INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Detecting Alzheimer's Disease using Interactional and Acoustic features from spontaneous speech

Shamila Nasreen[1,2] *, Julian Hough* [1] *, Matthew Purver*[1,3]

[1]Cognitive Science Group / Computational Linguistics Lab
School of Electronic Engineering and Computer Science
Queen Mary University of London, UK
[2]Department of Software Engineering, Mirpur University of Science and Technology, Pakistan
[3]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
{shamila.nasreen, j.hough, m.purver}@qmul.ac.uk

Abstract

Alzheimer's Disease (AD) is a form of Dementia that manifests in cognitive decline including memory, language, and changes in behavior. Speech data has proven valuable for inferring cognitive status, used in many health assessment tasks, and can be easily elicited in natural settings. Much work focuses on analysis using linguistic features; here, we focus on non-linguistic features and their use in distinguishing AD patients from similar-age Non-AD patients with other health conditions in the Carolinas Conversation Collection (CCC) dataset. We used two types of features: patterns of *interaction* including pausing behaviour and floor control, and *acoustic* features including pitch, amplitude, energy, and cepstral coefficients. Fusion of the two kinds of features, combined with feature selection, obtains very promising classification results: classification accuracy of 90% using standard models such as support vector machines and logistic regression. We also obtain promising results using interactional features alone (87% accuracy), which can be easily extracted from natural conversations in daily life and thus have the potential for future implementation as a non-invasive method for AD diagnosis and monitoring.

Index Terms: Alzheimer's disease, speech processing, acoustic features, Interactional patterns, computational paralinguistics

1. Introduction

Alzheimer's Disease (AD), the most prevalent form of Dementia, is an irreversible brain disorder associated with a gradual decline in cognitive functions of adults. Currently, it affects more than 5 million people in America every year. Its highest incidence is among adults due to age as a risk factor: one in every six individuals over the age of 80 is likely to develop AD and the number of cases over the age of 60 is doubling every 4–5 years [1]. Early recognition of cognitive decline could be helpful in managing pre-stage AD thus allowing better quality of life for elderly patients and their caregivers [2].

The most prominent associations with AD are disparity in language production, speech comprehension, impaired reasoning, and memory functions, resulting in reduced vocabulary, verbal fluency, and difficulty performing daily tasks related to semantic information [3]. This suggests that speech and natural language processing (NLP) methods could be suitable for use in the early recognition of impaired cognition and AD.

Currently, standard AD diagnosis methods include clinical assessments complemented with family history, neuropsychological tests (including the Mini-Mental State Examination (MMSE) [4] and many others), self-report questionnaires, MRI [5] and Positron Emission Tomography (PET) [6]. These methods are effective but are variously costly, invasive, time-consuming and/or stressful, require validation by neurologists, and must be performed in clinical settings. There is therefore a demand for extensible, less invasive methods that can reduce the burden on the health system and be reliably applied for AD diagnosis in more natural and less controlled environments.

The use of spontaneous speech to derive pathologically appropriate biomarkers for AD detection has therefore become a focus of research. Much of this work to date has focused on the properties of individual language, using various kinds of linguistic and acoustic features separately. Linguistic variables are used to describe the quantitative and qualitative aspects of language production, for example via the decline in lexical-semantic abilities, word comprehension, verbal fluency, and syntactic processing for particular kinds of tasks such as picture description [7, 8]. AD-related changes can also affect acoustic features of speech, suggesting that speech analysis could provide measures of early disease progression [9]. Several studies have used language-independent acoustic features only, achieving comparable accuracy to linguistic approaches [10, 11]; AD patients can show patterns of frequent hesitation, longer pauses, lower articulation and speech rates, and lower floor control ratio [12]. Other acoustic features including prosodic, energy-based, spectral, and spectral aspects (jitter, shimmer, harmonics-to-noise ratio, Mel-frequency cepstral coefficients (MFCCs) can also correlate with AD [13]. Some work has taken a multimodal approach to AD classification: Campbell et al. [14] examined two fusion strategies with linguistic features and acoustic features, achieving 75% accuracy. Shah et al. [2] used a weighted majority-vote ensemble algorithm for classification and chose the best performing language model with the three best performing acoustic models, giving final prediction accuracy of 83%.

Some other work focuses less on the individual and more on the properties of their interaction with others. Conversation Analysis (CA) studies show that dialogue with dementia has characteristic features that would be missed if analyzing only individual speech [15, 16], but these studies are generally qualitative and/or small-scale. Some computational work on dementia is starting to fill this gap, focusing on interaction patterns such as turn-taking behavior, disfluency, repair, repetition, and topic management. Luz et al. [17] use dialogue interaction features from the speech in a predictive model, with an impressive accuracy of 86%. Mirheidari et al. [18] go a step further, combining CA-inspired interaction features including turn-taking behavior with some acoustic and language features, to achieve

http://dx.doi.org/10.21437/Interspeech.2021-1526

a classification accuracy of 90%. Garcia et al. [19] develop a protocol based on dialogue conversations to investigate early behavioral signs of AD. This use of interaction cues has the potential to be more versatile in AD prediction and monitoring in more daily life settings than individual language tasks [20, 21]. However, work so far either looks only at interaction rather than combining it with other modalities (e.g. [17]) or relies on particular interactional settings such as interviews with chosen topics or question types (e.g. [18]). In this study, we address these issues, using a combination of dialogue interaction and acoustic features, and by analysing semi-structured interviews obtained in more natural settings. Our main contributions are:

- Evaluation of 31 dialogue interaction features extracted from the audio and transcripts of natural conversations.
- Analysis of various voicing, spectral, and energy-based aspects of speech, using feature selection techniques to reveal the most informative features.
- Analysis of different combinations of these features with feature selection and fusion strategies, showing that these can help distinguish between AD patients and Non-AD patients, achieving an overall accuracy of 87% with interactional features, and 90% with combining these with acoustic features.

2. Methodology

Our approach is to build a model based on interaction cues from dialogue conversations, combined with acoustic features, to predict whether an individual has AD or not. It consists of four main parts: feature engineering, feature selection, learning algorithms, and multimodal fusion strategy.

2.1. Feature Engineering

2.1.1. Interactional Features

Different temporal and interactional aspects of dialogue conversations are employed, grounded in Levinson's theory of pragmatics [22]. These include *short pauses (SP)*, *long pauses (LP)*, *gaps (GA)*, and *lapses (LA)*. *SPs* and *LPs* are silences within one individual's speech, with *SPs* less than 1.5 seconds and *LPs* greater than 1.5 seconds. Both *SPs* and *LPs* may occur either at a transition relevance place (TRP) or not, but no speaker change occurs; TRPs are places at which it would be appropriate for the turn at talk to pass from one speaker to another. A gap (*GA*) is silence at a speaker change (i.e. turn boundary, with speaker change from interviewer to patient *I-P* or vice versa *P-I*). A lapse (*LA*) is a longer delay in communication between two individuals, at a TRP, and after which one participant (usually interviewer in this case) initiates a new topic (Figure 1).

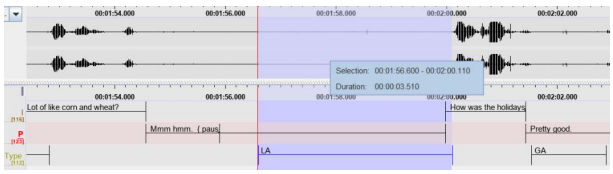

Figure 1: *A Lapse (LA) followed by new topic initiation.*

Further to these, we distinguish *attributable silences (ASs)*: silences during which a speaker change is strongly expected, to provide a response to the previous turn, but does not occur. In our dataset, this is usually when the interviewer *(I)* has asked a question, but no response is forthcoming from the patient *P*; a silence ensues, after which *I* takes the conversation floor again. Figure 2 shows an example, with the lack of response from *P* leading to an *AS* of 4.1 seconds following a question 'What other animals were there?'.

Table 1: *Interactional feature set*

Interaction features: Num_AS, Dur_AS, Num_LA, Dur_LA, Num_GA, Dur_GA, turn_switches_per_minute, num_overlaps, Num_GA(*I-P*), Dur_GA(*I-P*), Num_GA(*P-I*), Dur_GA(*P-I*), Num_LA(*I-I*), Dur_LA(*I-I*), Num_LA(*P-I*), Dur_LA(*P-I*).
Patient (*P*) features: *Num_SP, Dur_SP, Num_LP, Dur_LP, turn_length*, floor control ratio (*FCR*), standardized pause rate (*SPR*), transformed phonation rate (*TPR*), and *speech rate*.
Interviewer (*I*) features: *Num_SP, Dur_SP, Num_LP, Dur_LP, turn_length*, and *speech rate*.

Other features encode general characteristics of the interaction. We include the number of overlaps: the number of segments spoken simultaneously by both speakers, with the intuition that these may be attributed to speech initiation difficulties. We also include turn length (number of words per turn), floor control ratio (amount of time during which *P* speaks, relative to the total speech time of the conversation), standardized pause rate (ratio of total words spoken by *P* to the total pauses (including *SP & LP*)), phonation rate (total time spoken by *P* to total spoken time including *SP and LP* by *P*), and speech rate (number of words per minute). The annotation protocol for these interactional features is described in [23] for AD classification: 31 features in total are extracted based on this protocol from audio and transcript data as shown in Table 1.

Figure 2: *An Attributable Silence (AS) after an interviewer (I) question, followed by reformulation of the question by I.*

Features such as speech rate, SPR, TPR, turn lengths are used previously for Dementia analysis in the literature however, their combination with our defined interactional features such as *LA*, *AS*, *overlaps*, *SP, LP, GA*, and acoustic feature is unique.

2.1.2. Acoustic features

OpenSMILE v2.1 [24] was used to extract acoustic features from the audio recordings, for a total of 30 dialogue conversations. OpenSMILE is open-source software that has been previously used for AD classification using audio features [9, 25]. A set of 64 audio features was extracted and higher-order statistics (mean, standard deviation) were computed. Using the utterance timing information provided in the transcripts, we extracted the

participant's utterances (either *P* or *I*) and calculated average values of the features per utterance basis. A standard zero mean and variance normalization was applied to each feature. The detail of acoustic features is given in Table 2.

Table 2: *Acoustic feature set*

Type	Feature names
Frequency related	Fundamental frequency *(f_0)*, jitter, voicing probability
Energy, amplitude related	RMS energy, log RMS energy, shimmer, loudness, Harmonic to noise ratio *(HNR)*
Spectral parameters	4 Mel-frequency Cepstral coefficients *(MFCCs)* [1-4], *delta MFCCs* [1-4], *delta-delta MFCCs* [1-4]

2.2. Feature Selection

Feature selection (FS) reduces the dimensionality of the feature set by choosing a subset of relevant features. We used the recursive feature elimination (RFE) method, an iterative process that removes a specific number of features, and examines the effect on classification accuracy [26]. Those features making the least contribution are removed recursively until the desired number of features are left. We also utilized a pipeline with grid search to find the optimal value of a subset of features from both acoustic and interactional features.

2.3. Learning algorithms

Due to the low number of samples, compared to the dimensionality of the feature space, we use traditional machine learning classifiers rather than more complex neural networks, as the former has the potential to provide a rational trade-off between classification performance and run-time complexity and the risk of overfitting [27]. In this study, three traditional ML classifiers were used: Logistic Regression (LR), Support Vector Machines (SVM) and Random Forests (RF). We tested LR with a range of regularization parameters (0.1,10,100); SVM with RBF and polynomial kernels, cost C (0.1,100) and gamma (0.001,0.1); and RF with 60 trees of maximum depth of 5. The same hyperparameters were used for all experiments.

2.4. Fusion strategy

Two different fusion strategies were employed in this study. In the *early fusion (EF)* method, the values of each feature for both acoustic and interactional features are normalized using the standard scalar feature of scikit-learn [28] and then concatenated directly. The *late fusion (LF)* or decision-level strategy utilized the same normalization for each feature set, but with predictions made individually for each feature set. The prediction scores of each classifier are then combined using a standard soft voting ensemble method [29]: soft voting computes the average probability for each class over each component classifier and eventually bases the final prediction on maximum average probability. Among our classifiers, LR and RF provide prediction probabilities directly, while the SVM outputs were transformed into prediction probabilities using Praat scaling [27].

3. Experiments

Dataset The Carolinas Conversation Collection (CCC) [30] consists of audios and transcripts of conversations between health care professionals and patients with chronic diseases including AD. These are semi-structured interviews recorded in community centers and are a useful resource to explore interaction aspects of communication. Online access to the CCC was obtained after gaining ethical approval from our own institution and from the CCC administrators/hosts (Medical University of South Carolina, MUSC), and complying with MUSC's requirements for data handling and storage. For this study we selected dialogues for 30 patients: 15 diagnosed with AD (11 female, 4 male) and 15 non-AD patients (4 male, 11 females) with other chronic diseases such as diabetes, heart attack, broken leg, etc but not AD. Although the full CCC dataset contains 200 conversations with the non-AD group of patients and 400 conversations for the AD group, however, not all dialogues are available.[1] Dialogues of 38 AD patients were available and we randomly choose 15 dialogues of AD for this study. The total duration of interviews is 152 minutes for the Non-AD group and 179.7 minutes for the AD group.

Implementation metrics We set up our experiments to investigate which acoustic features and dialogue interaction features are most effective for predicting AD. Due to the fairly small dataset, we used leave-one-(patient)-out cross validation (LOOCV) to get a better estimation of generalization accuracy. The dataset is balanced in terms of classes; we choose precision, recall, F1-score, and accuracy as evaluation metrics.

Baseline Models We compared the performance of our model with Luz et al. [17]'s work on the same CCC corpus with dialogue interaction features. Luz et al.'s dataset is slightly bigger than ours (38 dialogues vs. 30). Although the features set are not directly comparable, they utilized only interactional aspects of conversation including dialogue duration, average turn duration, normalized duration, average number of words, and average words per minute from the spontaneous speech.

4. Results and Discussion

Effects of feature selection We performed RFE separately on both acoustic and interactional features, securing good classification results with 15 top interactional features and 42 acoustic features. Table 3 shows the ranked features from both feature sets, together with their Pearson's correlation (*r*) with the diagnosis class; due to space limitations, we only show the top 10 features.[2] The most significant acoustic feature was LogHNR, known to be important in acoustic analysis for the diagnosis of pathological voices; loudness, raw fundamental frequency, variation in jitter, intensity, and LogHNR all positively correlate with AD and have been reported as useful features in literature for Dementia [31, 9]. Among interactional features, lapses are positively correlated with AD, indicating that patients find trouble continuing topics, resulting in delays with interviewers initiating a new topic. Attributable silence (AS) duration is also positively correlated with AD, showing that AD patients exhibit more silences in response to questions; this fits with the findings of CA studies in the literature that this form of silence serves as a dispreferred response in communication [22, 32]. Turn lengths seem to negatively correlate with AD: AD patients produce less number of words in their turns with longer turn duration (5.91 vs 4.01 seconds), and consequently, produce more silences within the turns as compared to Non-AD. Standardized phonation time (SPT) and transformed phonation rate (TPR)

[1] See `https://carolinaconversations.musc.edu/`

[2] Detail can be found here: `https://osf.io/3fd8x/?view_only=8d864851fbd74be5b53c0ef86335a25a`

also show significant differences between the two groups, confirming findings in the literature [12].

Table 3: *Top-ranked features for distinguishing patients with AD from Non-AD patients. The third and fifth columns show Pearson's correlations with the AD class.*

Rank	Acoustic	*r*	Interactional	*r*
1	LogHNR_SD	0.60	*dur_LA (I-I)*	0.32
2	Voicing_Final_mean	0.58	*P_TPR*	-0.26
3	loudness_SD	0.56	*P_SPT*	-0.54
4	mfcc[2]_mean	-0.54	*P_turn_length*	-0.33
5	mfcc_de_de[3]_mean	0.49	*dur_AS (I_P)*	0.41
6	mfcc_de_de[3]_SD	-0.45	*dur_LA*	0.43
7	jitterDDP_SD	0.45	*dur_LA (P_))*	0.30
8	F0_raw_mean	0.44	*I_turn_length*	-0.40
9	intensity_SD	0.44	*dur_GA(P_I)*	0.38
10	LOgHNR_mean	0.42	*Num_GA*	0.49

Classification Results Table 4 then shows our results for the AD/Non-AD classification task. We show results for both feature sets individually and using the EF and LF fusion strategies; as a baseline, we compare against Luz et al. [17]'s method that utilized only interactional features. Combining interactional and acoustic features with the EF strategy seems to do best, and gives performance for all classifiers that improve over the baseline, and over the use of the individual feature sets. Using feature selection (FS) to select the 15 top-ranked interactional features and 42 top-ranked acoustic features increases performance further, giving an accuracy of 0.90 with both LR and SVM, and 0.87 with RF with the EF strategy. Contrary to our expectations, the LF method gives similar performance to EF when using feature selection, although it shows a significant drop in accuracy when using all features, with all three classifiers. Due to lower performance with this late fusion strategy, we only considered early fusion in our error analysis. The overall findings confirm our expectation that adding interactional features improves the results in comparison to more conventional acoustic features found in the literature. For each type of classifier, the performance improves when both features are combined. We note, though, that as we use a relatively small set of dialogues, it is unclear how well results will generalize.

Error analysis The results in Table 5 show that the EF strategy with top-ranked acoustic and interactional features obtains the highest precision and recall for both AD and non-AD classes with LR and SVM, with F1 scores of 0.90 for AD and 0.89 for Non-AD. Combining interactional and acoustic features particularly improves recall (0.93) of the AD class: acoustic features alone (with SVM) give recall 0.67 for the AD class, increasing to 0.73 with top-ranked features, while interactional features alone give 0.80 with all features and 0.87 with the top 15 features. Depending on the application the model is used for, false negatives or false positives for AD detection will be more or less desirable, but as it stands combining the most relevant features considerably reduces the false negatives of diagnosis whilst still marginally reducing the false positives.

5. Conclusion

This study investigating alternative means of AD diagnosis using features that may be easily computed upon daily conversations, without relying on specific neuropsychological assessments, would improve not only the diagnosis but also the monitoring of the disease. In this respect, using only these interactional type of features, the study achieves an accuracy of 87%, while combining interactional features with conventional acoustic features with early fusion improves the results with an accuracy of 90%. These results are in line with the current state of the art that uses alternative features based on fixed tasks such as picture description focusing on language characteristics of individuals and harder to derive from natural settings.

In future work, we intend to add more conversations from the CCC corpus, to address the issue of small sample size and investigate generalisability. In particular, we will automate the feature extraction process for interactional features keeping the context in consideration. We will also look into more interaction of communication in terms of the relevance of responses, dialogue act tagging to identify clarification requests, signal of non-understanding, several types of questions, and answers occurring in natural dialogues.

Table 4: *Multiple classifiers with different feature sets and fusion strategies, with all and top-ranked features with FS.*

Classifier	Features	Acc. [all]	Acc. [FS]	# features
Baseline Luz et al. [17]				
LR	Interactional	0.75	-	-
SVM	Interactional	0.83	-	-
RF	Interactional	0.81	-	-
Our Models				
LR	Acoustic	0.70	0.73	42
	Interactional	0.80	0.83	15
	Both (EF)	**0.87**	**0.90**	(42,15)
	Both (LF)	0.77	0.90	(42,15)
SVM	Acoustic	0.73	0.77	42
	Interactional	0.83	**0.87**	15
	Both (EF)	0.73	**0.90**	(42,15)
	Both (LF)	0.77	0.89	(42,15)
RF	Acoustic	0.80	0.83	42
	Interactional	0.73	0.73	15
	Both (EF)	**0.83**	**0.87**	(42,15)
	Both (LF)	0.73	0.87	(42,15)

Table 5: *Comparison of results for the AD classification, shown as precision, recall, F1, and accuracy per class.*

Model	No.	Class	Prec.	Rec.	F1	Acc.
SVM Acoustic	All	Non-AD AD	0.71 0.77	0.80 0.67	0.75 0.71	0.73
SVM Acoustic	42	Non-AD AD	0.75 0.79	0.80 0.73	0.77 0.76	0.77
SVM Interactional	All	Non-AD AD	0.81 0.86	0.87 0.80	0.84 0.83	0.83
SVM Interactional	15	Non-AD AD	0.87 0.87	0.87 0.87	0.87 0.87	0.87
SVM Both (EF)	(42,15)	Non-AD AD	0.93 0.88	0.87 0.93	0.89 0.90	0.90
LR Both (EF)	(42,15)	Non-AD AD	0.93 0.88	0.87 0.93	0.89 0.90	0.90

6. Acknowledgements

Purver is partially supported by the EPSRC under grant EP/S033564/1, and by the European Union's Horizon 2020 programme under grant agreements 769661 (SAAM, Supporting Active Ageing through Multimodal coaching) and 825153 (EMBEDDIA, Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

7. References

[1] Alzheimer's Society. (2020) Facts for the media. [Online]. Available: https://www.alzheimers.org.uk/about-us/news-and-media

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

[3] K. E. Forbes-McKay and A. Venneri, "Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task," *Neurological Sciences*, vol. 26, no. 4, pp. 243–254, 2005.

[4] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state": a practical method for grading the cognitive state of patients for the clinician," *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–198, 1975.

[5] J. Straiton, "Predicting Alzheimer's disease," 2019.

[6] J. Weller and A. Budson, "Current understanding of alzheimer's disease diagnosis and treatment. f1000research," 2018.

[7] V. Boschi, E. Catricala, M. Consonni, C. Chesi, A. Moro, and S. F. Cappa, "Connected speech in neurodegenerative language disorders: a review," *Frontiers in Psychology*, vol. 8, p. 269, 2017.

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

[9] H. Lin, C. Karjadi, T. F. Ang, J. Prajakta, C. McManus, T. W. Alhanai, J. Glass, and R. Au, "Identification of digital voice biomarkers for cognitive health," *Exploration of Medicine*, vol. 1, p. 406, 2020.

[10] J. Weiner, M. Angrick, S. Umesh, and T. Schultz, "Investigating the effect of audio duration on dementia detection using acoustic features." in *INTERSPEECH*, 2018, pp. 2324–2328.

[11] R. Chakraborty, M. Pandharipande, C. Bhat, and S. K. Kopparapu, "Identification of dementia using audio biomarkers," *arXiv preprint arXiv:2002.12788*, 2020.

[12] D. Beltrami, G. Gagliardi, R. Rossini Favretti, E. Ghidoni, F. Tamburini, and L. Calzà, "Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline?" *Frontiers in Aging Neuroscience*, vol. 10, p. 369, 2018.

[13] L. Ning and K. Luo, "Using text and acoustic features to diagnose mild cognitive impairment and Alzheimer's disease," 2020.

[14] E. L. Campbell, L. Docío-Fernández, J. J. Raboso, and C. García-Mateo, "Alzheimer's dementia detection from audio and text modalities," *arXiv preprint arXiv:2008.04617*, 2020.

[15] C. Elsey, P. Drew, D. Jones, D. Blackburn, S. Wakefield, K. Harkness, A. Venneri, and M. Reuber, "Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics," *Patient Education and Counseling*, vol. 98, no. 9, pp. 1071–1077, 2015.

[16] A. Varela Suárez, "The question-answer adjacency pair in dementia discourse," *International Journal of Applied Linguistics*, vol. 28, no. 1, pp. 86–101, 2018.

[17] S. Luz, S. de la Fuente, and P. Albert, "A method for analysis of patient speech in dialogue for dementia detection," *arXiv preprint arXiv:1811.09919*, 2018.

[18] B. Mirheidari, D. Blackburn, T. Walker, M. Reuber, and H. Christensen, "Dementia detection using automatic analysis of conversations," *Computer Speech & Language*, vol. 53, pp. 65–79, 2019.

[19] S. de la Fuente Garcia, C. W. Ritchie, and S. Luz, "Protocol for a conversation-based analysis study: Prevent-ed investigates dialogue features that may help predict dementia onset in later life," *BMJ open*, vol. 9, no. 3, p. e026254, 2019.

[20] S. de la Fuente Garcia, C. Ritchie, and S. Luz, "Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: A systematic review," *Journal of Alzheimer's Disease*, no. Preprint, pp. 1–27, 2020.

[21] A. Addlesee, A. Eshghi, and I. Konstas, "Current challenges in spoken dialogue systems and why they are critical for those living with dementia," *arXiv preprint arXiv:1909.06644*, 2019.

[22] S. C. Levinson, "Pragmatics cambridge university press," *Cambridge UK*, 1983.

[23] S. Nasreen, M. Rohanian, M. Purver, and J. Hough, "Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features," *Frontiers in Computer Science*, vol. 3, p. 49, 2021.

[24] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[25] T. Warnita, N. Inoue, and K. Shinoda, "Detecting alzheimer's disease using gated convolutional neural network from audio data," *arXiv preprint arXiv:1803.11344*, 2018.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[27] M. Taschwer, M. J. Primus, K. Schoeffmann, and O. Marques, "Early and late fusion of classifiers for the mediaeval medico task." in *MediaEval*, 2018.

[28] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.

[29] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[30] C. Pope and B. H. Davis, "Finding a balance: The Carolinas Conversation Collection," *Corpus Linguistics and Linguistic Theory*, vol. 7, no. 1, pp. 143–161, 2011.

[31] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana, "Speech in alzheimer's disease: Can temporal and acoustic parameters discriminate dementia?" *Dementia and Geriatric Cognitive Disorders*, vol. 37, no. 5-6, pp. 327–334, 2014.

[32] C. Wang, "A relevance-theoretic approach to turn silence," in *4th International Conference on Contemporary Education, Social Sciences and Humanities (ICCESSH 2019)*. Atlantis Press, 2019, pp. 1078–1084.