



ADCeleb: A Longitudinal Speech Dataset from Public Figures for Early Detection of Alzheimer’s Disease

Kunxiao Gao¹, Anna Favaro², Najim Dehak², Laureano Moro-Velázquez²

¹Department of Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, USA

²Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, USA

{kgao9, afavaro1}@jhu.edu

Abstract

Previous studies on early Alzheimer’s disease (AD) detection using speech have been limited by small sample sizes, scarce prodromal phase recordings, and minimal longitudinal data. Many also rely on standardized tasks that do not reflect natural language use. To address these gaps, we introduce *ADCeleb*, a longitudinal speech dataset comprising publicly available recordings from individuals who later disclosed an AD diagnosis. It includes samples from 40 individuals with prodromal AD and 40 cognitively normal controls (CNs), matched by age and sex, spanning 1 to 10 years before diagnosis. Classification experiments using multimodal models integrating speech and text achieved 0.72 accuracy in distinguishing AD from CNs in the 6- to 10-year pre-diagnosis window and 0.80 in the 1- to 5-year pre-diagnosis window. These results highlight the potential of speech-based technologies as non-invasive tools for early AD detection in real-world settings or for triage improvement in clinical trials.

Index Terms: Alzheimer’s disease, speech, language, data collection, early detection

1. Introduction

Alzheimer’s disease (AD) is a neurodegenerative disorder characterized by progressive cognitive decline and memory loss, significantly impacting patients’ quality of life and posing a growing burden on healthcare systems globally [1]. Despite advances in diagnostic technologies, early and accurate detection remains a critical challenge. Speech and language offer a non-invasive, cost-effective approach to early AD detection, as linguistic and acoustic changes might emerge years or even decades before diagnosis [2, 3]. Early signs include word-finding difficulties, reduced lexical diversity, and alterations in syntax, semantics, discourse, fluency, and acoustics [3, 4].

Observational studies have identified early markers of AD in speech and writing. Le et al. [5] analyzed the writings of Iris Murdoch, Agatha Christie, and Phyllis Dorothy James, finding that Murdoch’s novels exhibited signs of impoverished vocabulary and syntax before her dementia diagnosis. Berisha et al. [6] found that President Reagan, later diagnosed with AD, exhibited reduced lexical diversity, using fewer unique words and more low-imageability verbs over time. His speech also showed increased fillers and nonspecific words, patterns not observed in President Bush, suggesting early signs of dementia before diagnosis. More recently, Petty et al. [7] analyzed two corpora of spontaneous speech recordings from public figures who later developed AD and matched cognitively normal individuals (CNs). The first corpus included 10 AD-CN pairs, while the second contained 9 AD-CN pairs, with recordings spanning several decades, including 30 years before diagnosis. Their findings suggest that reduced lexical diversity and changes in pronoun usage are promising

early markers of cognitive decline. While the studies provide valuable insights into early linguistic changes, they remain largely observational. Other studies on early AD detection have primarily focused on detecting Mild Cognitive Impairment (MCI) using data collected post-diagnosis. Fraser et al. [8] combined MMSE scores with linguistic features, achieving 87% accuracy in MCI detection. More advanced methods have leveraged neural models such as BERT, HuBERT, and other Transformer architectures [9–12]. Moreover, multimodal approaches combining acoustic, linguistic, and demographic features reported improved classification performance over unimodal models using single modalities in isolation [3, 13, 14].

Previous longitudinal studies examining early signs of AD were often constrained by small sample sizes and focused on limited periods. The scarcity of datasets capturing the prodromal stage further hampers the development of reliable predictive models. Existing studies predominantly rely on cross-sectional data from individuals with MCI, but the heterogeneity in disease progression limits their prognostic utility for identifying prodromal AD. Moreover, many AD speech datasets are collected in controlled settings, ensuring consistency but failing to capture the variability of natural speech. While connected speech analysis provides a more ecologically valid and multidimensional assessment of language decline, most available datasets do not reflect real-world conversational dynamics, limiting their applicability for early detection. To address these limitations, we introduce *ADCeleb*¹, a novel dataset of spontaneous speech from individuals with AD, recorded from up to 10 years before diagnosis until the year preceding the year of diagnosis (YoD). It contains recordings from 40 public figures diagnosed with AD, matched with 40 CNs based on age, sex, and ethnicity. *ADCeleb* offers a resource for analyzing prodromal speech patterns and their temporal evolution.

To address the gap in prodromal speech-based AD detection, this study leverages *ADCeleb* to establish baseline classification performance by training acoustic and linguistic non-interpretable feature-based models (NIFMs) (see Section 2). We explore the complementarity of speech and text-based representations through multimodal fusion, assessing their impact on classification accuracy. Specifically, we aim to identify the prodromal stage at which integrating acoustic and linguistic features yields the greatest improvement over single-modality models in AD detection. While interpretable models could provide deeper insights into prodromal speech changes, this study focuses on NIFMs as prior studies demonstrated the superior performance of neural embeddings compared to hand-crafted features for early detection of AD and other neurological disorders [14, 15].

¹All metadata required to compile *ADCeleb* is publicly available on Zenodo, at this link: <https://zenodo.org/records/15515841>.

2. Methods

2.1. Dataset Description

ADCeleb encompasses 5347 recordings, amounting to approximately 25 h of speech, from 80 celebrities, split evenly between 40 diagnosed with AD and 40 CNs. The speech samples were extracted from videos uploaded to YouTube. The dataset includes various accents (e.g., British, American, Canadian), professions, and age groups, annotated in the dataset. The AD and CN groups are matched based on age, sex, and ethnicity. Additionally, to ensure that speech recordings are comparable across both groups, each CN participant was matched to an AD participant by demographics and birth year (± 5 years). This ensures that both groups' speech samples reflect similar linguistic and acoustic environments, minimizing confounding factors related to historical recording quality or linguistic shifts over time. The videos in the dataset feature a limited variety of complex multi-speaker acoustic environments, primarily including settings like calm studio interviews, public conferences, TV show talk delivered to large audiences, and informal talk in daily conversation. To accurately capture spontaneous language use, we manually reviewed the recording context and included only those featuring spontaneous speech. Public addresses and scripted monologues were excluded to eliminate read speech from the analysis. Notably, these videos have been subject to background noise, such as chatter, laughter, overlapping speech, reverberation, and variations in recording equipment quality and channel conditions.

Table 1 presents the overall dataset statistics, grouping data into two time intervals relative to the YoD of individuals with AD. CN individuals are assigned to corresponding intervals based on the YoD of their matched AD counterparts. The two intervals are defined as follows: *time interval -2* includes recordings from 6 to 10 years before the YoD, while *time interval -1* includes recordings from 1 to 5 years before the YoD. Figure 1 illustrates the distributions of the two groups according to sex, ethnicity, and nationality. During data collection, more videos featuring male celebrities were identified compared to female celebrities, primarily due to the greater availability of interview footage for male celebrities. Public figures, like celebrities or politicians, who have openly disclosed an AD diagnosis may be primarily 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 [16]. Moreover, ADCeleb only features prodromal data, as the limited availability of videos from celebrities after their public AD diagnosis restricted the collection of post-diagnosis recordings.

2.2. Data Collection Pipeline

ADCeleb was created following the same pipeline as ParkCeleb² [17], a novel longitudinal speech dataset from people with Parkinson's Disease, including speech samples from up to 20 years before to 10 years after the YoD. The main stages of the data collection pipeline are outlined below.

1. **Creation of a list of people with AD.** A list of 40 English-speaking public figures diagnosed with AD was created using publicly available information, such as the Wikipedia page *Category: People with Alzheimer's disease* [18].
2. **Identification of demographics.** We then collected demographic details for each subject, including YoD, sex, age, nationality, and profession.
3. **Audio download.** For each subject, we manually searched

²<https://zenodo.org/records/13954768>

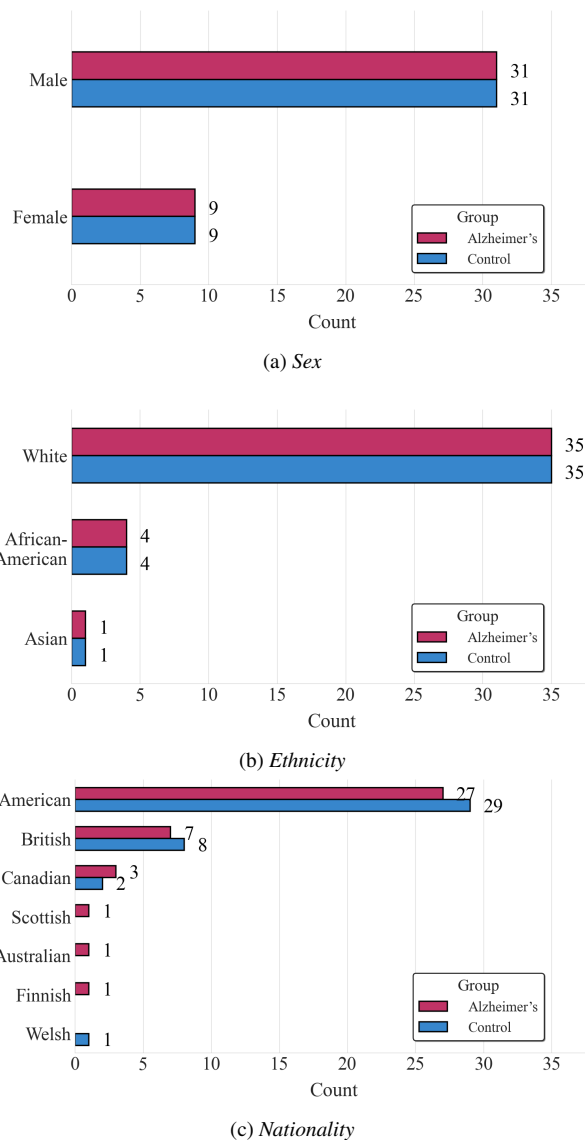


Figure 1: Demographic distributions of the dataset across (a) sex, (b) ethnicity, and (c) nationality.

for high-quality audio from television interviews and other sources on YouTube.

4. **Diarization:** WhisperX [19] was used for automatic speech recognition and speaker diarization to partition segments based on speakers' identity.
5. **Annotation of the target speaker and year.** Each speaker in the recordings was labeled as *target* or *non-target*, with intervals annotated relative to the YoD.
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.

2.3. Data Pre-processing and Feature Extraction

First, recordings were resampled at 16 kHz. Only recordings with a length greater than 8 s were considered, as we observed shorter audio segments were more likely to contain speech segments from other speakers, different from those of interest. The two

Time Interval	Group	# Speakers	# Recordings	# Recordings/Speaker	# Segments/Speaker	Segment Length (s)
-2	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
-1	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	8.01/16.62/150.00

Table 1: Dataset statistics for the AD and CN groups across time intervals -2 and -1, including the number of speakers, recordings per speaker, segments per speaker, and segment lengths (in seconds). Fields with three values represent the minimum, average, and maximum. Recordings refer to audio linked to videos, while segments denote portions containing the target speaker’s speech.

groups (i.e., AD and CN) were balanced in all the experiments, considering sample size, sex, age, and ethnicity. Second, we extracted non-interpretable acoustic and linguistic embeddings from spontaneous speech samples using foundation models. In particular, these representations are high-dimensional embeddings derived from state-of-the-art neural network models, which offer deeper insights into nuanced speech and language patterns. These models were employed purely as feature extractors, with their pre-trained weights remaining unchanged. A separate classifier was then trained on the embeddings for task-specific classification (see Section 3). Table 2 summarizes the extracted acoustic and linguistic embeddings.

3. Experimental Setup

Our experiments utilized a nested cross-validation scheme (NCV). In each iteration of the 10-fold cross-validation, the data were split so that speakers in training were never seen in testing. We applied Principal Component Analysis (PCA) for dimensionality reduction, followed by Probabilistic Linear Discriminant Analysis (PLDA) for binary classification (AD vs CN). The PCA transformation matrix was trained using the training embeddings and applied to training and testing embeddings in each iteration. The PLDA model³ was then trained on the PCA-reduced training subset. The mean embeddings of PCA-reduced AD and CN training samples were used as enrollment embeddings in PLDA scoring. A log-likelihood ratio was computed against the two enrollment embeddings for a given PCA-reduced testing embedding. This ratio was compared to the equal error rate threshold derived from the training subset. If the ratio exceeded the equal error rate threshold, the sample was classified as AD; otherwise, it was classified as CN.

Moreover, a fusion-based prediction framework was leveraged to integrate outputs from models trained using different features. Specifically, two feature groups were considered: acoustic and linguistic embeddings (see Section 2.3). Separate models were trained for each feature type, and the top three models for each type were selected based on the average F1-score from the inner loop of the NCV. For each speaker, the binary predictions (0 or 1) generated by the six selected models (three acoustic and three linguistic) on the test set 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 categorized as CN (1) and scores below the threshold categorized as AD (0). This fusion approach leverages complementary strengths across feature domains, improving robustness and generalizability, especially in noisy or small datasets. Model performance was evaluated in terms of accuracy (ACC), F1-score (F1), specificity (SPEC), sensitivity (SENS), and area under the ROC curve (AUC). Binary classification results were computed at the speaker level. To do so, each embedding underwent initial

³Implemented using SpeechBrain: <https://speechbrain.readthedocs.io/en/latest/API/speechbrain.processing.html>

Model	Description
Acoustic Embeddings	
X-vectors [20]	512-dimensional embeddings from a neural network trained on VoxCeleb datasets, capturing prosodic and fluency changes.
TRILLsson [21]	1024-dimensional embeddings from a CAP12 Conformer-based model, capturing non-lexical speech features.
Wav2Vec 2.0 [22]	768-dimensional embeddings pre-trained on LibriSpeech, capturing nuanced acoustic and phonetic features.
HuBERT [23]	768-dimensional embeddings extracted from the seventh layer of a HuBERT model, averaged over 10-second segments.
Whisper [24]	384-dimensional embeddings from the final encoder layer of the Whisper tiny model, capturing prosodic features.
Linguistic Embeddings	
XLM-RoBERTa [25]	768-dimensional embeddings capturing lexical and syntactic features.
Text2vec [26]	384-dimensional embeddings optimized for multilingual data, based on the CoSENT architecture.
Multilingual-e5-large [27]	1024-dimensional embeddings trained on multilingual data, capturing semantic and syntactic features.
LaBSE [28]	768-dimensional embeddings optimized for semantic and syntactic analysis.
DistilBERT-multilingual [29]	768-dimensional embeddings capturing syntactic and semantic relationships in multilingual data.
Distiluse-multilingual [30]	512-dimensional embeddings optimized for cross-lingual semantic analysis.
Cross-en-de-roberta [30]	768-dimensional cross-lingual embeddings optimized for English and German using transformer architecture.
BERT-multilingual [31]	768-dimensional embeddings trained on over 100 languages, capturing syntactic and semantic relationships.
Llama3 [32]	4096-dimensional embeddings trained on a multilingual corpus, capturing intricate syntactic, semantic, and contextual relationships.

Table 2: Summary of the acoustic and linguistic models used to extract embeddings for the classification tasks.

length normalization, followed by averaging across all segments corresponding to a given speaker. A final length normalization was applied to the aggregated embeddings.

4. Results and Discussion

4.1. Acoustic Modeling

The first section of Table 3 reports the performance of models leveraging acoustic embeddings, with Wav2Vec 2.0 achieving the highest accuracy: 0.67 in time interval -2 and 0.72 in time interval -1, reflecting the model’s contrastive pretraining on raw audios. The enhanced performance observed across models in the time interval closer to diagnosis aligns with the expectation that speech alterations associated with AD become more pronounced as the disease progresses. These results highlight the potential of NIFMs to detect subtle acoustic variations during the prodromal phase of AD, reinforcing their viability for early detection.

4.2. Linguistic Modeling

The second section of Table 3 presents the performance of NIFMs using linguistic embeddings, with Multilingual-e5-large achieving the highest accuracy of 0.73 in time interval -2 and Cross-en-de-roberta reaching 0.75 in time interval -1. While the improvement between these intervals is modest, it aligns with the progressive nature of the linguistic decline in AD, which becomes increasingly pronounced over time, facilitating differentiation between AD and CN groups, similar to trends observed in acoustic models. These findings underscore the effectiveness of NIFMs in detecting subtle linguistic changes associated with early cognitive decline, even during the prodromal stage, reinforcing their potential for early diagnosis and intervention. Moreover, the results suggest that linguistic representations may offer greater sensitivity and robustness than acoustic ones in prodromal AD detection, especially in the earlier stages. This supports the hypothesis that linguistic dysfunction precedes acoustic impairments in AD and dementia [33].

4.3. Fusion Based Modeling

The final section of Table 3 presents the performance of fusion-based models integrating predictions from acoustic and linguistic models. At time interval -2, the fusion model achieved an accuracy of 0.72, closely resembling the performance of the best unimodal linguistic model (0.73) in the same interval. This similarity suggests that linguistic markers alone may be sufficient for distinguishing AD from CNs in the earlier prodromal stages, as language impairments tend to emerge before noticeable speech alterations. However, at time interval -1, the fusion model outperformed both unimodal approaches, reaching an accuracy of 0.80. This improvement underscores the increasing relevance of acoustic features as individuals approach the YoD. While linguistic dysfunctions are often the earliest indicators of cognitive decline, speech production deteriorates over time, making acoustic markers more informative in later stages [33,34]. The multimodal approach leverages the complementary nature of these features, capturing a more comprehensive profile of AD-related speech and language changes.

5. Conclusion and Future Work

This study introduces ADCeleb, a novel speech dataset comprising recordings from 40 individuals with AD and 40 CNs, spanning from 10 years before diagnosis up to the year of official AD diagnosis. Through binary classification experiments, we examine the complementary role of acoustic and linguistic models in detecting prodromal AD. Our findings indicate that linguistic models outperform acoustic models in earlier stages when language impairments are more prominent than speech-related deficits. However, as individuals approach diagnosis, integrating

Feature Name	F1	ACC	AUC	SENS	SPEC
Acoustic					
Time Interval -2					
X-vectors	0.52	0.52	0.53	0.53	0.50
TRILLsson	0.53	0.53	0.56	0.57	0.50
HuBERT	0.50	0.50	0.52	0.53	0.47
Wav2vec 2.0	0.67	0.67	0.68	0.70	0.63
Whisper	0.53	0.53	0.56	0.50	0.57
Time Interval -1					
X-vectors	0.67	0.67	0.70	0.67	0.67
TRILLsson	0.65	0.65	0.69	0.63	0.67
HuBERT	0.72	0.72	0.76	0.77	0.67
Wav2vec 2.0	0.72	0.72	0.75	0.73	0.70
Whisper	0.70	0.70	0.71	0.73	0.67
Linguistic					
Time Interval -2					
Bert-based-multilingual-cased	0.60	0.60	0.71	0.53	0.67
Cross-en-de-roberta	0.70	0.70	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	0.72	0.72	0.72	0.73	0.70
Multilingual-e5-large	0.73	0.73	0.76	0.70	0.77
LaBSE	0.70	0.70	0.72	0.70	0.70
Text2vec	0.68	0.68	0.71	0.70	0.67
XLm-RoBERTa	0.68	0.68	0.66	0.73	0.63
Llama3	0.65	0.65	0.69	0.63	0.67
Time Interval -1					
Bert-based-multilingual-cased	0.53	0.53	0.55	0.53	0.53
Cross-en-de-roberta	0.75	0.75	0.78	0.70	0.80
Distilbert-base-multilingual-cased	0.56	0.57	0.59	0.63	0.50
Distiluse-base-multilingual-cased-v1	0.68	0.68	0.79	0.63	0.73
Multilingual-e5-large	0.72	0.72	0.79	0.70	0.73
LaBSE	0.66	0.67	0.74	0.57	0.77
Text2vec	0.68	0.68	0.77	0.7	0.67
XLm-RoBERTa	0.62	0.62	0.66	0.57	0.67
Llama3	0.73	0.73	0.78	0.73	0.73
Fusion					
Time Interval -2					
Distiluse-base-multilingual-cased-v1, Wav2vec 2.0, HuBERT, X-vectors, Multilingual-e5-large, Cross-en-de-roberta	0.74	0.72	0.77	0.80	0.63
Time Interval -1					
Multilingual-e5-large, Llama3, Wav2vec 2.0, Whisper, HuBERT, Cross-en-de-roberta	0.81	0.80	0.86	0.87	0.73

Table 3: *Per-speaker classification results for acoustic, linguistic, and fusion models across two time intervals. Metrics reported include F1-score (F1), accuracy (ACC), area under the ROC curve (AUC), sensitivity (SENS), and specificity (SPEC). Models are evaluated separately across two time intervals relative to the year of diagnosis (YoD). Bold values indicate the highest F1, ACC, and AUC scores in each time interval.*

acoustic and linguistic models enhances classification robustness by capturing both linguistic and paralinguistic changes associated with disease progression. The superior performance of multimodal models underscores the complementary nature of acoustic and linguistic markers, reinforcing the potential of speech-based, non-invasive screening tools for early AD detection and timely intervention. Future research will expand the dataset with additional recordings, allowing for a more detailed longitudinal analysis of speech changes across prodromal stages. Additionally, we will enhance data diversity by incorporating speakers from various linguistic, national, and ethnic backgrounds to analyze the robustness and generalizability of speech-based AD detection models. This will ensure their applicability across diverse populations, making them more inclusive for early diagnosis and monitoring.

6. References

- [1] I. Vigo, L. Coelho, and S. Reis, "Speech- and language-based classification of alzheimer's disease: A systematic review," *Bio-engineering*, vol. 9, no. 1, 2022.
- [2] W. Douglas and M. Scharre, "Preclinical, prodromal, and dementia stages of alzheimer's disease," *Pract. Neurol*, vol. 2019, pp. 36–42, 2019.
- [3] L. Calzà, G. Gagliardi, R. R. Favretti, and F. Tamburini, "Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia," *Computer Speech & Language*, vol. 65, p. 101113, 2021.
- [4] D. Smirnova, T. Smirnova, and P. Cumming, "Language impairments in dementia: From word-finding difficulties to everyday conversation in a dementia-friendly community," *Dementia Care: Issues, Responses and International Perspectives*, pp. 85–108, 2021.
- [5] X. Le, I. Lancashire, G. Hirst, and R. Jokel, "Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three british novelists," *Literary and linguistic computing*, vol. 26, no. 4, pp. 435–461, 2011.
- [6] V. Berisha, S. Wang, A. LaCross, and J. Liss, "Tracking discourse complexity preceding alzheimer's disease diagnosis: A case study comparing the press conferences of presidents ronald reagan and george herbert walker bush," *Journal of Alzheimer's Disease*, vol. 45, no. 3, pp. 959–963, 2015.
- [7] U. Petti, S. Baker, A. Korhonen, and J. Robin, "The generalizability of longitudinal changes in speech before alzheimer's disease diagnosis," *Journal of Alzheimer's Disease*, vol. 92, no. 2, pp. 547–564, 2023, PMID: 36776053.
- [8] K. C. Fraser, K. Lundholm Fors, M. Eckerström, and et al., "Improving the sensitivity and specificity of mci screening with linguistic information," in *LREC workshop: RaPID-2. Miyazaki, Japan*, 2018.
- [9] B. Mirheidari, Y. Pan, D. Blackburn, and et al., "Identifying cognitive impairment using sentence representation vectors." in *Interspeech*, 2021, pp. 2941–2945.
- [10] S. Amini, B. Hao, L. Zhang, M. Song, A. Gupta, C. Karjadi, V. B. Kolachalama, R. Au, and I. C. Paschalidis, "Automated detection of mild cognitive impairment and dementia from voice recordings: a natural language processing approach," *Alzheimer's & Dementia*, vol. 19, no. 3, pp. 946–955, 2023.
- [11] A. Pourramezan Fard, M. H. Mahoor, M. Alsuhaibani, and et al., "Linguistic-based mild cognitive impairment detection using informative loss," *Computers in Biology and Medicine*, vol. 176, p. 108606, 2024.
- [12] E. Kurtz, Y. Zhu, T. Driesse, B. Tran, J. A. Batsis, R. M. Roth, and X. Liang, "Early detection of cognitive decline using voice assistant commands," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [13] F. Tang, J. Chen, H. H. Dodge, and J. Zhou, "The joint effects of acoustic and linguistic markers for early identification of mild cognitive impairment," *Frontiers in digital health*, vol. 3, p. 702772, 2022.
- [14] 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*, 2024, pp. 972–976.
- [15] A. Favaro, Y.-T. Tsai, A. Butala, T. Thebaud, J. Villalba, N. Dehak, and L. Moro-Velázquez, "Interpretable speech features vs. dnn embeddings: What to use in the automatic assessment of parkinson's disease in multi-lingual scenarios," *Computers in Biology and Medicine*, vol. 166, p. 107559, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S010482523010247>
- [16] R. Gill and R. C. Gill, *Gender and the Media*. Polity, 2007.
- [17] A. Favaro, A. Butala, T. Thebaud, J. Villalba, N. Dehak, and L. Moro-Velázquez, "Unveiling early signs of parkinson's disease via a longitudinal analysis of celebrity speech recordings," *npj Parkinson's Disease*, vol. 10, no. 1, p. 207, 2024.
- [18] Category:People with Alzheimer's disease, "Category:people with alzheimer's disease," 2024, [Online; accessed 30-May-2024]. [Online]. Available: https://en.wikipedia.org/wiki/Category:People_with_Alzheimer%27s_disease
- [19] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *INTERSPEECH 2023*, 2023.
- [20] D. Snyder, D. Garcia-Romero, G. Sell, and et al., "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.
- [21] J. Shor and S. Venugopalan, "Trillsson: Distilled universal paralinguistic speech representations," in *Proceedings of Interspeech 2022*, 2022, pp. 356–360.
- [22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [23] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
- [24] A. Radford, J. W. Kim, T. Xu, and et al., "Robust speech recognition via large-scale weak supervision," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 28 492–28 518.
- [25] A. e. a. Conneau, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 8440–8451.
- [26] M. Xu, "text2vec: A tool for text to vector," 2023. [Online]. Available: <https://github.com/shibing624/text2vec>
- [27] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual e5 text embeddings: A technical report," *arXiv preprint arXiv:2402.05672*, 2024.
- [28] F. e. a. Feng, "Language-agnostic bert sentence embedding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 878–891.
- [29] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.
- [30] 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)*, 2019, pp. 3982–3992.
- [31] J. e. a. Devlin, "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)*, 2019, pp. 4171–4186.
- [32] A. Grattafiori, A. Dubey, A. Jauhri, and et al., "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.
- [33] L. Cummings, *Alzheimer's Dementia*. Cambridge University Press, 2020, p. 1–19.
- [34] 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.