THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Alzheimer's Dementia Detection through Spontaneous Dialogue with Proactive Robotic Listeners

Citation for published version:
Li, Y, Lai, C, Lala, D, Inoue, K & Kawahara, T 2022, Alzheimer's Dementia Detection through Spontaneous Dialogue with Proactive Robotic Listeners. in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction.* ACM, pp. 875-879, 2022 ACM/IEEE International Conference on Human-Robot Interaction, 7/03/22. https://doi.org/https://dl.acm.org/doi/abs/10.5555/3523760.3523896, https://doi.org/10.5555/3523760.3523896

Digital Object Identifier (DOI):
https://dl.acm.org/doi/abs/10.5555/3523760.3523896
10.5555/3523760.3523896

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 08. Jun. 2022

Alzheimer's Dementia Detection through Spontaneous Dialogue with Proactive Robotic Listeners

Yuanchao Li, Catherine Lai
School of Informatics
University of Edinburgh, UK
y.li-385@sms.ed.ac.uk, c.lai@ed.ac.uk

Divesh Lala, Koji Inoue, Tatsuya Kawahara
Graduate School of Informatics
Kyoto University, Japan
{lala, inoue, kawahara}@sap.ist.i.kyoto-u.ac.jp

***Abstract*—As the aging of society continues to accelerate, Alzheimer's Disease (AD) has received more and more attention from not only medical but also other fields, such as computer science, over the past decade. Since speech is considered one of the effective ways to diagnose cognitive decline, AD detection from speech has emerged as a hot topic. Nevertheless, such approaches fail to tackle several key issues: 1) AD is a complex neurocognitive disorder which means it is inappropriate to conduct AD detection using utterance information alone while ignoring dialogue information; 2) Utterances of AD patients contain many disfluencies that affect speech recognition yet are helpful to diagnosis; 3) AD patients tend to speak less, causing dialogue breakdown as the disease progresses. This fact leads to a small number of utterances, which may cause detection bias. Therefore, in this paper, we propose a novel AD detection architecture consisting of two major modules: an ensemble AD detector and a proactive listener. This architecture can be embedded in the dialogue system of conversational robots for healthcare.**

***Index Terms*—Alzheimer's dementia, speech and language processing, dialogue systems, human-robot interaction, digital health**

I. Introduction

With the rapid growth of the elderly population, Alzheimer's Disease (AD) has become a serious problem in today's aging society. AD is a neurodegenerative disease that progressively deteriorates memory, language, and cognitive abilities. Hence, early detection of AD for prevention is the most important to tackle the disease [1]. Many studies have proved that AD is recognizable from spontaneous speech [2]–[4], and both audio and transcript information contribute to the detection [5], [6]. These studies have greatly advanced the early detection of AD.

AD detection consists of two necessary steps: feature extraction and classification/regression. The ADReSS Challenge at INTERSPEECH 2020 [7] has presented a comparison study using several baseline feature sets, which include *emobase* [8], *ComParE* [9], *eGeMAPS* [10], *MRCG functionals* [11], and *Minimal* [12]. These feature sets include acoustic features such as energy, Mel-Frequency Cepstral Coefficients (MFCC), fundamental frequency (F0), and so on, as well as their statistical functionals. Besides, directly learning the mapping from raw speech signals using neural networks has emerged as a trend in current work [13]. Traditional machine learning approaches such as linear discriminant analysis, decision trees, nearest neighbor, random forests, and support vector machines are adopted for classification and regression using extracted features in the ADReSS Challenge. In addition, deep learning-based approaches are also being investigated for the same tasks and show great performance [14], [15].

Despite the progress, AD detection remains challenging due to several key issues. First of all, AD is a complex neurocognitive disorder that needs a professional medical diagnosis. Detection using computers with only utterance-level information is inappropriate and causes inaccurate results [3]. Thus, more information from dialogue aspects (e.g., turn-taking time) should be considered. Second, AD patients are usually not able to speak as fluently and clearly as non-AD people. Their utterances contain many disfluencies, such as fillers, false starts, repetitions, and so on [15], [16]. These disfluencies affect Automatic Speech Recognition (ASR) performance, and as a consequence, the error-prone ASR transcripts are not as reliable as manual transcripts. Third, AD patients tend to speak less as the disease progresses, which leads to a small number of utterances [17]. In that case, data scarcity can cause a serious detection bias problem.

Nowadays, robots with a dialogue function have been replacing human labor in several scenarios related to reception, presentation, elderly care, and the like [18]. Therefore, we propose a novel AD detection architecture that is expected to solve the above-mentioned problems. This proposed architecture can be embedded in the dialogue system of conversational robots, with the aim of breaking disciplinary boundaries to help AD patients.

II. Related Work

A. AD Detection from Speech

To automate AD detection, researchers have used various acoustic and lexical features. Traditionally, a number of acoustic-related low-level descriptors built upon prior knowledge have been handcrafted to represent AD. [4] evaluated several handcrafted feature sets designed for different computational paralinguistics tasks and proposed a novel Active Data Representation method using acoustic features of all

speech segments with a single fixed-dimension feature vector. Inspired by successful end-to-end approaches in speech and emotion recognition [19], [20], some recent works extract features directly from log-mel spectrograms instead of using handcrafted ones, resulting in better performance [21], [22].

Unlike from acoustic analysis, language research employs high-level lexical features for AD detection. For example, Lu's L2 Syntactic Complexity Analyzer computed 23 features that measure the syntactic complexity of the text, which include lengths of production units, the ratio of clauses to sentences, subordination, coordination, and particular structures [23]. Thanks to the rapid development of Natural Language Processing (NLP), current research is adopting pre-trained models such as GloVe [24], BERT [25], and RoBERTa [26] to extract lexical features [13], [15], [27], and such methods usually outperform handcrafted feature engineering [28].

Alongside acoustic and lexical features, some work also takes into consideration interactivity and disfluency features. Based on the finding that therapist-patient dialogues can be regarded as Markov chains [29], [30] presented descriptive statistics on dialogues, such as dialogue duration, turn duration, number of words, and words per minute. Disfluencies, a natural pattern in spontaneous dialogues, are necessary for detecting cognitive diseases [31], which leads to AD detection using self-repairs, editing terms, short pauses, and long pauses [14].

In this paper, we follow prior work and propose a novel detection approach that uses two-stage ensemble learning to integrate multiple classifiers for the final AD detection results.

B. Conversational Robots for Healthcare

Conversational robots have been used in various situations nowadays, and they have proven the importance of social Human-Robot Interaction (HRI) as a means of providing effective healthcare. Prior research has demonstrated that the elderly are eager to talk to robots because of the lack of social ties. Even simple social conventions like daily greetings from the robot can comfort the elderly with the assurance that the robot will always be there to respond [32]. This finding provides a basis to use conversational robots for AD detection. Recent work has investigated experiment design and strategies for AD detection using conversational robots. [33] indicated that even though the conversational robot presents a non-pharmacological treatment approach, its daily use can have a therapeutic effect on the behavioral and psychological symptoms of dementia in older adults. [34] explored how conversational robots can be used to support individuals with AD, and their results show that the robot was generally well-liked by AD patients and that it could capture their interest in dialogue. They also highlighted the robot's potential as a monitoring tool by analyzing how non-acoustic aspects of language change across participants with different AD degrees. [35] developed a prototype of a listener agent, and collected conversations between the agent and people with dementia for assessment using a conversation log system.

However, the above-mentioned research ignores the fact that AD patients tend to speak less as the disease progresses. This dialogue breakdown impedes the automatic AD detection from being applied in real life. Hence, there is a need for the system to maintain the conversation until an overall diagnosis is conducted. In this paper, we propose to use a proactive listener to resolve this issue.

III. Proposed Approach

A. Architecture Overview

The overview of the proposed architecture is shown in Fig. 1, and the processing steps are as follows. First, the spontaneous dialogue between the human and the robot takes place regardless of who starts it. The human's speech signals are perceived by the robot's microphones and then processed by the front end. It should be noted that to process the speech signals, there are two paths, one of which directly extracts acoustic features and models the feature sequence, and the other converts speech into ASR transcripts. We choose to use open-source tools to perform this step because they are convenient to implement in programmable robots.

Next, *proactive listener*, one of the major modules in the architecture, recognizes the type of user speech with a focus word extractor and a dialogue act tagger. This proactive listener is a variant of a previously proposed attentive listener [36], mainly by removing backchanneling and flexible turn-taking functions and adding a proactive initiator [37]. We define user speech into three types: question, statement, and silence. Each type has its corresponding response: answer, question, partial repeat, follow-up question, and topic introduction. The proactive responses aim to maintain the conversation to acquire more speech samples from the human user for a comprehensive diagnosis.

Meanwhile, *ensemble AD detector*, the other major module in the architecture, detects the AD degree using a two-stage ensemble learning approach. Inside this module, there are four classifiers conducting AD detection on four different modalities respectively: audio, language, disfluency, and interactivity. Then an ensemble learning model applies averaging and a majority vote in two stages on the four probability distributions to obtain the final decision. The output has four degrees: mild AD, moderate AD, severe AD, and non-AD, which will finally be recorded in medical logs for medical professionals to follow up on the disease conditions in long term.

B. Detail Description

In this section, we give a detailed description of the two major modules.

1) Proactive Listener: Our goal of implementing the proactive listener is to maintain spontaneous dialogues between the robot and the human, considering that the dialogues are easy to break down since AD patients tend to speak less as the disease progresses. The focus word extractor is one of the NLP functionalities that identifies the focus of an utterance based on a conditional random field classifier that uses part-of-speech tags and a phrase-level dependency tree [38]. The dialogue

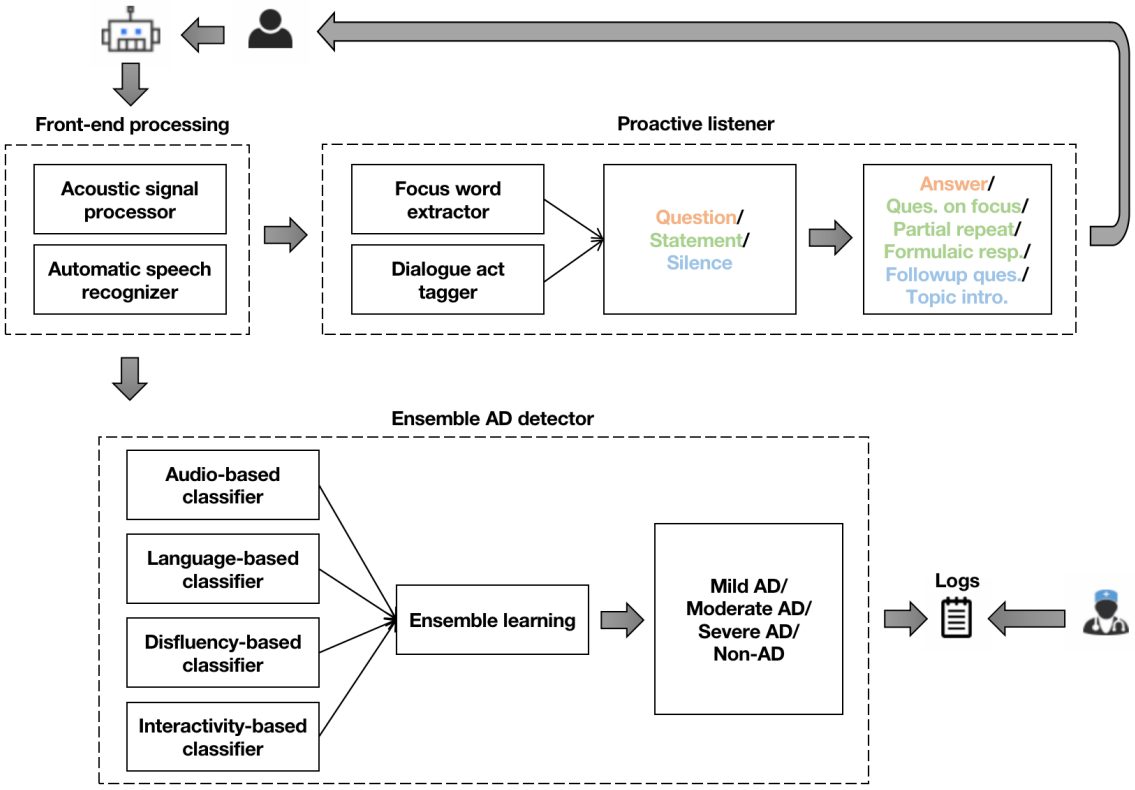

Fig. 1. The proposed AD detection architecture.

act tagger classifies user utterances into question, statement or silence, and works together with the focus word extractor.

For "question", the response will be generated based on adjacency pairs from a handcrafted question-answer database. Compared to statistical methods (e.g., neural dialogue models) for generating responses, carefully crafted rule-based system is better in this situation for control, since dialogues with AD patients are vulnerable. For "statement", the response can be a question or a partial repeat according to a decision tree scheme. We follow a previous approach using an n-gram language model to compute the joint probability of the focus noun being associated with each wh-question word (i.e., who, what, when, where) [36]. If the maximum joint probability of this noun and a question word is over a pre-defined threshold, then a question on the focus word is generated. Otherwise, a partial repeat is generated. If no focus noun, then a formulaic response is generated for rapport. For "silence", if it lasts five seconds, then a follow-up question related to the most recent topic will be responded. If the silence continues lasting another five seconds, then a topic introduction will be started. A dialogue example is shown in Table. 1.

2) Ensemble AD Detector: The ensemble AD detector conducts detection on four aspects: audio, language, disfluency, and interactivity.

I. *Audio-based classifier.* Following a prior HRI work [39], we first use the OpenSMILE toolbox [40] to extract ComParE [9] acoustic feature set as a baseline. We will also directly use raw speech in an end-to-end manner for comparison. In addition, we will also use a pre-trained acoustic model such as wav2vec [41] or HuBERT [42] to explore its efficiency in real-time application. We present the sequence of the extracted acoustic feature vectors as $[a_1, ..., a_M]$, where M is the sequence length.

II. *Language-based classifier.* We use word embedding to convert each word from the text as an embedding vector. We will also use a pre-trained language model such as GloVe [24],

TABLE I
A DIALOGUE EXAMPLE CONSISTING OF DIFFERENT UTTERANCE-RESPONSE TYPES. H: HUMAN, R: ROBOT

Utterance-response type	Example
Question-answer	H: How is the weather? R: It's raining outside.
Statement-ques. on focus	H: OK, I'll watch a movie then. R: Which movie?
Statement-partial repeat	H: Avengers, the newest one. R: Avengers?
Silence-followup ques.	H: [5s silence] R: What's your favorite movie?
Silence-topic intro.	H: [5s silence] R: Do you like music?
Statement-formulaic resp.	H: Yes, I like. R: That's good.

BERT [25] or RoBERTa [26] for the same reason above. We present the sequence of the extracted lexical feature vectors as $[l_1, ..., l_N]$, where N is the sequence length.

III. *Disfluency-based classifier.* We categorize disfluency into four categories: restart, repetition, correction, and filler. The traditional way to detect AD from disfluency is first predicting disfluency from feature inputs using SVMs or neural networks, and then detecting AD from disfluency patterns according to a linear regression mapping relationship. However, considering the ASR error in transcripts, which leads to inaccurate disfluency prediction, we directly detect AD from feature inputs. Following [43], we use one-hot vectors from the ASR transcripts as the lexical feature input. Following [44], we use voice activity, pitch, intensity, and spectral stability as the prosodic feature input. We define the extracted lexical and prosodic features as disfluency features and present their sequence as $[d_1, ..., d_P]$, where P is the sequence length.

IV. *Interactivity-based classifier.* Unlike other three classifiers that conduct utterance-level detection (one detection result per utterance), interactivity-based classifier conducts dialogue-level detection (one detection result per dialogue). We define six turn-pairs as a dialogue, and extract the following interactional features according to prior knowledge [45]: turn length (number of words per turn), floor control ratio (time amount during the human speaker speaks to the total speech time of the dialogue), standardized pause rate (ratio of total words to the total pauses), phonation rate (total time spoken to total spoken time including pause), and speaking rate (number of words per minute).

The interactivity-based classifier will be built on linear regression since there are only five feature types. The other three classifiers, as well as the dialogue act tagger, will be built on Gated Recurrent Unit (GRU) since it performs similarly to Long Short-Term Memory (LSTM) but is computationally cheaper and more efficient, which is crucial to real-life HRI. The GRU network consists of a forward GRU that reads the input from left to right and a backward GRU that reads the input reversely to better model the sequential structure of the feature sequences.

$$\overrightarrow{h}_t = GRU(\overrightarrow{x_t}, \overrightarrow{h}_{t-1}) \quad (1)$$

$$\overleftarrow{h}_t = GRU(\overleftarrow{x_t}, \overleftarrow{h}_{t-1}) \quad (2)$$

$$h_t = \overrightarrow{h}_t \oplus \overleftarrow{h}_t \quad (3)$$

where x_t and h_t are the input feature vector and the hidden state at time t respectively, and $\oplus$ is a concatenation operation.

Since the final result is not determined until a dialogue (six turn pairs) is over, the ensemble learning has two stages. In the first stage, there are six probability distributions in accordance to six turn-pairs from the three utterance-level classifiers. An averaging operation is conducted on the six probability distributions of each classifier, respectively, to obtain three AD detection results. Meanwhile, the interactivity-based classifier generates one AD detection result. In the second stage, a majority vote is applied on the four results, to produce the final diagnosis output.

IV. Discussion

In this paper, we propose a novel diagnosis architecture consisting of an ensemble AD detection module and a proactive listener module. The ensemble AD detection module integrates four classifiers which are based on audio, language, disfluency, and interactivity, respectively, utilizing both utterance and dialogue information for diagnosis. This module resolves the problem of variable spontaneous dialogue by using a two-stage ensemble learning approach to hierarchically balance the effects of four classifiers. The proactive listener module categorizes user speech into three classes: question, statement, and silence, for which particular response types are generated. This approach overcomes the limitation of small-amount speech samples and breakdown dialogues. With the proposed architecture, conversational robots are expected to be applied in healthcare as a solution to the shortage of medical professionals in the aging society.

A limitation we need to take into consideration is that models trained on "normal" speech don't necessarily extend to disordered speech. Each component will still need to be evaluated on AD speech, especially the ones which rely on ASR, such as disfluency and dialogue act. It might be necessary to rely more on prosody using complex models [46]. We need to keep in mind that how good the state-of-the-art of each of the components is. In our future work, we plan to firstly use the dementia dialogue database "Carolina Conversation Collection" [47] to analyze audio, language, disfluency, and interactivity and model the classifiers. An experimental evaluation using the proactive robotic listener will be conducted as our long-term goal.

References

[1] Ritchie, K., et al., 2017. The midlife cognitive profiles of adults at high risk of late-onset Alzheimer's disease: The PREVENT study. Alzheimer's & Dementia, 13(10), pp.1089-1097.

[2] Yu, B., Quatieri, T.F., Williamson, J.R. and Mundt, J.C., 2015. Cognitive impairment prediction in the elderly based on vocal biomarkers. In Sixteenth Annual Conference of the International Speech Communication Association.

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

[4] Haider, F., De La Fuente, S. and Luz, S., 2019. An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. IEEE Journal of Selected Topics in Signal Processing, 14(2), pp.272-281.

[5] Ambrosini, E., et al., 2019, July. Automatic speech analysis to early detect functional cognitive decline in elderly population. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 212-216). IEEE.

[6] Li J., et al., 2021, June. A comparative study of acoustic and linguistic features classification for alzheimer's disease detection. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6423-6427). IEEE.

[7] Luz, S., Haider, F., de la Fuente, S., Fromm, D. and MacWhinney, B., 2020. Alzheimer's dementia recognition through spontaneous speech: the ADReSS Challenge. arXiv preprint arXiv:2004.06833.

[8] Eyben, F., Wöllmer, M. and Schuller, B., 2010, October. Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia (pp. 1459-1462).

[9] Eyben, F., Weninger, F., Gross, F. and Schuller, B., 2013, October. Recent developments in opensmile, the munich open-source multimedia feature extractor. In Proceedings of the 21st ACM international conference on Multimedia (pp. 835-838).

[10] Eyben, F., et al., 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE transactions on affective computing, 7(2), pp.190-202.

[11] Chen, J., Wang, Y. and Wang, D., 2014. A feature study for classification-based speech separation at low signal-to-noise ratios. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(12), pp.1993-2002.

[12] Luz, S., 2017, June. Longitudinal monitoring and detection of Alzheimer's type dementia from spontaneous speech data. In 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS) (pp. 45-46). IEEE.

[13] Qin, Y., et al., 2021. Exploiting Pre-Trained ASR Models for Alzheimer's Disease Recognition Through Spontaneous Speech. arXiv preprint arXiv:2110.01493.

[14] Rohanian, M., Hough, J. and Purver, M., 2021. Alzheimer's Dementia Recognition Using Acoustic, Lexical, Disfluency and Speech Pause Features Robust to Noisy Inputs. arXiv preprint arXiv:2106.15684.

[15] Rohanian, M., Hough, J. and Purver, M., 2021. Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech. arXiv preprint arXiv:2106.09668.

[16] Chinaei, H., Currie, L.C., Danks, A., Lin, H., Mehta, T. and Rudzicz, F., 2017. Identifying and avoiding confusion in dialogue with people with Alzheimer's disease. Computational Linguistics, 43(2), pp.377-406.

[17] Dassa, A. and Amir, D., 2014. The role of singing familiar songs in encouraging conversation among people with middle to late stage Alzheimer's disease. Journal of music therapy, 51(2), pp.131-153.

[18] Li, Y., Ishi, C.T., Inoue, K., Nakamura, S. and Kawahara, T., 2019. Expressing reactive emotion based on multimodal emotion recognition for natural conversation in human–robot interaction. Advanced Robotics, 33(20), pp.1030-1041.

[19] Li, Y., Bell, P. and Lai, C., 2021. Fusing ASR Outputs in Joint Training for Speech Emotion Recognition. arXiv preprint arXiv:2110.15684.

[20] Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Bengio, C.L.Y. and Courville, A., 2017. Towards end-to-end speech recognition with deep convolutional neural networks. arXiv preprint arXiv:1701.02720.

[21] Meghanani, A., Anoop, C.S. and Ramakrishnan, A.G., 2021, January. An exploration of log-mel spectrogram and MFCC features for Alzheimer's dementia recognition from spontaneous speech. In 2021 IEEE Spoken Language Technology Workshop (SLT) (pp. 670-677). IEEE.

[22] Ilias, L., Askounis, D. and Psarras, J., 2021. Detecting Dementia from Speech and Transcripts using Transformers. arXiv preprint arXiv:2110.14769.

[23] Lu, X., 2010. Automatic analysis of syntactic complexity in second language writing. International journal of corpus linguistics, 15(4), pp.474-496.

[24] Pennington, J., Socher, R. and Manning, C.D., 2014, October. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[25] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[26] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[27] Rohanian, M., Hough, J. and Purver, M., 2021. Alzheimer's Dementia Recognition Using Acoustic, Lexical, Disfluency and Speech Pause Features Robust to Noisy Inputs. arXiv preprint arXiv:2106.15684.

[28] Balagopalan, A., Eyre, B., Rudzicz, F. and Novikova, J., 2020. To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection. arXiv preprint arXiv:2008.01551.

[29] Jaffe, J. and Feldstein, S. (1970). Rhythms of dialogue. Personality and Psychopathology. Academic Press, New York.

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

[31] Goberman, A.M., Blomgren, M. and Metzger, E., 2010. Characteristics of speech disfluency in Parkinson disease. Journal of Neurolinguistics, 23(5), pp.470-478.

[32] Sabelli, A.M., Kanda, T. and Hagita, N., 2011, March. A conversational robot in an elderly care center: an ethnographic study. In 2011 6th ACM/IEEE international conference on human-robot interaction (HRI) (pp. 37-44). IEEE.

[33] Yamazaki, R., Kase, H., Nishio, S. and Ishiguro, H., 2019, September. A conversational robotic approach to dementia symptoms: Measuring its effect on older adults. In Proceedings of the 7th International Conference on Human-Agent Interaction (pp. 110-117).

[34] Pou-Prom, C., Raimondo, S. and Rudzicz, F., 2020. A conversational robot for older adults with alzheimer's Disease. ACM Transactions on Human-Robot Interaction (THRI), 9(3), pp.1-25.

[35] Nonaka, Y., Sakai, Y., Yasuda, K. and Nakano, Y., 2012, September. Towards assessing the communication responsiveness of people with dementia. In International Conference on Intelligent Virtual Agents (pp. 496-498). Springer, Berlin, Heidelberg.

[36] Lala, D., Milhorat, P., Inoue, K., Ishida, M., Takanashi, K. and Kawahara, T., 2017, August. Attentive listening system with backchanneling, response generation and flexible turn-taking. In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (pp. 127-136).

[37] Milhorat, P., Lala, D., Inoue, K., Zhao, T., Ishida, M., Takanashi, K., Nakamura, S. and Kawahara, T., 2019. A conversational dialogue manager for the humanoid robot ERICA. In Advanced Social Interaction with Agents (pp. 119-131). Springer, Cham.

[38] Yoshino, K. and Kawahara, T., 2015. Conversational system for information navigation based on POMDP with user focus tracking. Computer Speech & Language, 34(1), pp.275-291.

[39] Li, Y., Zhao, T. and Shen, X., 2020, March. Attention-Based Multimodal Fusion for Estimating Human Emotion in Real-World HRI. In Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (pp. 340-342).

[40] Eyben, F., Wöllmer, M. and Schuller, B., 2010, October. Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia (pp. 1459-1462).

[41] Baevski, A., Zhou, H., Mohamed, A. and Auli, M., 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv preprint arXiv:2006.11477.

[42] Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R. and Mohamed, A., 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. arXiv preprint arXiv:2106.07447.

[43] G. Mesnil, X. He, L. Deng, and Y. Bengio, "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding." in INTERSPEECH, 2013, pp. 3771–3775.

[44] Hara, K., Inoue, K., Takanashi, K. and Kawahara, T., 2018. Prediction of turn-taking using multitask learning with prediction of backchannels and fillers. Listener, 162, p.364.

[45] Nasreen, S., Hough, J. and Purver, M., 2021. Detecting Alzheimer's Disease using Interactional and Acoustic features from Spontaneous Speech. Interspeech.

[46] Tran, T., 2020. Neural Models for Integrating Prosody in Spoken Language Understanding (Doctoral dissertation, University of Washington).

[47] Pope, C. and Davis, B.H., 2011. Finding a balance: The carolinas conversation collection.