<u>Original Paper</u>

# Automated Speech Markers of Alzheimer Dementia: Test of Cross-Linguistic Generalizability

Paula Andrea Pérez-Toro[1,2,3], PhD; Franco J Ferrante[4,5,6], Eng; Gonzalo Pérez[4,5,6], Eng; Boon Lead Tee[7,8,9], MD; Jessica de Leon[8], MD; Elmar Nöth[1], PhD; Maria Schuster[10], PhD; Andreas Maier[1], PhD; Andrea Slachevsky[11,12,13,14], MD; Maria Luisa Gorno-Tempini[8], MD; Agustín Ibáñez[4,7,9,15], PhD; Juan Rafael Orozco-Arroyave[1,2], PhD; Adolfo García[4,7,16], PhD

[1]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[2]GITA Lab, Facultad de Ingeniería, Universidad de Antioquia, Medellín, Colombia

[3]Massachusetts General Hospital, Boston, MA, United States

[4]Centro de Neurociencias Cognitivas, University of San Andrés, Buenos Aires, Argentina

[5]Consejo Nacional de Investigaciones Científicas y Técnicas, Ciudad Autónoma de Buenos Aires, Argentina

[6]Facultad de Ingeniería, University of Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina

[7]Global Brain Health Institute, University of California, San Francisco, CA, United States

[8]Memory and Aging Center, Department of Neurology, University of California, San Francisco, San Francisco, CA, United States

[9]Trinity College Dublin, Dublin, Ireland

[10]Department of Otorhinolaryngology, Ludwig-Maximilians-Universität München, Munich, Germany

[11]Neuropsychology and Clinical Neuroscience Laboratory (LANNEC), Faculty of Medicine, University of Chile, Santiago, Chile

[12]Geroscience Center for Brain Health and Metabolism (GERO), Faculty of Medicine, University of Chile, Santiago, Chile

[13]Memory and Neuropsychiatric Clinic (CMYN), Neurology Department, Hospital del Salvador, Santiago, Chile

[14]Servicio de Neurología, Departamento de Medicina, Clínica Alemana, Santiago, Chile

[15]Latin American Brain Health (BrainLat) Institute, Adolfo Ibáñez University, Santiago, Chile

[16]Departamento de Lingüística y Literatura, Facultad de Humanidades, Universidad de Santiago de Chile, Santiago, Chile

**Corresponding Author:**
Adolfo García, PhD
Global Brain Health Institute
University of California
675 Nelson Rising Lane
San Francisco, CA, 94158
United States
Phone: 1 4154763722
Email: adolfo.garcia@gbhi.org

## *Abstract*

**Background:** Automated speech and language analysis (ASLA) is gaining momentum as a noninvasive, affordable, and scalable approach for the early detection of Alzheimer disease (AD). Nevertheless, the literature presents 2 notable limitations. First, many studies use computationally derived features that lack clinical interpretability. Second, a significant proportion of ASLA studies have been conducted exclusively in English speakers. These shortcomings reduce the utility and generalizability of existing findings.

**Objective:** To address these gaps, we investigated whether interpretable linguistic features can reliably identify AD both within and across language boundaries, focusing on English- and Spanish-speaking patients and healthy controls (HCs).

**Methods:** We analyzed speech recordings from 211 participants, encompassing 117 English speakers (58 patients with AD and 59 HCs) and 94 Spanish speakers (47 patients with AD and 47 HCs). Participants completed a validated picture description task from the Boston Diagnostic Aphasia Examination, eliciting natural speech under controlled conditions. Recordings were preprocessed and transcribed before extracting (1) speech timing features (eg, pause duration, speech segment ratios, and voice rate) and (2) lexico-semantic features (lexical category ratios, semantic granularity, and semantic variability). Machine learning classifiers were trained with data from English-speaking patients and HCs, and then tested (1) in a within-language setting (with

English-speaking patients and HCs) and (2) in a between-language setting (with Spanish-speaking patients and HCs). Additionally, the features were used to predict cognitive functioning as measured by the Mini-Mental State Examination (MMSE).

**Results:** In the within-language condition, combined speech timing and lexico-semantic features yielded maximal classification (area under the receiver operating characteristic curve [AUC]=0.88), outperforming single-feature models (AUC=0.79 for timing features; AUC=0.80 for lexico-semantic features). Timing features showed the strongest MMSE prediction (R=0.43, $P<.001$). In the between-language condition, speech timing features generalized well to Spanish speakers (AUC=0.75) and predicted Spanish-speaking patients' MMSE scores (R=0.39, $P<.001$). Lexico-semantic features showed lower performance (AUC=0.64) and no significant MMSE prediction (R=–0.31, $P=.05$). The combined model did not improve results (AUC=0.65; R=0.04, $P=.79$).

**Conclusions:** These results suggest that while both timing and lexico-semantic features are informative within the same language, only speech timing features demonstrate consistent performance across languages. By focusing on clinically interpretable features, this approach supports the development of clinically usable ASLA tools.

## Introduction

Alzheimer disease (AD) is a neurodegenerative condition involving an insidious decline of semantic and episodic memory alongside other functions [1]. Its current prevalence, of 55 million, will likely triple by 2050, with cases increasing by 116% in high-income countries and 250% in low- or middle-income countries [2-4]. This underscores the need for noninvasive, scalable markers that facilitate disease detection and monitoring [5]. Automated speech and language analysis (ASLA) meets these requisites [6-8].

In ASLA studies, patients and healthy controls (HCs) are simply required to speak, be it through spontaneous (eg, memory description), semispontaneous (eg, picture description), or nonspontaneous (eg, paragraph reading) tasks [9]. Thereupon, their recordings and transcripts can be digitally analyzed to identify disease-sensitive features [10,11]. This approach has been leveraged to detect early-stage cases [6,8,12], support differential diagnosis [7], predict dementia onset [6], and capture cognitive decline [13] and brain atrophy patterns [14,15]. As ASLA also reduces testing time and costs, it represents a promising framework to foster global equity in dementia assessments [16]. However, such potential remains unmet, as most studies target features with low interpretability and tests of between-language generalizability are incipient [16,17]. This casts doubts on the approach's clinical utility and cross-linguistic validity.

Interpretability is often undermined by 2 analytical strategies. One involves targeting heterogeneous feature sets that mix domains known to be affected (eg, semantics) and often spared (eg, morphosyntax) in AD [8,18,19]. The other relies on black-box models (eg, transformers) operating on hidden layers without clinical significance—eg, deep learning models such as BERT (Bidirectional Encoder Representations from Transformers) or Wav2vec (Meta), which produce high-dimensional representations that do not correspond to well-defined linguistic or acoustic constructs [20,21]. Both strategies yield variable outcomes that cannot be readily integrated with core clinical knowledge about AD, reducing the framework's translational potential.

Moreover, over 40% of ASLA research on AD targets English speakers [22]. This is highly inequitable, as English is spoken by less than 20% of the world's population [16] and AD is most prevalent in non-Anglophone countries [2-5]. Importantly, several language domains can be affected by this disease in English speakers and spared in users of other languages [22]. Thus, not all ASLA results from English-speaking cohorts may generalize well to other language groups.

Promisingly, recent developments allow the circumvention of both issues. Robust, interpretable results have been obtained in ASLA studies targeting (1) speech timing features (eg, pause duration) as proxies of lexical search effort [23-26]; and (2) lexico-semantic features, including word class features (proportion of different lexical categories) [27-29], semantic granularity (conceptual precision when naming entities) [7,30], and semantic variability (conceptual distance across successive words) [30]. Such features can reveal distinct processing demands and strategies in AD, potentially yielding maximal results when combined than when framed in isolation [8,31].

Also, these features can be tested for cross-cultural validity by training classifiers with data from English-speaking participants and testing them on users of a different language, as recently proposed [22] and preliminarily attempted in promising studies and challenges [32-34]. In particular, a 0-shot approach, involving no cross-linguistic calibration or transfer, creates highly stringent, unbiased conditions for this examination. Conceivably, cross-linguistic generalizability could be higher for speech timing features (as word retrieval effort should increase in AD irrespective of the language) than for lexico-semantic features (which vary widely between English and other languages) [35,36]. As proposed elsewhere, such features are interpretable because they reflect dysfunctions of semantic memory, a core system affected in early disease stages. This allows results to be understood in relation to well-established neuropsychological models of AD [4,5].

Against this background, we examined whether interpretable ASLA features yield cross-linguistic AD markers. Using speech

timing and lexico-semantic features from a validated task [17,19,24], we first trained classifiers with English-speaking patients and HCs, and then tested them in a within-language and a between-language setting (with English and Spanish speakers as testing folds, respectively). In each case, we ran separate classifiers for timing features, lexico-semantic features, and their combination. Finally, we explored whether the most sensitive features in each setting could predict patients' cognitive decline, indexed through the Mini-Mental State Examination (MMSE) [37]. Relative to current literature [38-40], our approach is unique in its integration of clinically motivated acoustic and linguistic features with unimodal and fusion-based classifiers for 0-shot within- and between-language classification and severity prediction. Adding novelty, our design involves Latino individuals, a large, underserved population with a high prevalence of AD [41,42].

For the within-language setting, we hypothesized that interpretable features would enable patient identification in both modalities, with the best outcomes resulting from their combination. For the between-language setting, we anticipated better cross-linguistic generalization for timing than for lexico-semantic features. Finally, considering recent findings [43,44], we anticipated that the most discriminatory features in each setting would reliably predict patients' MMSE scores. By testing these hypotheses, we seek to further ASLA research on AD with a translational, cross-cultural ethos.

## Methods

### Study Design

Our study involved an opportunistic design. Preregistration was not feasible as the protocols' stimuli, prompts, speech recording settings, and accompanying measures were established before the present investigation was conceived. Still, the analysis plan was established in the National Institutes of Health–approved project R01AG075775 (aim 1c), led by the corresponding author. Moreover, the current study extends and refines a previously reported analytical strategy [32], focused on embeddings as opposed to interpretable features. The current study's methods are diagrammed in Figure 1.

**Figure 1.** Study design. (A) This study involved English- and Spanish-speaking patients with AD and HCs, matched for sociodemographic variables (with patient groups also matched for MMSE scores). (B) Participants were recorded as they described the Cookie Theft picture. (C) Audios and transcripts were preprocessed with standard procedures before the extraction of speech timing and lexico-semantic features. (D) A machine learning classifier was trained with data from English-speaking patients with AD and HCs through a pipeline with decision trees and sparse linear models. (E) Classification performance was assessed in a within-language setting (with English-speaking patients and HCs) and in a between-language setting (with Spanish-speaking patients and HCs), considering 1=speech timing features, 2=lexico-semantic features, and 3=their combination via early fusion. The results were visualized via ROC curves and score distribution plots. (F) We computed correlations between the patients' actual MMSE scores and those predicted by each feature set. AD: Alzheimer disease; AUC: area under the curve; HC: healthy control; MMSE: Mini-Mental State Examination; RMSE: root-mean-squared error; ROC: receiver operating characteristic.
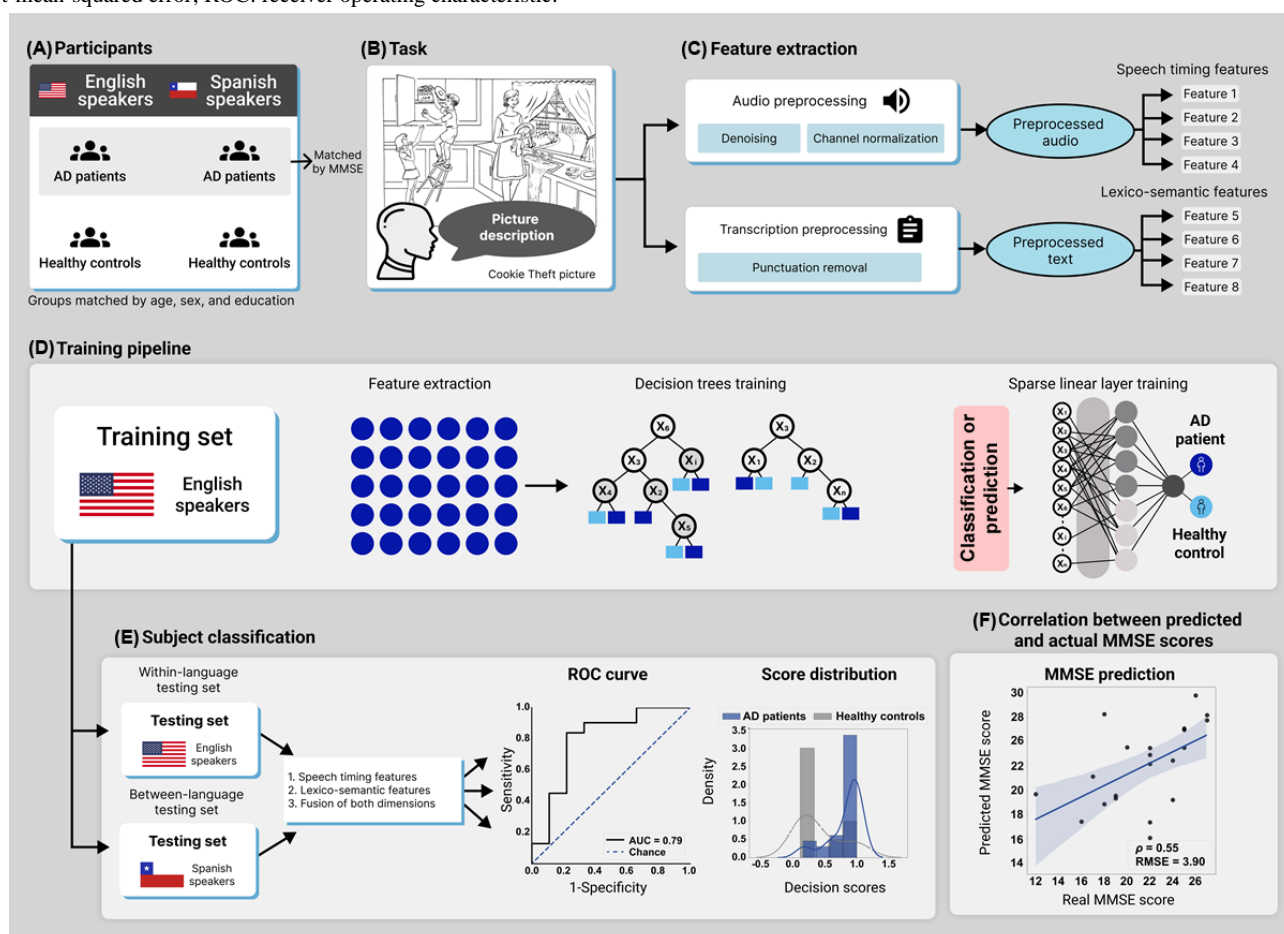
## Participants

This study used a cross-sectional design based on a convenience sample. The initial datasets comprised 242 native English speakers for the English dataset and 198 native Spanish speakers for the Spanish dataset. Native English speaker data came from the Pitt corpus [45], a widely used resource in ASLA research on AD [17], enabling global challenges such as Alzheimer's Dementia Recognition Through Spontaneous Speech and its variations [1]. Native Spanish speakers belonged to a Chilean cohort at the Hospital del Salvador's Memory and Neuropsychiatry Clinic [7,30]. Fifty-six participants were removed due to faulty recordings or incomplete metadata. Participants from both datasets were selected using custom-made Python code to ensure demographic similarity for subsequent analyses (Figure 1A), leading to the removal of an additional 173 participants. The final sample involved 211 participants across both datasets, all without any missing data. The English-language dataset (n=117), for the within-language setting, included 58 patients with AD and 59 HCs. The Spanish-language dataset (n=94), for the between-language setting, encompassed 47 patients with AD and 47 HCs [46].

All English-speaking participants completed a comprehensive neuropsychiatric evaluation, a semistructured psychiatric interview, and a neuropsychological assessment [47,48]. As the Pitt corpus's protocol began before the establishment of NINCDS-ADRDA (National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association) [49] and *DSM* (*Diagnostic and Statistical Manual of Mental Disorders*) [50] criteria, inclusion in the patient group was determined by frank

evidence of progressive cognitive and functional decline alongside abnormal MMSE scores (sample's mean 19.98, SD 4.79). Use of this dataset was strategic to (1) enable direct comparisons between our results and literature benchmarks while (2) guaranteeing that the same task was used in both languages. With regards to the Spanish-language dataset, AD diagnoses were made by expert neurologists following updated, validated procedures based on NINCDS-ADRDA criteria [22,51]. Outcomes on the MMSE [37] revealed abnormal cognitive scores (mean 22.11, SD 3.87) [52]. The HC group in both datasets consisted of cognitively preserved, functionally independent individuals. No participant had a history of (other) neurological disorders, psychiatric conditions, primary language deficits, or substance abuse, and all of them had normal or corrected-to-normal vision and hearing.

All 4 groups were matched for sex, age, and years of education, and dialect was consistent across participants in each language group (Table 1). MMSE scores were not matched between English- and Spanish-speaking patients with AD as dementia cutoffs differ between such populations [53]. Clinical and cognitive measures for the English and the Spanish datasets were administered following formal procedures of the University of Pittsburgh School of Medicine and harmonized protocols of the Multi-Partner Consortium to Expand Dementia Research in Latin America, respectively. The feasibility and acceptability of the Spanish data overall protocol are verified by its continuous successful application in multiple studies and populations [7,10,30,54-58] (see below for details). Participants provided written informed consent under the Declaration of Helsinki. The current study was approved by the institutional ethics committee.

[XSL•FO]

**RenderX**

**Table 1.** Participants' demographic and neuropsychological profiles.

| | English speakers | | Spanish speakers | | Statistics | | Pairwise comparisons | | |
|---|---|---|---|---|---|---|---|---|---|
| | Patient with AD[a] (N=58) | Healthy controls (N=59) | Patients with AD (N=47) | Healthy controls (N=47) | Value | P value | Groups | Estimate | P value |
| **Demographic data** | | | | | | | | | |
| Sex (female:male) | 35:23 | 36:23 | 24:23 | 33:14 | $\chi^2$=3.62 | .31[b] | N/A[c] | N/A | N/A |
| Age, mean (SD) | 65.21 (6.90) | 66.12 (7.08) | 70.19 (6.93) | 64.32 (17.48) | F=5.68 | .02[d] | | | |
| | | | | | | | EN[e] AD vs EN HCs[f] | –0.48 | .97[g] |
| | | | | | | | ES AD vs ES[h] HCs | 2.78 | .06[g] |
| | | | | | | | EN AD vs ES AD | –2.47 | .11[g] |
| | | | | | | | EN HCs vs ES HCs | 0.90 | .85[g] |
| Years of education, mean (SD) | 13.79 (2.32) | 13.46 (2.01) | 13.31 (4.42) | 14.53 (3.96) | F=3.02 | .10[d] | N/A | N/A | N/A |
| **Neuropsychological data** | | | | | | | | | |
| MMSE[i], mean (SD) | 19.98 (4.79) | N/A | 22.11 (3.87) | N/A | N/A | N/A | EN AD vs ES AD | –3.15 | .02[f] |

[a]AD: Alzheimer disease.

[b]P values calculated via the chi-squared test.

[c]N/A: not applicable.

[d]P values calculated via independent measures ANOVA, with results showing the interaction between condition (patients and controls) and language (English and Spanish).

[e]EN: English.

[f]HC: healthy control.

[g]P values for pairwise comparisons of significant interaction effects were calculated via independent measures ANOVA with Scheffé correction; crucially, no pairwise comparison was significant despite the significant interaction for age.

[h]ES: Spanish.

[i]MMSE: Mini-Mental State Examination.

## Experimental Task

Speech was elicited through the Cookie Theft picture task (Figure 1B) from the Boston Diagnostic Aphasia Examination [59], as described in previous reports of both datasets [7,17,30,32,43,46,60,61]. The stimulus is a black-and-white drawing of a mother and her 2 children performing several actions in a kitchen. Participants were asked to view the scene and describe its events and elements in as much detail as possible, in whichever order they preferred. No time limit was imposed. Examiners intervened only if participants needed assistance or stopped talking without mentioning central details. Importantly, this picture has proven useful in capturing speech and language abnormalities in English-speaking [62,63] and Spanish-speaking [30,58,64] cohorts.

The English recordings were captured in .mp3 format at a sample rate of 44,100 Hz and 16-bit resolution, following a comprehensive protocol overseen by the Alzheimer and Related Dementias Study at the University of Pittsburgh School of Medicine [45]. The Spanish recordings were obtained in a quiet environment using laptop computers fitted with noise-canceling microphones [30]. These recordings were captured in .wav format at a sample rate of 44,100 Hz and 16-bit resolution, all managed through Cool Edit Pro 2.0.

## Data Preprocessing and Feature Extraction

### Overview

Audio recordings and their transcripts were preprocessed and subjected to validated feature extraction procedures, as detailed below (Figure 1C).

### Audio Preprocessing and Extraction of Speech Timing Features

Recordings were preprocessed following a validated pipeline [65]. Segments including the examiner's voice were manually removed from the English and the Spanish recordings. The participants' speech signals were converted to .wav format and then resampled at a rate of 16 kHz through SoX (Sound Exchange) [66]. The output was then denoised using a recurrent neural network model with complex linear coding [67], and

manual inspection confirmed the absence of speech artifacts. Microphone-related biases were eliminated via channel and mean cepstral normalization [65].

Speech timing features were extracted at the whole-recording and the single-word level (Tables 2 and 3). Whole-recording features were calculated for the following variables: pause ratio, duration ratio, speech segment ratio, speech segment duration ratio, and voice ratio. Duration information was derived from an energy-based voice activity detection (VAD) algorithm to detect the presence of human speech and differentiate it from silence segments (indicating pauses), following these steps: (1) denoising using a recurrent neural network filter [67], (2) amplitude normalization, (3) VAD, and (4) segmentation and feature extraction. This pipeline follows validated procedures reported in prior literature [68]. To this end, the log energy is calculated on 25-ms windows with a 10-ms hop size, the DC (mean value of a signal) level of the energy is subtracted, and the signal is then smoothed by convolving it with a Gaussian window of 10-ms [69]. Speech segments were detected by identifying frames with log-energy levels above a heuristically defined (yet systematic) threshold by analyzing the average signal energy in silent and speech-labeled segments. In contrast, pauses were determined by frames with log-energy levels below

the threshold [70]. To calculate voice rate, we identified voice segments as those whose fundamental frequency differed from 0 [65] and computed the number of voiced segments per second. We note that the primary features used in our analysis are timing-based (eg, pause duration and speech rate), which are minimally affected by MP3 compression artifacts [71]. Indeed, MP3 compression on suprasegmental acoustic measures (eg, fundamental frequency, pitch range, and level) shows minimal measurement errors, typically below 2%, on compression bitrates in the 56-320 kbps range [72]. Although spectral distortions introduced by compression may influence certain acoustic features, our preprocessing pipeline includes noise-reduction and normalization steps that help mitigate these effects [67]. Importantly, denoising and energy-based pause detection improve performance metrics such as word error rate and character error rate [73], indicating that such preprocessing enhances reliability without artificially distorting pause-related features. Word-level features consisted of (normalized) word duration and (normalized) word count. Audio segments and their corresponding transcription units were time-aligned [74] as in previous works [75,76]. The single-word features were computed taking into account (1) syllables, (2) all words, (3) stop words, and (4) content words separately, as well as (5) all possible combinations thereof.

**Table 2.** List of speech timing features.

| Whole-recording level | Single-word level |
| --- | --- |
| Pause duration ratio | N/A[a] |
| Speech segment duration ratio | Word duration |
| Pause ratio | Normalized word duration |
| Pause duration | Word count |
| Speech segment duration | Normalized word count |
| Voiced rate | N/A |

[a]N/A: not applicable.

**Table 3.** List of vocabulary selection features.

| Word-class features | Semantic features |
| --- | --- |
| Noun ratio | N/A[a] |
| Verb ratio | Granularity |
| Adjective ratio | Semantic variability |
| Adverb ratio | N/A |

[a]N/A: not applicable.

All whole-recording and word-level features were expressed in seconds, and when possible, represented in terms of 6 statistical values (mean, SD, skewness, kurtosis, minimum value, and maximal value), as in previous research [65].

### Transcript Preprocessing and Extraction of Vocabulary Selection Features

We used previously reported transcriptions from both datasets. English recordings were transcribed using CHAT (Codes for the Human Analysis of Transcripts), a standardized system

developed by the CHILDES (Child Language Data Exchange System) project [77] for transcribing and analyzing spoken language data. It uses structured tiers to represent participant information, transcribed speech, translations, and annotations. Special characters used by this procedure (eg, ?, –, (), +) were automatically removed from each text. Spanish recordings were transcribed via an automatic speech-to-text service [78] and manually revised. The rare occurrences of unintelligible words were discarded. In all cases, examiners' interventions were removed (including the silent segment following each

intervention). Transcriptions were analyzed regarding word-class and semantic features.

First, word-class features were calculated as the ratio of each content word type: nouns, verbs, adjectives, and adverbs—namely, words which, unlike functional categories, involve representational content and thus tap on core conceptual processes affected by AD [79]. We calculated these proportions relative to the text's overall word count and the number of content words only. Second, we calculated 2 types of semantic features. We used the Natural Language Toolkit library in Python to estimate each word's granularity using WordNet, a hierarchical graph whose nodes branch out from the highest hypernym *entity* to more specific concepts (such as *animal*, *dog*, and *bulldog*, in increasing order of granularity). Granularity is calculated as the smaller number of nodes between a word and the root node *entity*. For instance, words in bin-3 are closer to the "entity" than those in bin-10, indicating that the former refers to more general concepts [7]. We extracted distributional statistics and the proportion of words with low, intermediate, and high granularity over the total number of words. Third, in line with reported procedures [30], semantic variability was established across: (1) content words, (2) not-repeated adjacent words, and (3) not-repeated adjacent content words. We allocated individual words to vectors within the vocabulary through the fastText (Meta) model [66], pretrained with an extensive corpus. Specific pretrained monolingual models were used for each language (cc.en.300.bin for English and cc.es.300.bin for Spanish), both downloaded from the official fastText website. These were trained separately on each language's Common Crawl and Wikipedia corpora under an identical configuration (Continuous Bag of Words) with position-weights, 300 dimensions, character n-grams (length 5), window size 5, and 10 negatives. Distances between adjacent vectors were stored in a time series. Semantic consistency was determined by calculating the variance of this combined time series. A higher value in semantic consistency is observed in texts where consecutive words represent distant concepts. As in previous NLP research [6], variables were analyzed whenever feasible, considering their mean, SD, skewness, kurtosis, minimum value, and maximal value.

## Data Analysis

Within- and between-language analyses were performed following identical steps (Figure 1D). The within-language analysis used 80% of participants from each group for model training, and the remaining 20% were set aside as a hold-out sample for testing. The between-language analysis used the same training model as the English-language experiment and used the entire Spanish-speaking sample for testing. In all cases, we considered three classification scenarios based on (1) acoustic features only, (2) linguistic features only, and (3) both types of features combined via early fusion (Figure 1E). We trained the model with dedicated speaker-independent training and test splits to ensure robustness. Note that the Spanish-speaking cohort was used exclusively as an independent testing set.

We used a modified algorithm that combines decision trees with a partially connected, sparse multilayer perceptron (MLP)

network [80,81]. This network's structure is based on information from a group of decision trees trained to discriminate AD. It is not constructed with a fully connected layer; the interconnections are guided by the decision tree classifier preceding the MLP. Each neuron specializes in processing a specific tree, taking input from features generated by the decision tree. Importantly, the MLP's decision tree component improves interpretability, which is important for clinical relevance, and its sparsely connected nature results in fewer parameters, reducing the risk of overfitting. Feature importance is assessed by counting the neurons connected to a specific feature (the more connections, the greater the importance).

The models underwent a bootstrapping strategy, with an 80/20 split. Features were normalized as *z*-scores relative to the train set parameters. To reduce the risk of overfitting, several strategies were implemented during model training [82]. Ridge regularization [83] was applied to penalize large weight magnitudes, encouraging simpler models that generalize better to unseen data. Early stopping [84] was used with a patience of 10 epochs, halting training when validation performance ceased to improve and preventing the model from overadapting to the training data. Within each feature set (speech timing, linguistic, and fusion), we report the best-performing combination of trees $\in \{10,20,40,60,80,100\}$ and depth $\in \{4,6,8,10\}$ ("depth" refers to the number of levels or splits in a single decision tree, indicating its complexity, while "number of trees" refers to the quantity of decision trees used in an ensemble model, such as a random forest or XGBoost [Extreme Gradient Boosting]) across both settings (within-language and between-language), based on the area under the receiver operating characteristic curve (AUC) values. Given the class imbalance problem, we used weighting techniques to assign a higher penalty to misclassification of the minority class, thus adjusting the cross-entropy loss within the neural network [85]. Details about hyperparameter tuning are offered in Multimedia Appendix 1 (Sections S3 and S4). The corresponding code implementations are available on GitHub [86]. Pairwise DeLong tests were conducted to assess whether AUC differences across models were statistically meaningful.

Additionally, we examined whether speech timing and linguistic features in each language were predictive of patients' MMSE scores (Figure 1F). We report the best correlation across both languages in each case based on the Spearman correlation coefficient (ρ) and the root-mean-squared error (RMSE). This approach indicates how close the predicted values are to the true values on average.

## Ethical Considerations

Data collection procedures of the English dataset (from the publicly available Pitt Corpus) were approved by the Institutional Review Board of the University of Pittsburgh. Participants provided informed consent in accordance with the protocols established by the Alzheimer and Related Dementias Study at the University of Pittsburgh School of Medicine. Data from the Chilean cohort were collected with approval from the Ethics Committee of the Servicio de Salud Metropolitano Oriente, Santiago (project code: ANID-Fondap 15150012). All

participants provided written informed consent under the 1964 Declaration of Helsinki. The Spanish dataset's ethical approval allowed for data to be integrated with international datasets for projects involving the principal investigators (note that these data were analyzed by our local research team rather than shared with external collaborators). The English dataset is open for cross-national analyses, as shown in previous papers and challenges [17,34,43]. Both datasets were deidentified via the removal of personal information (names, occupations, and geographical locations). Data shared in our repository was limited to deidentified feature sets; there being no raw audio recordings (in particular, the speech timing features we targeted cannot be leveraged to infer a speaker's identity). Importantly, the picture description task we used elicits no self-referential information, as verified manually in each of the transcriptions. Additionally, we ensured harmonization of ethical oversight across both sites by aligning our procedures with international best practices for secondary use of sensitive speech data. Privacy and data protection were guaranteed by (1) robust deidentification, (2) restricting analyses to derived features that cannot be reverse-engineered into intelligible speech or personal identity, and (3) refraining from sharing raw audio between sites. These safeguards are consistent with concerns highlighted in the speech technology and legal communities regarding General Data Protection Regulation compliance and privacy protection [87]. Accordingly, ethical clearance covered the full scope of our cross-linguistic analyses and ensured compliance with data governance requirements at both sites. No compensation was offered or given to participants.

## Results

### Within-Language Results

As shown in Figure 2, within-language (English-to-English) analyses revealed similar classification based on timing features from all words and stop words (AUC=0.79, Figure 2A) and based on lexico-semantic features (AUC=0.80, Figure 2B). The fusion of both dimensions yielded maximal discrimination (AUC=0.88, Figure 2C). The outcomes of these 3 classifiers did not differ significantly (speech timing vs lexico-semantic features: $P=.96$; lexico-semantic features vs fusion: $P=.55$; speech timing vs fusion: $P=.52$). Further, the MMSE scores of English-speaking patients correlated significantly with those predicted by within-language regressions based on speech timing features ($\rho=0.430$, $P<.001$, RMSE=5.620), but not with those based on lexico-semantic features ($\rho=0.271$, $P=.39$, RMSE=7.898) or the combination of both feature sets ($\rho=0.088$, $P=.79$, RMSE=6.257).

**Figure 2.** Within-language results. The ROC curves (top row) and the classification scores (middle row) show outcomes for classifying AD patients and HCs in the English-to-English setting. (A) Discrimination was good when based on speech timing and (B) lexico-semantic features alone, and (C) excellent when combining both modalities via early fusion. The bottom row shows correlations between actual and predicted MMSE scores. Associations were significant and moderately strong for all 3 feature sets. AD: Alzheimer disease; AUC: area under the curve; HC: healthy control; MMSE: Mini-Mental State Examination; RMSE: root mean squared error; ROC: receiver operating characteristic curve.
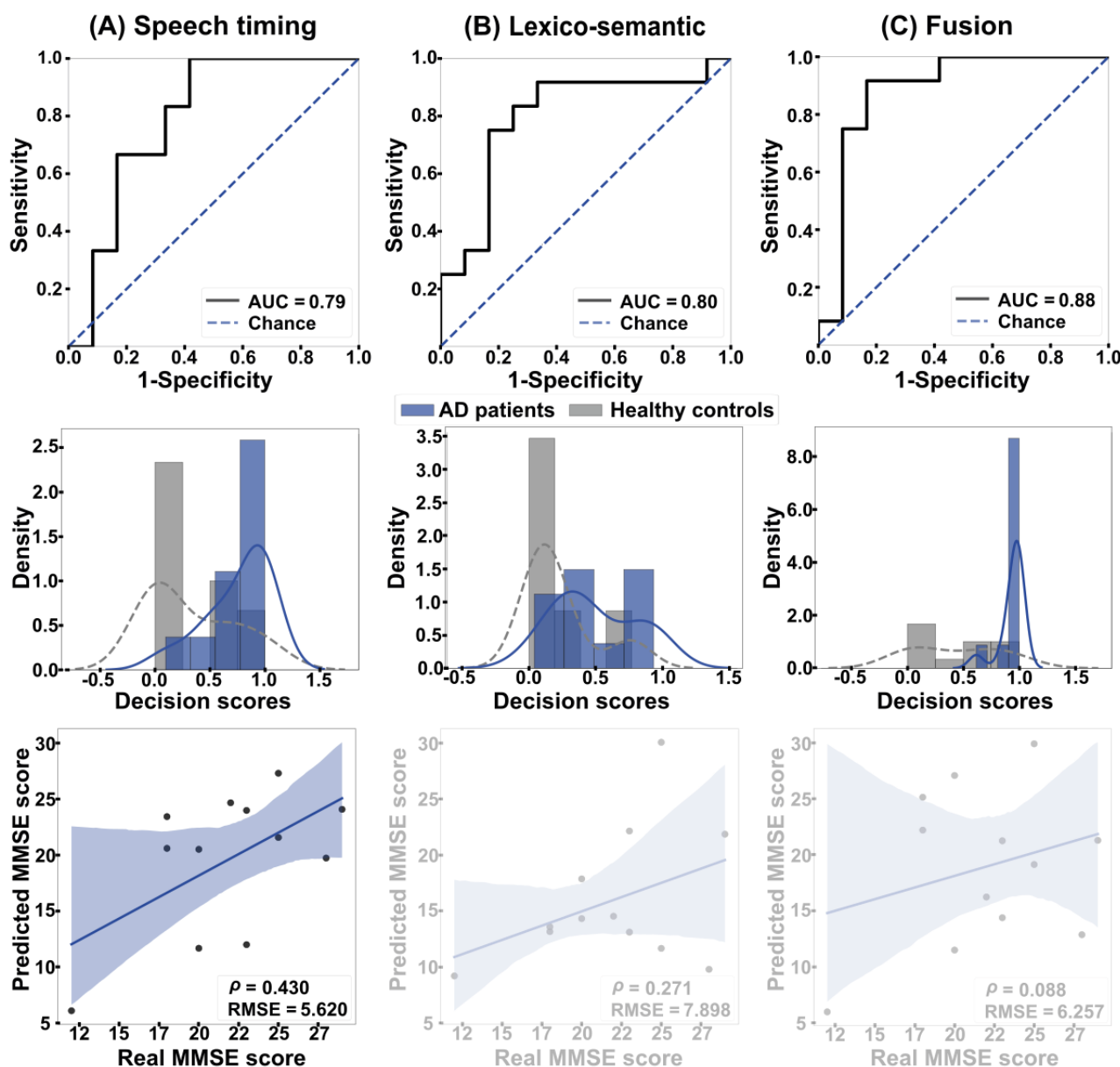


## Between-Language Results

Between-language (English-to-Spanish) analyses (Figure 3) yielded maximal discrimination for the speech timing classifier, based on stop words (AUC=0.75, Figure 3A). Lower outcomes were obtained for lexico-semantic features, considering semantic information (AUC=0.64, Figure 3B), and the fusion of both dimensions (AUC=0.65, Figure 3C). The speech timing model performed significantly better than lexico-semantic features ($P$=.015) and the fusion of both dimensions ($P$=.005), while the latter 2 models did not differ significantly ($P$=.90). The MMSE scores of Spanish-speaking patients correlated significantly with those predicted by between-language regressions based on speech timing features ($\rho$=0.389, $P$<.001, RMSE=6.248). Correlations were not significant for regressions based on lexico-semantic features ($\rho$=–0.306, $P$=.05, RMSE=6.189) or the combination of both feature sets ($\rho$=0.042, $P$=.79, RMSE=6.470).

**Figure 3.** Between-language results. The ROC curves (top row) and the classification scores (middle row) show outcomes for the classification between AD patients and HCs in the English-to-Spanish setting. (A) Discrimination was good when based on speech timing and (B) poor when based on either lexico-semantic features or (C) the combination of both dimensions via early fusion. The bottom row shows correlations between actual and predicted MMSE scores. (A) A strong, significant association emerged when using speech timing features, while (B) nonsignificant correlations were obtained when using lexico-semantic features or (C) the combination of both dimensions. AD: Alzheimer disease; AUC: area under the curve; HC: healthy control; MMSE: Mini-Mental State Examination; RMSE: root mean squared error; ROC: receiver operating characteristic curve.
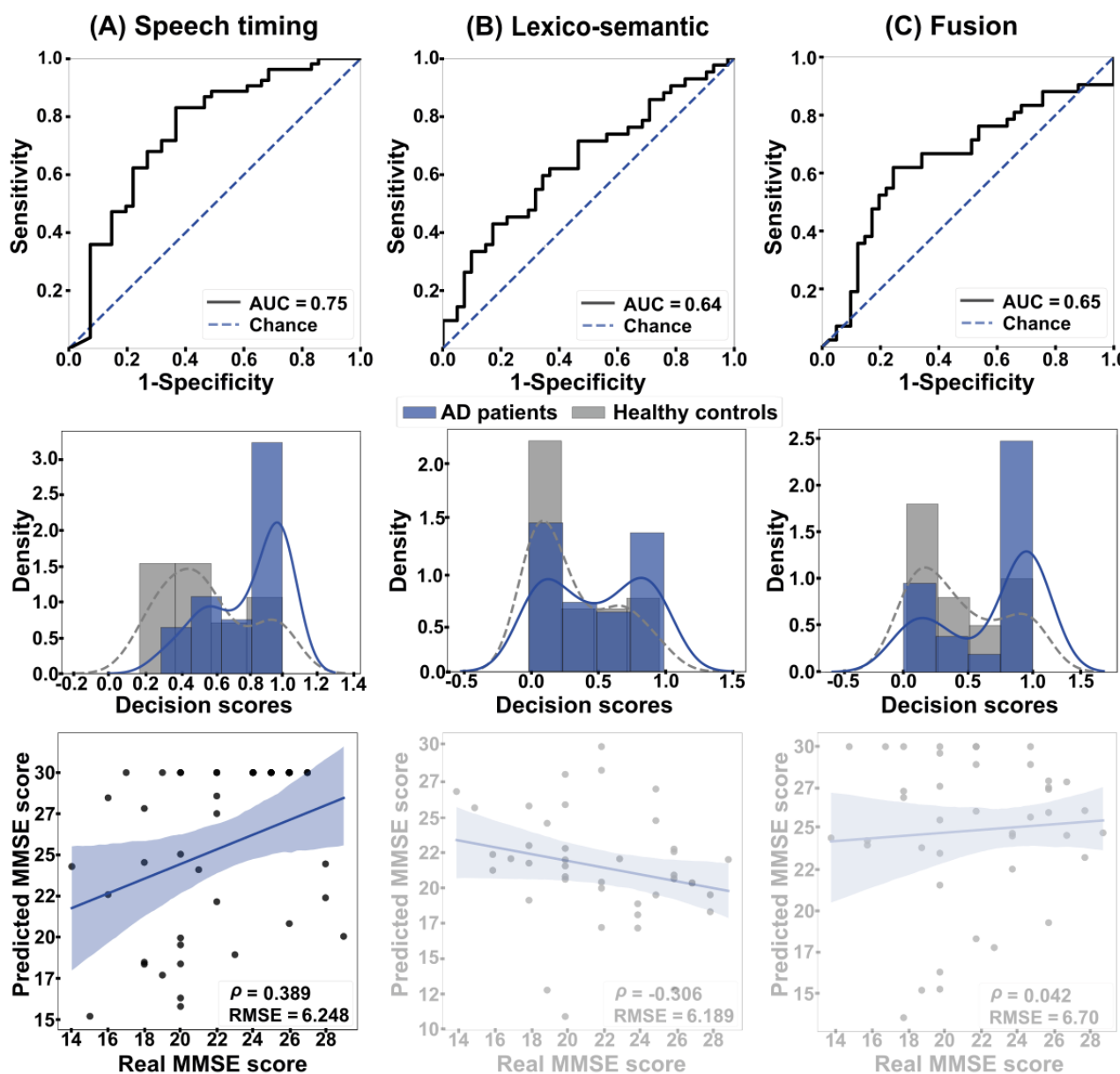


To benchmark against commonly used classifiers, we rerun our cross-linguistic analysis using a support vector machine and a support vector regressor. For the support vector machine, the best performance in Spanish was obtained with lexico-semantic features (AUC=0.603), while speech timing and combinations of features performed worse (AUC≤0.458). For the support vector regressor, correlations between predictions and true labels were very low and nonsignificant across feature sets (eg, lexico-semantic features: ρ=−0.075, *P*=.64; speech timing: ρ=−0.008, *P*=.96; combination of features: ρ=−0.131, *P*=.499), with RMSEs ≥4.46.

## *Discussion*

This study used interpretable ASLA features to identify persons with AD in within- and between-language settings. Trained on data from English speakers, our models detected English-speaking patients through timing and lexico-semantic features (especially when combined), with timing features yielding the best prediction of MMSE scores. Conversely, testing on Spanish speakers showed that only speech timing features were useful for identifying patients and capturing their MMSE scores. These findings inform the quest for clinically relevant and cross-linguistically valid ASLA features, as discussed below.

Within-language classification reached AUC values of 0.79 and 0.80 when based on speech timing and lexico-semantic features, respectively. These classification scores resemble those of other acoustic [17,24,33] and linguistic [18,64,88] studies while meeting the requirements of interpretability. As proposed in previous AD research, longer pauses and syllables likely reflect reduced word retrieval efficiency [89]. Extended silences and phonated segments are typically observed when intended words are not easily retrieved, either because they are not optimally entrenched in memory or because reduced contextual awareness complicates decisions among other lexical candidates. Compatibly, the observed lexico-semantic alterations replicate previous findings from AD research [7,30]. These include high semantic variability and low semantic granularity, pointing to greater variance in conceptual proximity across words and reduced conceptual precision [7,30]. Such patterns also seem absent in other neurodegenerative disorders [7,30], highlighting their role as distinct candidate markers of AD.

Of note, joint analysis of both dimensions yielded an AUC of 0.88, corroborating that within-language AD detection is higher for multimodal (acoustic plus linguistic) than unimodal (eg, linguistic only) ASLA approaches [8,31]. Importantly, these results approximate others derived from uninterpretable audio- and text-level embeddings (eg, Wav2vec and transformer-based models such as BERT) [60,90], suggesting that focus on interpretable ASLA features does not entail a loss of sensitivity. Briefly, the integration of strategic acoustic and linguistic features can optimize AD identification while illuminating patients' cognitive dysfunctions.

Further, timing features successfully captured patients' MMSE scores in the within-language setting. This replicates previous works based on timing and other audio-derived features [88,91,92], underscoring the importance of speech rhythm as a candidate marker of the disorder [93]. Specific speech dimensions, then, seem related to overall cognitive status in AD—a promising finding given the vast intra- and interindividual cognitive heterogeneity across patients [94].

Additional insights were obtained through between-language analyses. Identification of AD patients was significantly better for timing features (AUC=0.75) than for lexico-semantic or fusion features (AUCs<0.65). Accordingly, we propose that low word retrieval efficiency (captured by timing metrics) would represent a general cognitive trait of AD, affecting verbal production irrespective of the patient's language. Indeed, this dimension has enabled AD detection in separate studies targeting languages as diverse as English, Spanish, and Mandarin Chinese [25,95-97]. Conversely, lower cross-linguistic generalization for lexico-semantic patterns would follow from the vast vocabulary differences across typologically distant languages [35,36]. Indeed, specific linguistic patterns may exhibit different and even opposite alterations in AD depending on the patients' language—for example, pronouns are overused by English-speaking patients and underused by patients who speak Bengali, a language with a more complex pronoun system [16]. Importantly, features yielding moderate AUCs can still support clinical decision-making when interpretability and generalizability are prioritized [98]. Briefly, timing features

seem to outperform lexico-semantic features as potential cross-linguistically valid markers of AD.

The poor cross-linguistic performance of lexico-semantic features invites reflection for further research. First, weak classification and MMSE prediction results may reflect the inadequacy of the specific features we used rather than an overarching futility of lexico-semantic measures for between-language studies. In fact, it might even be that these very features do generalize robustly between other (eg, typologically close) language pairs. Further, beyond our study's goals, cross-linguistic generalizability might be enhanced through alternative strategies, including transfer learning and model training with data from multiple languages [99].

This notion is reinforced by MMSE estimation results. In the between-language setting, MMSE scores were captured by timing features, but not by the lexico-semantic or fusion models. Compatibly, a model trained on silence features and acoustic embeddings from English-speaking patients with AD and HCs yielded good prediction of MMSE scores in a Greek-speaking sample [33]. Thus, speech timing features emerge as promising candidates not only for identifying patients but also for estimating symptom severity across typologically different languages.

Note that our approach cannot ascertain the cause of observed anomalies. In particular, speech timing anomalies may certainly reflect cognitive deficits, but they may also reflect motor speech dysfunction. While motor deficits are rarely reported in AD, they may be present in several cases [100], inviting new studies integrating relevant measures for patient stratification or covariance analysis. The same analytical approaches could be used with strategic measures (eg, picture naming response times) to test the conjecture that speech timing metrics reflect lexical retrieval effort.

Our results bear clinical implications. First, the features used here, as opposed to scores from black-box models, may enhance patient phenotyping and therapeutic decision-making. Indeed, they can be linked to specific cognitive dysfunctions that practitioners are well versed in. Second, they indicate that some, but not all, findings from Anglophone samples may be useful for assessing AD in other Indo-European language communities. The point is important because of the overrepresentation of English in ASLA studies [16,46] and the tendency to derive language-agnostic conclusions from the Pitt Corpus (the English-language dataset used here, which has been leveraged by numerous studies) [17,19,32,44,46] and wide-ranging research challenges [17]. Third, our approach rests entirely on automated methods, increasing cost-efficiency and scalability while minimizing the potential for human error. In this sense, our pipeline has been deployed in a recent version of the Toolkit to Examine Lifelike Language, a speech testing app used in several clinics worldwide [11,101,102]. This allows for its widespread implementation in actual screening, diagnostic, and monitoring scenarios—a key milestone for bridging the gap between basic research and real-world applications.

Yet, our work is not without limitations. First, the 2 datasets we used differed in terms of acquisition conditions. Although we included gold-standard audio normalization steps ensuring

comparability, it would be useful to replicate our approach with harmonized cross-linguistic protocols. Second, AD in the English dataset was established before the release of contemporary diagnostic criteria. Although we chose patients with AD-consistent MMSE scores and ensured their sociodemographic matching with same-language controls as well as Spanish-speaking groups, future works should replicate our approach with Anglophone cohorts fulfilling NINCDS-ADRDA criteria or other validated standards. Third, we acknowledge that the limited sample size, especially in the Spanish cohort, restricts statistical power and constrains our capacity to rigorously assess model generalization. Future works should aim for substantially larger groups and include models trained with Spanish speakers and tested on English speakers—thus illuminating their potential for bidirectional generalizability while enabling alternative analytical approaches, based on cross-linguistic calibration or transfer. Fourth, while we controlled for sociodemographic variables, and participants' dialects were consistent within language groups, our datasets lacked data on education quality, cultural communication norms, and other factors influencing speech production. Future studies should be strategically designed to tackle these factors.

Fifth, we lacked data on when and how examiners intervened during participants' descriptions. While all examiners' segments were removed, future works should contemplate the potential impact of this factor. Sixth, our cross-sectional design reveals features that are sensitive to patients' current cognitive status, but moot on their clinical trajectories. Longitudinal datasets should be leveraged to examine whether these features can predict AD progression. Seventh, though vital for enhancing VAD and downstream processing [7,17], denoising in our preprocessing pipeline might have introduced speech artifacts. These were mitigated through manual revision of all processed audio files and careful tuning of the algorithm's threshold parameters to minimize unintended distortions. Yet, future

studies with more controlled acoustic conditions (eg, in recording booths) could replicate our approach without applying denoising. Eighth, while our 0-shot framework provides insights into the cross-linguistic generalizability of interpretable features, we recognize that performance could likely be improved through few-shot learning strategies, particularly if small amounts of target-language data are available [16,17]. Prior studies leveraging embeddings have demonstrated such gains [32], though often at the cost of transparency. Future research should explore hybrid approaches that incorporate limited target-language samples to enhance model robustness while preserving clinical interpretability.

Finally, note that we aimed to examine whether a model trained in 1 language can generalize to another using interpretable features. Thus, we avoided multilingual large language models or Wav2vec-style embeddings due to their lack of clinical interpretability. Likewise, we did not aim to train a model with interpretable features from both languages, as this would have hindered insights into cross-linguistic generalizability. Such strategies could be leveraged in further research with different goals. Here, evidence that an English-only model can be applied to a lower-resource language helps in understanding which mainstream (English-based) results might hold promise for global speech and language frameworks. Thus, future works should compare our target features with embedding-based features to establish whether interpretability involves a trade-off with sensitivity.

In sum, this study reveals 2 novel empirical patterns for ASLA research on dementia. Interpretable timing and lexico-semantic features support AD detection and cognitive decline estimation among English speakers, but only timing features generalize well from English to Spanish speakers. This suggests that timing features are more language-agnostic than lexico-semantic features. Further work can aid the search for cross-linguistic and language-specific ASLA markers of brain health.

## Acknowledgments

## Data Availability

The demographic information and feature matrices extracted for both datasets are available online [103].

## Authors' Contributions

PAP-T handled work on the methodology, software, and visualization, curated the data, and wrote the original draft. FJF and GP worked on the methodology, formal analysis, writing of the original draft, and visualization. BLT, JdL, EN, MS, AM, and MLG-T reviewed and edited the writing. AS acquired resources and curated the data. AI acquired resources, and reviewed and edited the writing. JRO-A carried out work for the methodology, software, and visualization, and wrote the original draft. AG put effort into the conceptualization, methodology, visualization, acquiring resources, reviewing and editing the writing, supervising this study, administering this project, and acquiring funding for this study.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Lists, hyperparameter tuning, and results.
[DOCX File , 22 KB-Multimedia Appendix 1]

## References

1.  Ozel‑Kizil ET, Bastug G, Kirici S. Semantic and episodic memory performances of patients with Alzheimer's disease and minor neurocognitive disorder. Alzheimer's Dementia. 2020;16(S6):e039310. [doi: 10.1002/alz.039310]
2.  Li X, Feng X, Sun X, Hou N, Han F, Liu Y. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990-2019. Front Aging Neurosci. 2022;14:937486. [FREE Full text] [doi: 10.3389/fnagi.2022.937486] [Medline: 36299608]
3.  Nandi A, Counts N, Chen S, Seligman B, Tortorice D, Vigo D, et al. Global and regional projections of the economic burden of Alzheimer's disease and related dementias from 2019 to 2050: a value of statistical life approach. EClinicalMedicine. 2022;51:101580. [FREE Full text] [doi: 10.1016/j.eclinm.2022.101580] [Medline: 35898316]
4.  Schwarzinger M, Dufouil C. Forecasting the prevalence of dementia. Lancet Public Health. 2022;7(2):e94-e95. [doi: 10.1016/s2468-2667(21)00277-2]
5.  Nichols E, Vos T. The estimation of the global prevalence of dementia from 1990‑2019 and forecasted prevalence through 2050: an analysis for the global burden of disease (GBD) study 2019. Alzheimer's Dementia. 2021;17(s10):e051496. [doi: 10.1002/alz.051496]
6.  Eyigoz E, Mathur S, Santamaria M, Cecchi G, Naylor M. Linguistic markers predict onset of Alzheimer's disease. EClinicalMedicine. 2020;28:100583. [FREE Full text] [doi: 10.1016/j.eclinm.2020.100583] [Medline: 33294808]
7.  Ferrante FJ, Migeot J, Birba A, Amoruso L, Pérez G, Hesse E, et al. Multivariate word properties in fluency tasks reveal markers of Alzheimer's dementia. Alzheimer's Dementia. 2024;20(2):925-940. [FREE Full text] [doi: 10.1002/alz.13472] [Medline: 37823470]
8.  König A, Satt A, Sorin A, Hoory R, Toledo-Ronen O, Derreumaux A, et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimer's Dementia (Amst). 2015;1(1):112-124. [FREE Full text] [doi: 10.1016/j.dadm.2014.11.012] [Medline: 27239498]
9.  Boschi V, Catricalà E, Consonni M, Chesi C, Moro A, Cappa SF. Connected speech in neurodegenerative language disorders: a review. Front Psychol. 2017;8:269. [FREE Full text] [doi: 10.3389/fpsyg.2017.00269] [Medline: 28321196]
10. García AM, Escobar-Grisales D, Correa JCV, Bocanegra Y, Moreno L, Carmona J, et al. Detecting Parkinson's disease and its cognitive phenotypes via automated semantic analyses of action stories. NPJ Parkinson's Dis. 2022;8(1):163. [FREE Full text] [doi: 10.1038/s41531-022-00422-8] [Medline: 36434017]
11. García AM, Johann F, Echegoyen R, Calcaterra C, Riera P, Belloli L, et al. Toolkit to examine lifelike language (TELL): an app to capture speech and language markers of neurodegeneration. Behav Res Methods. 2024;56(4):2886-2900. [FREE Full text] [doi: 10.3758/s13428-023-02240-z] [Medline: 37759106]
12. Laske C, Sohrabi HR, Frost SM, López-de-Ipiña K, Garrard P, Buscema M, et al. Innovative diagnostic tools for early detection of Alzheimer's disease. Alzheimer's Dementia. 2015;11(5):561-578. [doi: 10.1016/j.jalz.2014.06.004] [Medline: 25443858]
13. Al-Hameed S, Benaissa M, Christensen H, Mirheidari B, Blackburn D, Reuber M. A new diagnostic approach for the identification of patients with neurodegenerative cognitive complaints. PLoS One. 2019;14(5):e0217388. [FREE Full text] [doi: 10.1371/journal.pone.0217388] [Medline: 31125389]
14. Jonell P, Moëll B, Håkansson K, Henter GE, Kucherenko T, Mikheeva O, et al. Multimodal capture of patient behaviour for improved detection of early dementia: clinical feasibility and preliminary results. Front Comput Sci. 2021;3:642633. [doi: 10.3389/fcomp.2021.642633]

15. Riley KP, Snowdon DA, Desrosiers MF, Markesbery WR. Early life linguistic ability, late life cognitive function, and neuropathology: findings from the Nun Study. Neurobiol Aging. 2005;26(3):341-347. [doi: 10.1016/j.neurobiolaging.2004.06.019] [Medline: 15639312]

16. García AM, de Leon J, Tee BL, Blasi DE, Gorno-Tempini ML. Speech and language markers of neurodegeneration: a call for global equity. Brain. 2023;146(12):4870-4879. [FREE Full text] [doi: 10.1093/brain/awad253] [Medline: 37497623]

17. Luz S, Haider F, de la Fuente Garcia S, Fromm D, MacWhinney B. Editorial: Alzheimer's dementia recognition through spontaneous speech. Front Comput Sci. 2021;3:780169. [FREE Full text] [doi: 10.3389/fcomp.2021.780169] [Medline: 35291512]

18. Calzà L, Gagliardi G, Rossini Favretti R, Tamburini F. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. Comput Speech Lang. 2021;65:101113. [doi: 10.1016/j.csl.2020.101113]

19. Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. J Alzheimer's Dis. 2015;49(2):407-422. [doi: 10.3233/jad-150520]

20. Hu Z, Wang Z, Jin Y, Hou W. VGG-TSwinformer: transformer-based deep learning model for early Alzheimer's disease prediction. Comput Methods Programs Biomed. 2023;229:107291. [doi: 10.1016/j.cmpb.2022.107291] [Medline: 36516516]

21. Liu N, Luo K, Yuan Z, Chen Y. A transfer learning method for detecting Alzheimer's disease based on speech and natural language processing. Front Public Health. 2022;10:772592. [FREE Full text] [doi: 10.3389/fpubh.2022.772592] [Medline: 35493375]

22. García AM, Carrillo F, Orozco-Arroyave JR, Trujillo N, Vargas Bonilla JF, Fittipaldi S, et al. How language flows when movements don't: an automated analysis of spontaneous discourse in Parkinson's disease. Brain Lang. 2016;162:19-28. [doi: 10.1016/j.bandl.2016.07.008] [Medline: 27501386]

23. Ivanova O, Meilán JJG, Martínez-Sánchez F, Martínez-Nicolás I, Llorente TE, González NC. Discriminating speech traits of Alzheimer's disease assessed through a corpus of reading task for Spanish language. Comput Speech Lang. 2022;73:101341. [doi: 10.1016/j.csl.2021.101341]

24. Liu J, Fu F, Li L, Yu J, Zhong D, Zhu S, et al. Efficient pause extraction and encode strategy for Alzheimer's disease detection using only acoustic features from spontaneous speech. Brain Sci. 2023;13(3):477. [FREE Full text] [doi: 10.3390/brainsci13030477] [Medline: 36979287]

25. Pastoriza-Domínguez P, Torre I, Diéguez-Vide F, Gómez-Ruiz I, Geladó S, Bello-López J, et al. Speech pause distribution as an early marker for Alzheimer's disease. Speech Commun. 2022;136:107-117. [doi: 10.1016/j.specom.2021.11.009]

26. Pérez-Toro PA, Arias-Vergara T. Multilingual speech and language analysis for the assessment of mild cognitive impairment: outcomes from the taukadial challenge. Proc Interspeech. 2024;2024:982-986. [doi: 10.21437/Interspeech.2024-2115]

27. Krstev I, Pavikjevikj M, Toshevska M, Gievska S. Multimodal data fusion for automatic detection of Alzheimer's disease. 2022. Presented at: International Conference on Human-Computer Interaction; June 26-July 1, 2022:79-94; Gothenburg, Sweden.

28. Robin J, Xu M, Balagopalan A, Novikova J, Kahn L, Oday A, et al. Automated detection of progressive speech changes in early Alzheimer's disease. Alzheimer's Dementia (Amst). 2023;15(2):e12445. [FREE Full text] [doi: 10.1002/dad2.12445] [Medline: 37361261]

29. Saranpää AM, Kivisaari SL, Salmelin R, Krumm S. Moving in semantic space in prodromal and very early Alzheimer's disease: an item-level characterization of the semantic fluency task. Front Psychol. 2022;13:777656. [FREE Full text] [doi: 10.3389/fpsyg.2022.777656] [Medline: 35265001]

30. Sanz C, Carrillo F, Slachevsky A, Forno G, Gorno Tempini ML, Villagra R, et al. Automated text-level semantic markers of Alzheimer's disease. Alzheimer's Dementia (Amst). 2022;14(1):e12276. [FREE Full text] [doi: 10.1002/dad2.12276] [Medline: 35059492]

31. Matthews KA, Xu W, Gaglioti AH, Holt JB, Croft JB, Mack D, et al. Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015-2060) in adults aged ≥65 years. Alzheimers Dement. 2019;15(1):17-24. [FREE Full text] [doi: 10.1016/j.jalz.2018.06.3063] [Medline: 30243772]

32. Pérez-Toro PA, Klumpp P, Hernandez A, Arias-Vergara T, Lillo P. Alzheimer's Detection From English to Spanish Using Acoustic and Linguistic Embeddings. Incheon, Korea. Interspeech; 2022:2483-2487.

33. Mei K, Ding X, Liu Y, Guo Z, Xu F. The ustc system for Adress-m Challenge. 2023. Presented at: ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 June 10:1-2; Rhodes Island, Greece. [doi: 10.1109/icassp49357.2023.10094714]

34. Luz S, Haider F, Fromm D, Lazarou I, Kompatsiaris I, MacWhinney B. Multilingual Alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge. 2023. Presented at: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); June 04-June 10, 2023:05562; Rhodes, Greece. [doi: 10.1109/icassp49357.2023.10433923]

35. Blasi DE, Henrich J, Adamou E, Kemmerer D, Majid A. Over-reliance on English hinders cognitive science. Trends Cogn Sci. 2022;26(12):1153-1170. [FREE Full text] [doi: 10.1016/j.tics.2022.09.015] [Medline: 36253221]

36. Kemmerer D. Concepts in the Brain: The View from Cross-Linguistic Diversity. Oxford, England. Oxford University Press; 2019.

37. Folstein MF, Robins LN, Helzer JE. The mini-mental state examination. Arch Gen Psychiatry. 1983;40(7):812. [doi: 10.1001/archpsyc.1983.01790060110016] [Medline: 6860082]

38. Ding K, Chetty M, Noori Hoshyar A, Bhattacharya T, Klein B. Speech based detection of Alzheimer's disease: a survey of AI techniques, datasets and challenges. Artif Intell Rev. 2024;57(12):325. [doi: 10.1007/s10462-024-10961-6]

39. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. 1975;12(3):189-198. [doi: 10.1016/0022-3956(75)90026-6] [Medline: 1202204]

40. Shakeri A, Farmanbar M. Natural language processing in Alzheimer's disease research: systematic review of methods, data, and efficacy. Alzheimer's Dementia (Amst). 2025;17(1):e70082. [doi: 10.1002/dad2.70082] [Medline: 39935888]

41. Parra MA, Baez S, Allegri R, Nitrini R, Lopera F, Slachevsky A, et al. Dementia in Latin America: assessing the present and envisioning the future. Neurology. 2018;90(5):222-231. [FREE Full text] [doi: 10.1212/WNL.0000000000004897] [Medline: 29305437]

42. Salgado AMR, Acosta I, Kim DJ, Zitser J, Sosa AL, Acosta D, et al. Prevalence and impact of neuropsychiatric symptoms in normal aging and neurodegenerative syndromes: a population-based study from Latin America. Alzheimer's Dementia. 2023;19(12):5730-5741. [FREE Full text] [doi: 10.1002/alz.13384] [Medline: 37427840]

43. Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge. Brno, Czechia. Interspeech: ISCA; 2021.

44. Zhu Y, Obyat A, Liang X, Batsis JA, Roth RM. WavBERT: Exploiting semantic and non-semantic speech using Wav2vec and BERT for dementia detection. Interspeech. 2021;2021:3790-3794. [FREE Full text] [doi: 10.21437/interspeech.2021-332] [Medline: 37063977]

45. Becker JT, Boller F, Lopez OL, Saxton J, McGonigle KL. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. Arch Neurol. 1994;51(6):585-594. [doi: 10.1001/archneur.1994.00540180063015] [Medline: 8198470]

46. de la Fuente Garcia S, Ritchie CW, Luz S. Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. J Alzheimer's Dis. 2020;78(4):1547-1574. [doi: 10.3233/jad-200888]

47. Becker JT, Huff FJ, Nebes RD, Holland A, Boller F. Neuropsychological function in Alzheimer's disease: pattern of impairment and rates of progression. Arch Neurol. 1988;45(3):263-268. [doi: 10.1001/archneur.1988.00520270037018] [Medline: 3341951]

48. Lopez OL, Becker JT, Brenner RP, Rosen J, Bajulaiye OI, Reynolds CF. Alzheimer's disease with delusions and hallucinations: neuropsychological and electroencephalographic correlates. Neurology. 1991;41(6):906. [doi: 10.1212/wnl.41.6.906] [Medline: 2046938]

49. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of department of health and human services task force on Alzheimer's disease. Neurology. 1984;34(7):939. [doi: 10.1212/wnl.34.7.939] [Medline: 6610841]

50. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Washington DC. American Psychiatric Association; 1987.

51. Santamaría-García H, Baez S, Reyes P, Santamaría-García JA, Santacruz-Escudero JM, Matallana D, et al. A lesion model of envy and Schadenfreude: legal, deservingness and moral dimensions as revealed by neurodegeneration. Brain. 2017;140(12):3357-3377. [FREE Full text] [doi: 10.1093/brain/awx269] [Medline: 29112719]

52. Torralva T, Roca M, Gleichgerrcht E, LÓPEZ P, MANES F. INECO frontal screening (IFS): a brief, sensitive, and specific tool to assess executive functions in dementia–CORRECTED VERSION. J Int Neuropsychol Soc. 2009;15(5):777-786. [doi: 10.1017/s1355617709990415]

53. Quiroga P, Albala C, Klaasen G. [Validation of a screening test for age associated cognitive impairment, in Chile]. Rev Med Chile. 2004;132(4):467-478. [doi: 10.4067/s0034-98872004000400009] [Medline: 15382519]

54. Ferrante FJ, Grisales DE, López MF, Lopes da Cunha P, Sterpin LF, Vonk JM, et al. Cognitive phenotyping of Parkinson's disease patients via digital analysis of spoken word properties. Mov Disord. 2025. [doi: 10.1002/mds.70005] [Medline: 40798929]

55. Gattei CA, Ferrante FJ, Sampedro B, Sterpin L, Abusamra V, Abusamra L, et al. Semantic memory navigation in HIV: conceptual associations and word selection patterns. Clin Neuropsychologist. 2025;39(6):1598-1614. [doi: 10.1080/13854046.2024.2417844] [Medline: 39439192]

56. Lopes da Cunha P, Ruiz F, Ferrante F, Sterpin LF, Ibáñez A, Slachevsky A, et al. Automated free speech analysis reveals distinct markers of Alzheimer's and frontotemporal dementia. PLOS One. 2024;19(6):e0304272. [FREE Full text] [doi: 10.1371/journal.pone.0304272] [Medline: 38843210]

57. Toro-Hernández FD, Migeot J, Marchant N, Olivares D, Ferrante F, González-Gómez R, et al. Neurocognitive correlates of semantic memory navigation in Parkinson's disease. NPJ Parkinsons Dis. 2024;10(1):15. [FREE Full text] [doi: 10.1038/s41531-024-00630-4] [Medline: 38195756]

58. Pérez-Toro PA, Vásquez-Correa JC, Arias-Vergara T, Klumpp P. Acoustic and linguistic analyses to assess early-onset and genetic Alzheimer's disease. 2021. Presented at: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 13, 2021:8338-8342; ON, Canada. [doi: 10.1109/icassp39728.2021.9414009]

59. Goodglass H, Kaplan E, Barressi B. Cookie Theft Picture, Boston Diagnostic Aphasia Examination. Philadelphia, PA. Lea Febiger; 1983.

60. Balagopalan A, Eyre B, Robin J, Rudzicz F, Novikova J. Comparing pre-trained and feature-based models for prediction of Alzheimer's disease based on speech. Front Aging Neurosci. 2021;13:635945. [FREE Full text] [doi: 10.3389/fnagi.2021.635945] [Medline: 33986655]

61. Pérez-Toro PA, Bayerl SP, Arias-Vergara T. Influence of the interviewer on the automatic assessment of Alzheimer's disease in the context of the ADReSSo challenge. Interspeech. 2021:3785-3789. [doi: 10.21437/interspeech.2021-1589]

62. Berube SK, Goldberg E, Sheppard SM, Durfee AZ, Ubellacker D, Walker A, et al. An analysis of right hemisphere stroke discourse in the modern Cookie Theft picture. Am J Speech Lang Pathol. 2022;31(5S):2301-2312. [doi: 10.1044/2022_ajslp-21-00294]

63. Cummings L. Describing the Cookie Theft picture: sources of breakdown in Alzheimer's dementia. Pragmatics. 2019;10(2):153-176. [doi: 10.1075/ps.17011.cum]

64. Campbell EL, Mesía RY, Docío-Fernández L, García-Mateo C. Paralinguistic and linguistic fluency features for Alzheimer's disease detection. Comput Speech Lang. 2021;68:101198. [doi: 10.1016/j.csl.2021.101198]

65. García AM, Arias-Vergara T, C Vasquez-Correa J, Nöth E, Schuster M, Welch AE, et al. Cognitive determinants of dysarthria in Parkinson's disease: an automated machine learning approach. Mov Disord. 2021;36(12):2862-2873. [doi: 10.1002/mds.28751] [Medline: 34390508]

66. Barras B. SoX: Sound eXchange. Flash informatique. 2012;(9):3-6. [doi: 10.4324/9780080459684-14]

67. Schröter H, Rosenkranz T, Escalante-B AN, Aubreville M, Maier A. Clcnet: Deep learning-based noise reduction for hearing aids using complex linear coding. 2020. Presented at: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020 May 04:6949-6953; Canada. [doi: 10.1109/icassp40776.2020.9053563]

68. Mak MW, Yu HB. A study of voice activity detection techniques for NIST speaker recognition evaluations. Comput Speech Lang. 2014;28(1):295-313. [doi: 10.1016/j.csl.2013.07.003]

69. Rabiner L, Schafer R. Theory and Applications of Digital Speech Processing. New Jersey, US. Prentice Hall Press; 2010.

70. Skodda S, Schlegel U. Speech rate and rhythm in Parkinson's disease. Mov Disord. 2008;23(7):985-992. [doi: 10.1002/mds.21996] [Medline: 18383114]

71. Niebuhr O, Siegert I. A digital "flat affect"? Popular speech compression codecs and their effects on emotional prosody. Front Commun. 2023;8:972182. [doi: 10.3389/fcomm.2023.972182]

72. Fuchs R, Maxwell O. The effects of mp3 compression on acoustic measurements of fundamental frequency and pitch range. Proc Speech Prosody. 2016;2016:523-527. [doi: 10.21437/speechprosody.2016-107]

73. Panayotov V, Chen G, Povey D, Khudanpur S. Librispeech: an asr corpus based on public domain audio books. IEEE; 2015. Presented at: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2015 Apr 25:5206-5210; Canada. [doi: 10.1109/icassp.2015.7178964]

74. Kisler T, Reichel U, Schiel F. Multilingual processing of speech via web services. Comput Speech Lang. 2017;45:326-347. [doi: 10.1016/j.csl.2017.01.005]

75. Paschen L, Delafontaine F, Draxler C, Fuchs S, Stave M, Seifart F. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). 2020. Presented at: Proceedings of the Twelfth Language Resources and Evaluation Conference; 2020 May 10:2657-2666; France.

76. Strunk J, Schiel F, Seifart F. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. 2014. Presented at: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); May 10, 2014:3940-3947; Reykjavik, Iceland.

77. MacWhinney B. The CHILDES Project: The Database. London. Psychology Press; 2000.

78. Turn speech into text using Google AI. Google Cloud. URL: https://cloud.google.com/speech-to-text [accessed 2025-09-24]

79. García AM, DeLeon J, Tee BL. Della Sala S, editor. Neurodegenerative Disorders of Speech and Language: Non-Language-Dominant Diseases. Oxford, England. Elsevier; 2022:66-80.

80. Humbird KD, Peterson JL, Mcclarren RG. Deep neural network initialization with decision trees. IEEE Trans Neural Netw Learn Syst. 2019;30(5):1286-1295. [doi: 10.1109/tnnls.2018.2869694]

81. Rodríguez-Salas D, Mürschberger N, Ravikumar N, Seuret M, Maier A. Mapping ensembles of trees to sparse, interpretable multilayer perceptron networks. SN Comput Sci. 2020;1:252. [doi: 10.1007/s42979-020-00268-y]

82. Ying X. An overview of overfitting and its solutions. IOP Publishing; 2019. Presented at: Journal of Physics: Conference Series; May 10, 2019:022022; Bristol, England.

83. Goodfellow I, Bengio Y, Courville A. Regularization for deep learning. Deep Learning. 2016:216-261. [doi: 10.1007/978-1-4842-3790-8_5]

84. Hagiwara KJN. Regularization learning, early stopping and biased estimator. Neurocomputing. 2002;48(1-4):937-955. [doi: 10.1016/s0925-2312(01)00681-6]

85. Cui Y, Jia M, Lin TY, Song Y, Belongie S. Class-balanced loss based on effective number of samples. 2019. Presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 20, 2019:9268-9277; USA. [doi: 10.1109/cvpr.2019.00949]

86. GitHub. URL: https://github.com/PauPerezT/Crosslingual_AD_Descriptors [accessed 2025-10-04]

87. Nautsch A, Jasserand C, Kindt E, Todisco M, Trancoso I, Evans NJPI. The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps Towards a Common Understanding. Graz, Austria. INTERSPEECH; 2019.

88. Shah Z, Sawalha J, Tasnim M, Qi S, Stroulia E, Greiner R. Learning language and acoustic models for identifying Alzheimer's dementia from speech. Front Comput Sci. 2021;3:624659. [doi: 10.3389/fcomp.2021.624659]

89. Kavé G, Goral M. Word retrieval in connected speech in Alzheimer's disease: a review with meta-analyses. Aphasiology. 2018;32(1):4-26. [doi: 10.1080/02687038.2017.1338663]

90. Zhu Y, Liang X, Batsis JA, Roth RM. Exploring deep transfer learning techniques for Alzheimer's dementia detection. Front Comput Sci. 2021;3:624683. [FREE Full text] [doi: 10.3389/fcomp.2021.624683] [Medline: 34046588]

91. Fu Z, Haider F, Luz S. Predicting mini-mental status examination scores through paralinguistic acoustic features of spontaneous speech. IEEE; 2020. Presented at: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); July 20, 2020:5548-5552; Canada. [doi: 10.1109/embc44109.2020]

92. Sun L, Zheng J, Li J, Qian C. Exploring MMSE score prediction model based on spontaneous speech. SEKE. 2022:347-350. [doi: 10.18293/seke2022-062]

93. Bertola L, Mota NB, Copelli M, Rivero T, Diniz BS, Romano-Silva MA, et al. Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, mild cognitive impairment and normal elderly controls. Front Aging Neurosci. 2014;6:185. [FREE Full text] [doi: 10.3389/fnagi.2014.00185] [Medline: 25120480]

94. Qiu Y, Jacobs DM, Messer K, Salmon DP, Feldman HH. Cognitive heterogeneity in probable Alzheimer disease: clinical and neuropathologic features. Neurology. 2019;93(8):e778-e790. [doi: 10.1212/wnl.0000000000007967]

95. Lofgren M, Hinzen W. Erratum to 'Breaking the flow of thought: increase of empty pauses in the connected speech of people with mild and moderate Alzheimer's disease'. J Commun Disord. 2023;101:106299. [FREE Full text] [doi: 10.1016/j.jcomdis.2022.106299] [Medline: 36628800]

96. Qiao Y, Xie X, Lin G, Zou Y, Chen S, Ren R, et al. Computer-assisted speech analysis in mild cognitive impairment and Alzheimer's disease: a pilot study from Shanghai, China. J Alzheimer's Dis. 2020;75(1):211-221. [doi: 10.3233/jad-191056]

97. Szatloczki G, Hoffmann I, Vincze V, Kalman J, Pakaski M. Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. Front Aging Neurosci. 2015;7:195. [FREE Full text] [doi: 10.3389/fnagi.2015.00195] [Medline: 26539107]

98. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian J Intern Med. 2013;4(2):627-635. [FREE Full text] [Medline: 24009950]

99. García AM, Tee BL, de Leon J, Alladi S, Gorno-Tempini ML. Language disorders in understudied languages: Opening editorial. Cortex. 2025;188:A1-A6. [doi: 10.1016/j.cortex.2025.05.009] [Medline: 40484550]

100. Al-Harrasi AM, Iqbal E, Tsamakis K, Lasek J, Gadelrab R, Soysal P, et al. Motor signs in Alzheimer's disease and vascular dementia: detection through natural language processing, co-morbid features and relationship to adverse outcomes. Exp Gerontol. 2021;146:111223. [doi: 10.1016/j.exger.2020.111223] [Medline: 33450346]

101. García AM, Ferrante FJ, Pérez G, Ponferrada J, Sosa Welford A, Pelella N, et al. Toolkit to examine lifelike language v.2.0: Optimizing speech biomarkers of neurodegeneration. Dement Geriatr Cogn Disord. 2025;54(2):96-108. [FREE Full text] [doi: 10.1159/000541581] [Medline: 39348797]

102. Perez‐Toro PA, Ferrante FJ, Pérez GN, Slachevsky A, Nöth E, Schuster M, et al. A cross‐linguistic test of automated speech and language analysis for detecting Alzheimer's disease: machine learning evidence from English and Spanish speakers. In: Alzheimer's Dementia. 2024. Presented at: Alzheimer's Association International Conference; July 12, 2026; London, United Kingdom. [doi: 10.1002/alz.084717]

103. Automated speech markers of Alzheimer's dementia: a test of cross-linguistic generalizability. OSF Home. URL: https://osf.io/n4eq5/?view_only=5fae6fb60bd64033802062d2b505b868 [accessed 2025-09-24]

## Abbreviations

**AD:** Alzheimer disease

**ASLA:** automated speech and language analysis

**AUC:** area under the receiver operating characteristic curve

**BERT:** Bidirectional Encoder Representations from Transformers

**CHAT:** Codes for the Human Analysis of Transcripts

**CHILDES:** Child Language Data Exchange System

**DSM:** Diagnostic and Statistical Manual of Mental Disorders

**HC:** healthy control

**MLP:** multilayer perceptron

**MMSE:** Mini-Mental State Examination

**NINCDS-ADRDA:** National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association

**RMSE:** root-mean-squared error

**SoX:** Sound Exchange
**VAD:** voice activity detection
**XGBoost:** Extreme Gradient Boosting

XSL•FO

**RenderX**