the variability of their speech intensity over time. The decline in intensity variability in the AD group could indicate a reduced ability to control vocal emphasis and expressiveness, as intensity modulation often involves cognitive and motor functions. As cognitive decline progresses in AD, individuals may struggle with both the planning and execution of speech, leading to a more monotonous vocal pattern [13]. A reduction in intensity standard deviation

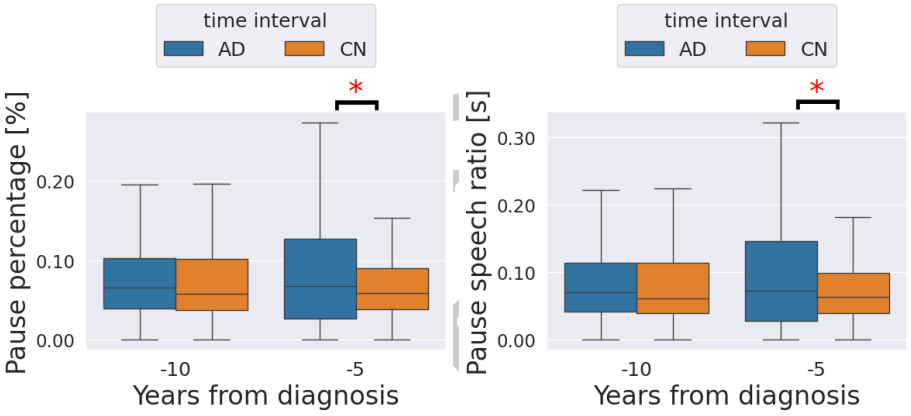
suggests that as AD progresses, individuals may lose this control, leading to a flatter, more monotonous speech pattern, which has been identified in previous studies as a characteristic of AD-related speech impairment [11]. This loss of intensity variation may be attributed to impairments in brain areas responsible for motor control and cognitive coordination, such as the basal ganglia and frontal lobe, which are commonly affected by AD. In contrast, the lack of significant change in intensity variability in the CN group suggests that healthy older adults generally maintain their ability to vary vocal intensity over time. Since intensity modulation is closely linked to expressive and emotional communication, this decrease in intensity variability may also indicate early signs of social withdrawal and reduced engagement, often observed in AD patients. Monitoring changes in intensity variability could, therefore, contribute to the early detection of AD-related speech changes.

The boxplots for pitch variability in figure 7 reveal a significant increase in pitch variability for the AD group over time ($p = 0.013$), while no significant change is observed for the CN group. The increase in pitch variability for the AD group may reflect a loss of control over vocal parameters, a phenomenon often linked to disruptions in the neural circuits that regulate motor and cognitive processes involved in speech [73]. As AD progresses, cognitive impairments can affect the ability to modulate pitch smoothly, leading to fluctuations that might manifest as greater pitch variability. This phenomenon is consistent with findings that associate AD with a decline in the coordination of speech-motor control, resulting in less stable vocal features over time [8]. Higher pitch variability may also reflect emotional or expressive dysregulation, commonly observed in AD patients. Changes in emotional regulation due to AD-related brain degeneration may lead to increased pitch variability, as individuals struggle to maintain consistent vocal tone [11]. In the CN group, the lack of significant change in pitch variability suggests that healthy aging does not severely impact pitch control to the same extent. While normal aging may result in slight changes to voice quality, the ability to maintain stable pitch variability is often preserved, reflecting preserved motor control mechanisms in older adults. This difference highlights how neurodegenerative processes in AD can impact vocal control, leading to distinctive patterns of pitch fluctuation that differ from normal aging.

Between-group The objective of the between-group analysis in this study was to compare speech measures between individuals with AD and CN individuals at two different time intervals. By contrasting speech performance across these two groups, this analysis highlights specific linguistic and acoustic deviations that can distinguish AD-related cognitive decline from the typical aging process, potentially providing insights into early diagnostic markers and progression patterns unique to AD.

Feature Name	Time Interval -2	Time Interval -1
Pause percentage	–	0.036
Pause speech ratio [sec]	–	0.036
Speech rate	–	< 0.001
Articulation rate	–	< 0.001
Average syllable duration [sec]	-	< 0.001
Intensity (std)[dB]	< 0.001	< 0.001
Pitch (std) [Hz]	0.02	< 0.001

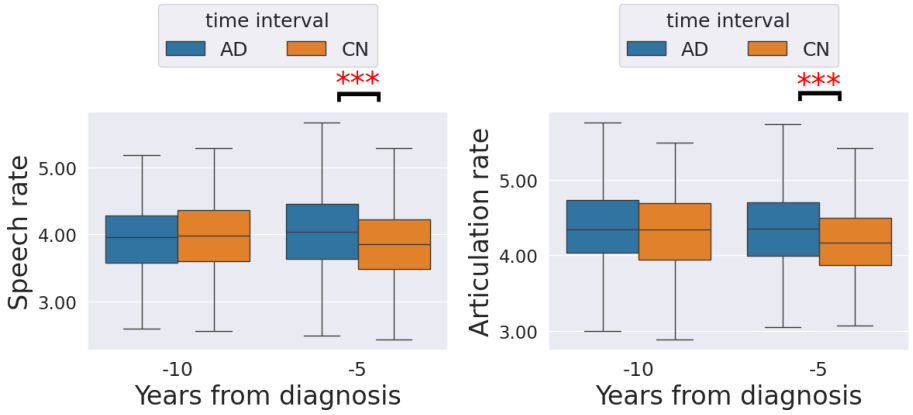
Table 10: The results of the between-group analysis for acoustics features are presented, where statistical significance was determined using the Mann-Whitney U test. For each feature showing significant differences, the table provides the relevant experimental group, adjusted p-value in both considered intervals in this study. Note: – stands for nonsignificant results $p \geq 0.05$.



(a) *(b)*

Figure 8: Boxplot of Pause Percentage [%] (a) and Pause Speech Ratio [s] (b) for AD and CN groups in the two time intervals considered in the between-group analysis.

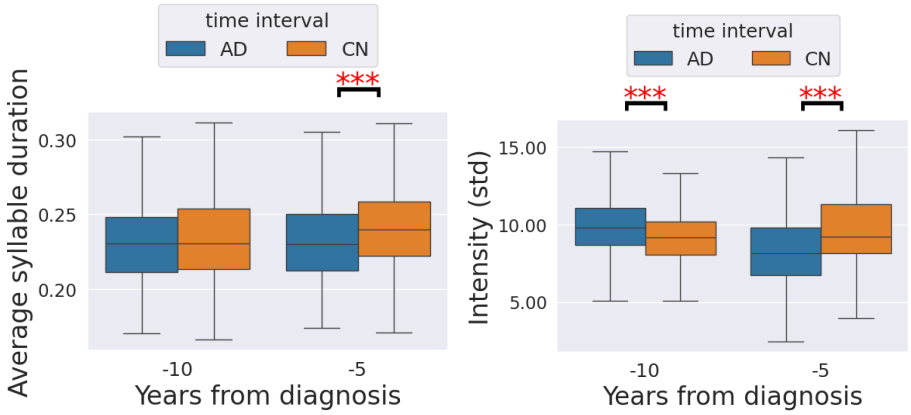
The boxplots for pause percentage and pause speech ratio highlight a significant divergence between AD and CN groups in time interval -1 ($p = 0.04$), aligning with previous within-group findings. While the CN group decreases pauses over time, reflecting typical aging adaptations that enhance speech fluency, the AD group exhibits a slight increase in these measures, indicating deteriorating speech control associated with cognitive decline. This contrast is particularly evident at time interval -1, where the CN group's improved fluency diverges from the AD group's worsening pause patterns. These findings suggest that the increase in pauses for AD patients may stem from challenges in word retrieval and sentence planning, reflecting broader impairments in cognitive and motor control that are characteristic of Alzheimer's pathology.



(a) *(b)*

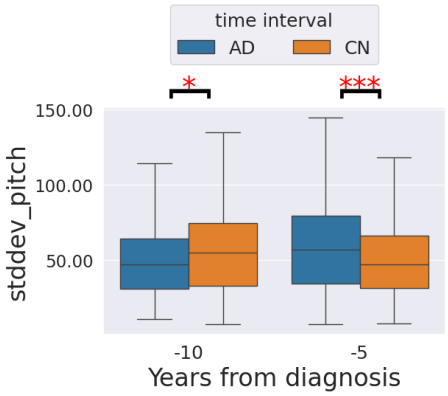
Figure 9: Boxplot of Speech Rate (a) and Articulation Rate (b) for AD and CN groups in the two time intervals considered in the between-group analysis.

Similarly, speech rate, articulation rate, and average syllable duration reveal a pattern similar to the previous findings. At the time interval -2, there is no significant difference between the AD and CN groups in these measures. However, at the time interval -1 (5 years before diagnosis), a significant difference emerges, primarily due to changes in the CN group over time. Specifically, the CN group shows a tendency toward reduced speech and articulation rates, as well as increased syllable duration, which could reflect adaptive strategies or normal aging processes aimed at maintaining speech clarity and efficiency [71]. In contrast, the AD group does not exhibit substantial changes across these time intervals, suggesting that they may have already reached a reduced level of speech fluency earlier in the disease progression and were also affected by choice of diarization criteria and properties of the AD group recordings. This aligns with previous within-group findings where the CN group displayed greater adaptive changes, whereas the AD group showed more stable but impaired performance, possibly due to early declines in cognitive and motor control over speech [13] [72]. These between-group differences underscore the subtle speech deterioration associated with AD that differentiates it from normal aging.



(a) *(b)*

Figure 10: Boxplot of Average syllable duration (a) and Intensity std (b) for AD and CN groups in the two time intervals considered in the between-group analysis.



(a)

Figure 11: Boxplot of stddev pitch (a) for AD and CN groups in the two time intervals considered in the between-group analysis.

In time interval -10, the AD group exhibits a higher variability in intensity compared to the CN group, which may reflect individual differences in speaking style or vocal energy that are unrelated to pathological changes. However, by time interval -5, there is a marked decrease in intensity variability in the AD group, while the CN group remains relatively stable. This sharp decline in intensity (std) for the AD group suggests a deterioration in their ability to modulate vocal intensity, possibly due to weakened motor control and reduced cognitive engagement in speech as AD progresses [13]. This finding is consistent with the within-group analysis, where AD individuals showed a notable decline in intensity (std) over time, while the CN group maintained stable vocal modulation. For pitch variability, significant differences are present at both time intervals, largely driven by the increasing trend in the AD group over time. In time interval -2,

the AD and CN groups have more comparable levels of pitch variability. However, by time interval -1, the AD group shows a clear increase in pitch variability, while the CN group's pitch remains relatively unchanged. This increase in pitch variability for the AD group may be attributed to declining motor control and coordination, which can cause irregularities in pitch modulation, reflecting the cognitive-motor deterioration that accompanies AD [73]. The trend suggests that, as AD progresses, individuals may exhibit greater pitch fluctuations, possibly due to difficulties in maintaining smooth speech patterns. Together, the findings from intensity (std) and stddev pitch reveal distinct trajectories for the AD group compared to the CN group, emphasizing the unique speech alterations associated with Alzheimer's disease. The CN group remains stable in both intensity and pitch modulation over time, indicative of typical aging processes that preserve vocal control. In contrast, the AD group displays a marked reduction in intensity variability and an increase in pitch variability, signaling progressive motor and cognitive decline in speech production. These results underscore the potential of using variability in intensity and pitch as biomarkers to differentiate AD from normal aging.

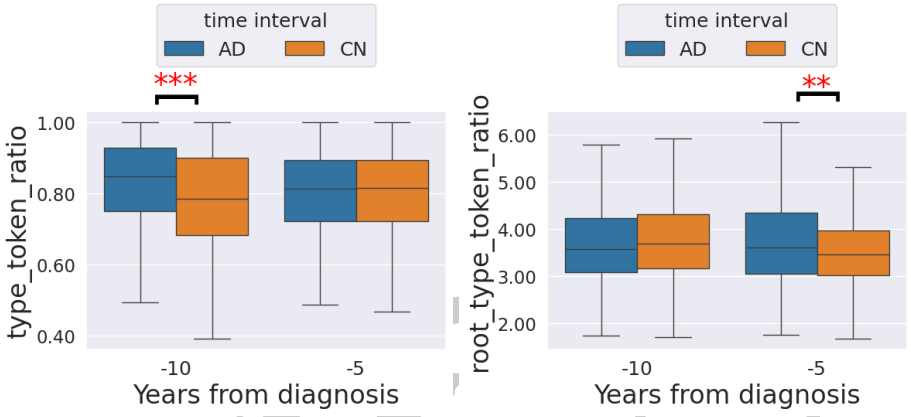
5.6.2 Linguistics

Within-group The objective of this study was to examine the trajectory of linguistic features in individuals with AD over time, focusing on language changes that may reflect cognitive decline. By comparing the linguistic performance of individuals with AD to a control group matched for age, sex, and nationality, we aimed to identify language patterns specific to AD and differentiate them from normal aging-related changes. This within-group analysis provides valuable insights into how vocabulary diversity, syntactic complexity, and coherence may deteriorate as the disease progresses, potentially serving as early markers of cognitive impairment in AD.

The absence of significant differences between time intervals in AD and CN groups suggests that the linguistic features selected in this study may lack the sensitivity needed to capture subtle language changes associated with AD progression. This finding aligns with our classification results, where interpretable linguistic features yielded similar F1 and accuracy values (0.62 and 0.63) for time intervals -2 and -1, pointing to limited variability over time. Moreover, these accuracy scores are relatively low compared to the 0.75 achieved using non-interpretable features, which incorporate more complex and nuanced linguistic patterns that interpretable features may not fully capture [13, 22, 32]. The challenge with interpretable linguistic features lies in their simplicity and limited scope. While metrics such as type-token ratio or phrase counts offer insight into vocabulary diversity and syntactic structure, they may not encompass the depth of syntactic, semantic, and pragmatic changes seen in cognitive decline [73, 74]. Non-interpretable features derived from deep learning embeddings, by contrast, can capture complex interactions between language elements, thus potentially identifying patterns missed by more traditional measures [58]. This disparity underscores the need for richer linguistic feature sets when analyzing cognitive changes in AD, as a limited feature set may fail to detect the gradual, nuanced decline observed in AD-related language deterioration.

Feature Name	Time Interval -2	Time Interval -1
Type Token Ratio	< 0.001	–
Root Type Token Ratio	–	0.004

Table 11: The results of the between-group analysis for linguistics features are presented, where statistical significance was determined using Mann-Whitney U test. For each feature showing significant differences, the table provides the relevant experimental group, adjusted p-value in both considered intervals in this study. Note: – stands for nonsignificant results $p \geq 0.05$.



(a) *(b)*

Figure 12: Boxplot of type token ratio (a) and root type token ratio (b) for AD and CN groups in the two time intervals considered in the between-group analysis.

Between-group The observed significant differences in Type-Token Ratio (TTR) ($p< 0.001$) at time interval -2 and Root Type-Token Ratio (RTTR) at time interval -1 ($p = 0.004$) between AD and CN groups may reflect distinct aspects of vocabulary use and lexical diversity. TTR, which measures lexical diversity by comparing the number of unique words (types) to the total number of words (tokens), tends to decrease as language becomes more repetitive or constrained [75]. In this case, the AD group decreases TTR over time, suggesting a shift toward more repetitive language or difficulty accessing diverse vocabulary, as commonly seen with Alzheimer's disease progression [13]. On the other hand, RTTR, defined as the ratio of types to the square root of tokens, adjusts for text length, making it less sensitive to variations in the overall number of words. The slight increase in RTTR for the AD group could indicate a higher relative diversity within shorter or simpler speech structures. This pattern may arise if individuals with AD rely on frequently used vocabulary within concise sentences, creating an impression of relative diversity on a smaller scale without genuinely expanding vocabulary range [39]. This combination of decreasing TTR and increasing RTTR suggests that as AD progresses, individuals may simplify their speech while retaining a narrow vocabulary range. This adaptation could reflect a compensatory strategy where simpler sentences allow for relatively diverse word use at the local level, even if overall lexical richness is limited [74]. Besides, it also matches the results we found in the within-group experiment and classification task, where we do not observe

too many significant changes between time intervals -2 and -1. Moreover, these findings align with our within-group and classification analyses, where no substantial linguistic change was observed between time intervals -2 and -1. This consistency across analyses suggests that the interpretable linguistic features considered in this study may not capture all linguistic patterns relevant to early AD detection.

6 Additional Remarks and Limitations

In this section, we discuss the limitations encountered in this study. During the diarization process, there were instances where segments attributed to the target speaker occasionally included brief portions from other interlocutors. To mitigate this, we filtered for speech segments that were at least 8 seconds long, which helped to minimize this issue. Nevertheless, the diarization output was not entirely accurate in all instances. Moreover, due to the nature of some AD target recordings, which often exhibit longer durations with frequent pauses, the WhisperX model occasionally split these segments into very short ones under $8\,\mathrm{sec}$. These shorter segments were filtered out based on our criteria, leading to potential data loss. This filtering, as illustrated in our longitudinal analysis boxplots, may have resulted in limited data availability for certain AD cases. Additionally, certain videos displayed inconsistencies between the upload date and the actual recording date. To address this, we cross-referenced video metadata, such as titles and descriptions, to more accurately infer the original recording year. However, it is possible that some discrepancies remained undetected, potentially leading to the incorrect classification of videos within specific experimental time intervals.

In addition, while we made significant efforts to balance the experimental groups by factors such as age, sex, and ethnicity, perfect alignment across all these dimensions was challenging due to data limitations. The dataset also does not include information regarding participants' medication schedules, such as how recently medication was taken or the duration of its peak effect in the periods following diagnosis. This omission is important, as taking medication before speaking or participating in interviews could enhance speech and language performance, whereas the lack of medication might negatively impact speech production. Furthermore, information on disease severity is missing from the dataset, which would have been valuable for a more comprehensive longitudinal analysis, enabling us to understand how the selected features change with different stages of disease progression.

For our experimental design, we categorized the data into two broader time intervals. While a more granular approach, with shorter time windows (such as one- or two-year intervals), could capture individual speech variations with greater precision, we chose this interval length to reliably maintain a minimum number of recordings per speaker within each interval. This approach allowed for more robust analysis by ensuring sufficient data representation across time periods.

The dataset inherently contains various types of real-world noise, such as background sounds, reverberation, laughter, overlapping speech, room acoustics, and inconsistencies in recording quality and channel noise. Consequently, these factors can impact the values of both interpretable and non-interpretable features. Given these limitations, the findings of this study can be considered somewhat similar to a simulation. Conducting

experiments in controlled acoustic environments, such as a laboratory, could provide a more accurate assessment of changes in interpretable features and potentially enhance classification accuracy—an improvement that would be especially advantageous for clinical applications.

Finally, as demonstrated in the longitudinal analysis of linguistic features, the interpretable features selected in this study appear insufficient to comprehensively capture the full spectrum of linguistic patterns associated with Alzheimer's disease. However, the high classification performance achieved with NIFMs underscores the critical role that linguistic characteristics play in detecting prodromal-stage AD. This suggests that linguistic features hold considerable potential as early indicators of cognitive decline. Future studies should aim to expand the range of linguistic features, including more nuanced and diverse metrics, to better encapsulate the complexity of language impairment in AD. By doing so, researchers could improve diagnostic accuracy, providing more robust tools for early detection and monitoring of AD progression.

7 Conclusions and Future Work

This study introduces a novel longitudinal speech dataset, ADCeleb, comprising speech samples from 40 individuals diagnosed with AD and 40 CNs matched by age and sex. The dataset includes samples from the AD group spanning from 10 years before the YoD up to the YoD itself. We conducted binary classification experiments to establish baseline performance on this dataset and explore key research questions (see Section 3.2). Additionally, a longitudinal analysis, including within-group analysis and between-group analysis, was performed to observe the progression of various speech features over time, beginning from the prodromal phase and extending up to YoD. The experimental findings lead us to conclude the following:

1. In our binary classification experiments, both IFMs and NIFMs for acoustic and linguistic features yielded promising classification outcomes when applied to pre-diagnosis data. Classification performance generally improved as data from later time intervals closer to the YoD were included, reflecting more pronounced AD-related speech changes and increasing emergence of neurological impairments patterns characteristic of the disease. Notably, NIFMs for linguistic features demonstrated superior overall performance, underscoring the significant role of linguistic changes in detecting early AD-related impairments. These emerging patterns not only highlight the early onset of subtle language alterations associated with AD but also underscore the potential for developing more effective early detection models that could capture the nuanced, progressive nature of language decline in AD.

2. The longitudinal analysis of acoustic features identified specific speech characteristics in the AD group that evolved from the prodromal stage (10 years before diagnosis) up to the YoD. Within the AD group, we observed a significant decline in intensity variability alongside a marked increase in pitch variability from the prodromal phase to YoD, changes that were not mirrored in the CN group. This absence of similar trends in the CN group suggests that these changes in the

AD group are not merely effects of normal aging but are likely tied to disease progression. Additionally, while the CN group exhibited significant decreases in pause percentage and pause speech ratio, no comparable changes were observed for these features in the AD group, indicating that AD individuals may experience early cognitive-linguistic decline with minimal impact on motor control of pauses. The CN group further showed significant declines in speech rate and articulation rate, alongside an increase in average syllable duration, while the AD group showed no significant results for these features. This pattern suggests that these metrics could serve as early indicators of cognitive decline specifically linked to AD. However, strict diarization criteria applied during analysis may have led to some data loss, particularly in the AD group, potentially impacting the robustness of longitudinal findings in datasets like ADCeleb.

3. The longitudinal analysis of linguistic features identified specific patterns in the AD group that changed from the 10 years before diagnosis through to the YoD. In the within-group analysis, no significant changes were found in either the AD or CN group, likely due to the limited scope of interpretable linguistic features available in this study, which may not comprehensively capture the complexity of linguistic changes. The between-group analysis supports these findings, showing that only some of the lexical richness metrics consistently revealed significant differences at time intervals -2 and -1, respectively. This pattern suggests that, as AD progresses, individuals may tend to simplify their language, maintaining a limited vocabulary range. These results emphasize the need for a broader set of linguistic features to better detect subtle language changes associated with AD progression.

The significance of this study lies in its pioneering approach to understanding the early linguistic and acoustic markers associated with AD, particularly in the prodromal phase. By leveraging a longitudinal design that spans up to 10 years before diagnosis, this study provides insights into the subtle yet progressive changes in speech patterns that may signal the onset of AD. The results emphasize the potential of using both interpretable and non-interpretable features to detect early cognitive decline, highlighting linguistic simplifications and specific acoustic variations that differentiate AD from normal aging. This research is crucial for advancing early detection capabilities in AD. Developing models based on post-diagnosis data and validating them with pre-diagnosis samples creates a foundational tool for proactive healthcare interventions. Early identification of prodromal AD allows for timely, personalized treatments that align with the precision medicine framework to improve individual patient outcomes. Additionally, these findings offer a non-invasive, objective method to monitor and assess cognitive changes, enhancing their applicability in real-world healthcare settings. The study also holds substantial value for clinical research, as identifying specific linguistic and acoustic patterns linked to AD facilitates targeted participant selection for clinical trials. By focusing on individuals at the earliest stages of AD, this approach maximizes trial efficacy and contributes significantly to neurodegenerative disease research. Ultimately, the insights from this study pave the way for more refined, accessible tools for early AD screening, treatment, and clinical research, underscoring the transformative potential of speech analysis in neurodegenerative disease management.

Future research should address the limitations encountered in this study to further enhance our understanding of prodromal speech markers in AD. First, while this study relied on naturalistic, real-world data, future work would benefit from data collection in controlled environments to minimize background noise, reverberation, and other external variables. Such controlled conditions could yield more precise acoustic and linguistic measurements, leading to a clearer understanding of speech changes specific to AD. Additionally, expanding the range of interpretable linguistic features is critical. The current set of features, while informative, may not capture the full complexity of linguistic patterns associated with early cognitive decline. Incorporating a broader array of linguistic metrics could improve sensitivity in detecting subtle changes during the prodromal phase, allowing researchers to better differentiate between AD-related and normal aging. Another important avenue for future studies is the systematic metadata collection, including disease severity, medication history, and specific recording details such as upload and recording dates. This information would allow for a more nuanced longitudinal analysis by accounting for potential influences on speech and language performance. In summary, future research efforts incorporating controlled data collection, a more comprehensive feature set, and detailed metadata will provide a more robust foundation for early AD detection and monitoring, ultimately advancing clinical and research applications in neurodegenerative disease management.

References

[1] I. Vigo, L. Coelho, and S. Reis, "Speech- and language-based classification of alzheimer's disease: A systematic review," *Bioengineering*, vol. 9, no. 1, 2022.

[2] C. De Looze, A. Dehsarvi, L. Crosby, A. Vourdanou, R. F. Coen, B. A. Lawlor, and R. B. Reilly, "Cognitive and structural correlates of conversational speech timing in mild cognitive impairment and mild-to-moderate alzheimer's disease: Relevance for early detection approaches," *Frontiers in Aging Neuroscience*, vol. 13, 2021.

[3] P. H. F. B. Maysa Luchesi Cera, Karin Zazo Ortiz and T. Minett, "Phonetic and phonological aspects of speech in alzheimer's disease," *Aphasiology*, vol. 32, no. 1, pp. 88–102, 2018.

[4] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. 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*, vol. 7, 2015.

[5] M. L. Cera, K. Z. Ortiz, P. H. F. Bertolucci, T. Tsujimoto, and T. Minett, "Speech and phonological impairment across alzheimer's disease severity," *Journal of Communication Disorders*, vol. 105, p. 106364, 2023.

[6] S. Nasreen, M. Rohanian, J. Hough, and M. Purver, "Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features," *Frontiers in Computer Science*, vol. 3, 2021.

[7] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," 2020.

[8] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, G. Szatlóczki, E. Biró, F. Zsura, M. Pakaski, and K. János, "Automatic detection of mild cognitive impairment from spontaneous speech using asr," 09 2015.

[9] K. Lopez-de Ipiña, J. Solé-Casals, H. Eguiraun Martinez, J. Alonso, C. Travieso, M. Ecay, A. Ezeiza, N. Barroso, P. Martinez-Lage, and B. Beitia, "Feature selection for spontaneous speech analysis to aid in alzheimer's disease diagnosis: A fractal dimension approach," *Computer Speech & Language*, vol. 30, 08 2014.

[10] J. Cummings, G. Lee, A. Ritter, and K. Zhong, "Alzheimer's disease drug development pipeline: 2018," *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, vol. 4, no. 1, pp. 195–214, 2018.

[11] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana, "Speech in Alzheimer's Disease: Can Temporal and Acoustic Parameters Discriminate Dementia?," *Dementia and Geriatric Cognitive Disorders*, vol. 37, pp. 327–334, 01 2014.

[12] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. Robert, and R. David, "Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, p. 112–124, 03 2015.

[13] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease: JAD*, vol. 49, no. 2, pp. 407–422, 2016.

[14] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. A. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2081–2090, 2011.

[15] K. Forbes-McKay and A. Venneri, "Detecting subtle spontaneous language decline in early alzheimer's disease with a picture description task," *Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, vol. 26, pp. 243–54, 10 2005.

[16] V. Taler and N. Phillips, "Language performance in alzheimer's disease and mild cognitive impairment: A comparative review," *Journal of clinical and experimental neuropsychology*, vol. 30, pp. 501–56, 08 2008.

[17] S. Ahmed, A.-M. Haigh, C. Jager, and P. Garrard, "Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease," *Brain : a journal of neurology*, vol. 136, 10 2013.

[18] P. Garrard, L. M. Maloney, J. R. Hodges, and K. Patterson, "The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author," *Brain*, vol. 128, pp. 250–260, 12 2004.

[19] M. Nicholas, L. K. Obler, M. L. Albert, and N. Helm-Estabrooks, "Empty speech in alzheimer's disease and fluent aphasia," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 3, pp. 405–410, 1985.

[20] S. Luz, F. Haider, S. de la Fuente Garcia, D. Fromm, and B. Macwhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," 10 2020.

[21] F. Haider, S. de la Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2020.

[22] M. Yancheva, K. Fraser, and F. Rudzicz, "Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies* (J. Alexandersson, E. Altinsoy, H. Christensen, P. Ljunglöf, F. Portet, and F. Rudzicz, eds.), (Dresden, Germany), pp. 134–139, Association for Computational Linguistics, Sept. 2015.

[23] L. Ilias and D. Askounis, "Context-aware attention layers coupled with optimal transport domain adaptation and multimodal fusion methods for recognizing dementia from spontaneous speech," *Knowledge-Based Systems*, vol. 277, p. 110834, 2023.

[24] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "Unsupervised domain adaptation for dysarthric speech detection via domain adversarial training and mutual information minimization," in *Interspeech 2021*, pp. 2956–2960, 2021.

[25] A. Wisler, V. Berisha, A. Spanias, and J. Liss, "Noise robust dysarthric speech classification using domain adaptation," in *2016 Digital Media Industry & Academic Forum (DMIAF)*, pp. 135–138, 2016.

[26] K. Croot, J. R. Hodges, J. Xuereb, and K. Patterson, "Phonological and articulatory impairment in alzheimer's disease: A case series," *Brain and Language*, vol. 75, no. 2, pp. 277–309, 2000.

[27] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181–197, 2019.

[28] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei, "Text embeddings by weakly-supervised contrastive pre-training," 2024.

[29] A. White and K. Witty, "Men's under use of health services – finding alternative approaches," *Journal of Men's Health*, vol. 6, no. 2, pp. 95–97, 2009.

[30] Category:People with Alzheimer's disease, "Category:people with alzheimer's disease," 2024. [Online; accessed 30-May-2024].

[31] M. Bain, "Whisperx." `https://github.com/m-bain/whisperX`, 2022.

[32] A. Favaro, T. Cao, N. Dehak, and L. Moro-Velazquez, "Leveraging universal speech representations for detecting and assessing the severity of mild cognitive impairment across languages," in *Proc. Interspeech 2024*, pp. 972–976, 2024.

[33] V. Boschi, E. Catricalà, M. Consonni, C. Chesi, A. Moro, and S. F. Cappa, "Connected speech in neurodegenerative language disorders: A review," *Frontiers in Psychology*, vol. 8, 2017.

[34] J. Yuan, X. Cai, Y. Bian, Z. Ye, and K. Church, "Pauses for detection of alzheimer's disease," *Frontiers in Computer Science*, vol. 2, 2021.

[35] P. Pastoriza-Domínguez, I. G. Torre, F. Diéguez-Vide, I. Gómez-Ruiz, S. Geladó, J. Bello-López, A. Ávila Rivera, J. A. Matías-Guiu, V. Pytel, and A. Hernández-Fernández, "Speech pause distribution as an early marker for alzheimer's disease," *Speech Communication*, vol. 136, pp. 107–117, 2022.

[36] A. Anikin, S. Barreda, and D. Reby, "A practical guide to calculating vocal tract length and scale-invariant formant patterns," *Behavior Research Methods*, vol. 56, pp. 5588–5604, 2024.

[37] J. M. Hillenbrand and M. J. Clark, "The role of f_0 and formant frequencies in distinguishing the voices of men and women," *Attention, Perception, Psychophysics*, vol. 71, no. 5, pp. 1150–1166, 2009.

[38] A. C. Lammert and S. S. Narayanan, "On short-time estimation of vocal tract length from formant frequencies," *PLOS ONE*, vol. 10, pp. 1–23, 07 2015.

[39] M. A. Covington and J. D. McFall, "Cutting the gordian knot: The moving-average type–token ratio (mattr)," *Journal of Quantitative Linguistics*, vol. 17, no. 2, pp. 94–100, 2010.

[40] H. Chen, "A lexical network approach to second language development," *Humanities and Social Sciences Communications*, vol. 10, p. 735, 2023.

[41] J. S. Yang, C. Rosvold, and N. Bernstein Ratner, "Measurement of lexical diversity in children's spoken language: Computational and conceptual considerations," *Frontiers in Psychology*, vol. 13, 2022.

[42] Çağrı Çöltekin and T. Rama, "What do complexity measures measure? correlating and validating corpus-based measures of morphological complexity," *Linguistics Vanguard*, vol. 9, no. s1, pp. 27–43, 2023.

[43] P. Mccarthy and S. Jarvis, "Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment," *Behavior research methods*, vol. 42, pp. 381–92, 05 2010.

[44] S. Jarvis, "Capturing the diversity in lexical diversity," *Language Learning*, vol. 63, 03 2013.

[45] N. Gao and Q. He, "A dependency distance approach to the syntactic complexity variation in the connected speech of alzheimer's disease," *Humanities and Social Sciences Communications*, vol. 11, no. 1, p. 995, 2024.

[46] K. D. Mueller, R. L. Koscik, B. P. Hermann, S. C. Johnson, and L. S. Turkstra, "Declines in connected language are associated with very early mild cognitive impairment: Results from the wisconsin registry for alzheimer's prevention," *Frontiers in Aging Neuroscience*, vol. 9, 2018.

[47] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, pp. 341–345, 01 2001.

[48] C. D. D. M. P. T. I. Ildikó Hoffmann, Dezso Nemeth and J. Kálmán, "Temporal parameters of spontaneous speech in alzheimer's disease," *International Journal of Speech-Language Pathology*, vol. 12, no. 1, pp. 29–34, 2010. PMID: 20380247.

[49] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

[50] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," 2021.

[51] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech 2017*, pp. 2616–2620, 2017.

[52] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech 2018*, pp. 1086–1090, 2018.

[53] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech 2020*, pp. 5036–5040, 2020.

[54] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.

[55] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 12449–12460, Curran Associates, Inc., 2020.

[56] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "Superb: Speech processing universal performance benchmark," 2021.

[57] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[58] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[59] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021.

[60] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[61] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 8440–8451, Association for Computational Linguistics, July 2020.

[62] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 3982–3992, Association for Computational Linguistics, Nov. 2019.

[63] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," 2022.

[64] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2019.

[65] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[66] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[67] A. Balagopalan, B. Eyre, J. Robin, F. Rudzicz, and J. Novikova, "Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech," *Frontiers in Aging Neuroscience*, vol. 13, 2021.

[68] M. Yu, M. R. Gormley, and M. Dredze, "Combining word embeddings and feature embeddings for fine-grained relation extraction," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1374–1379, 2015.

[69] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, pp. 591–611, 1965.

[70] A. Favaro, C. Motley, T. Cao, M. Iglesias, A. Butala, E. S. Oh, R. D. Stevens, J. Villalba, N. Dehak, and L. Moro-Velázquez, "A multi-modal array of interpretable features to evaluate language and speech patterns in different neurological disorders," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 532–539, IEEE, 2023.

[71] M. A. Shafto and L. K. Tyler, "Language in the aging brain: The network dynamics of cognitive decline and preservation," *Science*, vol. 346, no. 6209, pp. 583–587, 2014.

[72] K. D. Mueller, R. L. Koscik, A. LaRue, L. R. Clark, B. Hermann, S. C. Johnson, and M. A. Sager, "Verbal Fluency and Early Memory Decline: Results from the Wisconsin Registry for Alzheimer's Prevention," *Archives of Clinical Neuropsychology*, vol. 30, pp. 448–457, 05 2015.

[73] S. Ash, C. McMillan, D. Gunawardena, B. Avants, B. Morgan, A. Khan, P. Moore, J. Gee, and M. Grossman, "Speech errors in progressive non-fluent aphasia," *Brain and Language*, vol. 113, no. 1, pp. 13–20, 2010.

[74] K. D. Mueller, R. L. Koscik, L. S. Turkstra, S. K. Riedeman, A. Larue, L. Clark, B. P. Hermann, M. A. Sager, and S. C. Johnson, "Connected language in late middle-aged adults at risk for alzheimer's disease.," *Journal of Alzheimer's disease : JAD*, vol. 54 4, pp. 1539–1550, 2016.

[75] P. M. McCarthy and S. Jarvis, "vocd: A theoretical and empirical evaluation," *Language Testing*, vol. 24, no. 4, pp. 459–488, 2007.