2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3503-7149-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMBC53108.2024.10782834

Alzheimer's Disease Classification From Speech Pause Distributions With Context Information

Geet Khatri[1], Reza Soleimani[1], Katarina L. Haley[2], Adam Jacks[2], Edgar Lobaton[1], *Senior Member, IEEE*

***Abstract*—Alzheimer's disease (AD) is known to affect the lengths and frequencies of certain kinds of pauses in speech. Previous studies have used features based on pause lengths for AD classification. We conjecture that in addition to using pause lengths, it is beneficial to incorporate the "context" behind each pause, i.e., what is being said before and after each pause. We propose an AD detection method based on this idea. As part of the proposed method, pause lengths and context are extracted from the raw audio using automatic speech recognition (ASR) and forced alignment. Then, statistical summaries of pause lengths with context information are extracted from the transcripts and used as features for classification. Our results indicate that incorporating the context significantly improves classification performance compared to using pause lengths alone, with classification accuracy of up to 81%. Additionally, the proposed features largely preserve privacy.**

***Index Terms*—Alzheimer's disease, dementia, speech, ASR, forced alignment**

I. Introduction

Screening and early identification of Alzheimer's disease (AD) can help those affected by it get timely treatment, including the newly FDA-approved drug Lecanemab [1], which is given intravenously and removes the culprit protein deposits from the brain in *early* stages of the disease progression. Thus, it is crucial to identify signs of AD as early as possible.

One early feature of AD is difficulties thinking of words when speaking, especially lower frequency words that carry precise meanings. One of the ways this is seen in a person's speech is that they pause as they try to remember the word. For this reason, their speech is expected to have more pauses before more difficult words. They may also have difficulties remembering a story or formulating their thoughts. Previous studies have sought to quantify these effects [2]–[4].

Various approaches have been proposed to detect AD from speech using acoustic features, linguistic features or combinations of both. To a degree, the language difficulties associated with AD can be captured using just text-based linguistic features. However, audio data contains information that text alone cannot capture, including prosody. This information can be useful for detecting or tracking progression of AD. For example, one study [5] incorporated pauses into text-based features, which improved performance over using plain text.

*This work was supported by National Science Foundation (NSF) under awards IIS-1915599 and IIS-2037328.

[1]Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695, USA.; Corresponding author: edgar.lobaton@ncsu.edu.

[2]Division of Speech and Hearing Sciences, Department of Allied Health Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC 27559, USA.

Manual transcription is cumbersome and not a viable option in many real-world applications. When only audio is available, text-based inference is still possible through ASR, which transcribes speech from raw audio. This enables the use of multimodal approaches which integrate acoustic features extracted from audio with linguistic features extracted from ASR-generated transcripts. Examples of ASR models include Silero [6], wav2vec 2.0 [7], HuBERT [8] and Whisper [9]. Typically, ASR does not provide phoneme- or word-level timestamps. Forced alignment can be used to align the ASR-generated transcript to the original audio. Forced alignment has been applied to AD [10] and mild cognitive impairment [11] for multimodal inference.

Deep learning (DL) models are often used for classification and regression due to their flexibility and performance. DL models applied to dementia include multi-layer perceptrons (MLPs) [12], long short-term memory (LSTM) models [13]–[17], convolutional neural networks (CNNs) [18], and transformers like BERT [10], [14], [16], [19]–[23], ERNIE [23], GPT-3 [24] and the vision transformer (ViT) [19]. Many of these models are currently among the state-of-the-art for dementia detection. However, DL models typically require huge amounts of data and tend to be computationally demanding during training. Moreover, DL models that directly take raw audio or transcripts as input can lead to privacy concerns.

In this paper, we propose an AD detection method based on the effect of AD on pauses during speech. We conjecture that in addition to using pause lengths, it is beneficial to incorporate the "context" behind each pause, i.e., what is being said before and after each pause. To keep compute requirements low and address privacy concerns, we limit the context for each pause to words immediately before and after it. The order of words in the transcript is not retained after feature extraction. Potentially sensitive words are not retained either. Our results indicate that incorporating the context behind pauses significantly improves classification performance compared to using pause lengths alone. The rest of this paper is organized as follows: Section II explains the methodology followed, Section III describes the experiments and results, and Section IV consists of remarks and interpretations of the results.

II. Methodology

For our analysis, we will make use of the *Cookie Theft* picture description task from the Pitt corpus in DementiaBank [25]. The corpus contains audio recordings and their transcripts from 104 control participants and 208 participants with Alzheimer's disease (AD). In *Cookie Theft*, there are

243 recordings for the "control" group and 305 recordings for the "dementia" (AD) group. The corpus also contains CHAT-format transcripts [26] with phrase-level timestamps. We obtain the "ground truth" transcripts by processing the CHAT transcripts. To ensure that the proposed method can be used even when only raw audio is available, we use ASR and forced alignment to transcribe the recordings with word-level timestamps.

Our main objective is to develop a classifier between individuals in the "control" and "dementia" groups using the raw audio recordings. Prior to this, we perform tests of statistical significance to better understand what kinds of pauses are relevant for AD detection, which in turn motivates the use of pause distributions for classification. For each kind of pause, we perform a one-sided t-test for the hypothesis that the mean pause length is greater for the "dementia" group.

To build the classifier, we need to first preprocess the audio to recover the transcriptions, extract the desired features, and use those features as input to a machine learning classifier. As can be observed in Fig. 1, during training, the feature extraction pipeline includes pause length extraction, statistical summarization and feature selection.

Aside from our main objective of differentiating between "control" and "dementia" groups, we also aim to answer the following questions through our experiments. How do ASR-generated transcripts compare to the "ground truth" transcripts for inference? Is it better to group pauses by words or by Part-of-Speech (PoS) tags? Is it better to use histogram-based features or quantile-based features? Which features in particular are the most informative? Which classifiers work well for this problem?

A. Preprocessing: Text Transcription and Force Alignment

The feature extraction step relies on transcripts with word-level timestamps. For the ground truth transcripts, only the timestamped phrases are extracted from the CHAT transcripts in the corpus. Other details such as morphological information are removed. ASR and forced alignment are performed using Silero's pre-trained speech-to-text model and the built-in alignment functionality in its decoder [6]. The model is used out-of-the-box without fine-tuning. As a pre-processing step for ASR and forced alignment, all recordings are converted to mono and downsampled to 16 kHz. Part-of-speech (PoS) tagging is performed by the averaged perceptron model [27], [28] based on the Penn Treebank tagset [29].

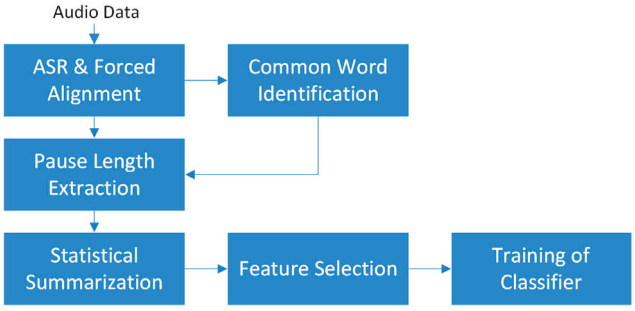

Fig. 1. Block diagram of training. K-fold cross-validation is performed for validation. When the "ground truth" transcripts from the dataset are used, the ASR + forced alignment step is not needed.

Transcripts from the same participant are merged because we ultimately want to extract statistical summaries for participants instead of individual transcripts.

B. Feature Extraction

Pause Extraction. Once we have transcripts with word-level timestamps, we categorize pauses based on what words are adjacent to them. No minimum length is set while defining "pauses", i.e., if there's no silence between two words, the pause is considered to be 0 seconds long. We try out two different approaches for categorizing pauses, described below.

1) Pauses Before/After Common Words: The algorithm for extracting the features is as follows:

a) Identify the W most common words in the dataset, where W is fixed.
b) For each participant's transcript and for each of the W most common words, extract the following:
   i) List of lengths of all pauses occurring immediately before the word in the transcript. Denote this category of pauses by the tuple (word, "before").
   ii) List of lengths of all pauses occurring immediately after the word in the transcript. Denote this category of pauses by the tuple (word, "after").

   Note that for each transcript, each tuple (word, pause position) forms a distinct category of pauses.

We use only common words to ensure that there are enough samples in each category to extract crude statistical features that are not sparse. This should inform the choice of W.

2) Pauses Before/After Part-of-Speech (PoS) Tags: This type of categorization is similar except that all the words are replaced by their equivalent PoS tags, and the iteration is over the list of unique PoS tags instead of common words. Each tuple (tag, pause position) forms a distinct category. In this case, we refer to the number of unique PoS tags as W. The idea behind using PoS tags instead of words is that it may be beneficial to group together words that play similar roles in a sentence.

Based on the above description, the number of categories per transcript is $2W$ for either approach.

Statistical Summarization. Now we describe the statistical features extracted from the pause lengths. Ideally, we want these features to represent pause distributions. Motivated by this, we use two different types of features (separately).

1) Histogram-Based Features: For each category (word/tag, pause position), the features are normalized histogram frequencies, minimum and maximum of pause lengths. The number of bins per histogram, b, is kept fixed, so it is important to explicitly provide range information, which is why minimum and maximum are added as features. In this case, the total number of features per sample is $2W(b+2)$.

2) Quantile-Based Features: The features are q-quantiles of pause lengths, including the 0^{th} and q^{th} quantiles. The total number of features per sample is $2W(q+1)$.

Feature Selection. Univariate feature selection is done by selecting the N best features in terms of mutual information scores, where N is fixed. This step uses only training data so that the testing data does not bias the classifier.

C. Classification

A set of candidate classifiers are trained and evaluated using the training data as part of K-fold cross-validation. The model with the best average accuracy across folds is selected. Finally, the selected model is evaluated on the testing set using accuracy, F1 score, precision and recall. If there is a gap between the testing performance and the average cross-validation performance, the model is discarded.

III. Experiments and Results

A. Tests of Statistical Significance

For each category (word/tag, pause position) described in Section II-B, a one-sided t-test was performed for the hypothesis that the mean pause length is greater for the "dementia" group. $W = 75$ was chosen for common words. For the transcripts in the Pitt corpus, we observed that:

- For 71% of the word categories and 85% of the tag categories, the average pause length is greater for the "dementia" group than for the "control" group. Many of these differences are statistically significant ($p < 0.05$). Fig. 2 shows the most significant of these differences.
- The words with the most statistically significant differences ($p < 0.015$) for the means of pause length before utterance are: "and", "is", "can", "oh", "cookie", "she's", "there's", "he", "he's", "okay", "anything" and "off".
- The words with the most statistically significant differences ($p < 0.015$) for the means of pause length after utterance are: "is", "jar", "cookies", "and", "the", "uh", "mother", "for", "girl", "boy", "to", "um" and "window".
- There are particularly noticeable differences between the two groups in terms of pauses before interjections (UH), pauses before wh-adverbs (WRB), pauses after "to" (TO), and pauses before and after conjunctions (CC) [all with $p \leq 0.012$]. See Fig. 2 for other PoS tags.

These results indicate that pause lengths with word alignment, especially in conjunction with other types of features, could be good indicators of signs of AD. Based on this, we use features derived from these pauses for classification.

B. Classification

For classification, features were extracted according to the method described in Section II-B. The number of participants in the dataset was 290. The train-test split was 80/20. For K-fold cross-validation, K was 5. The following candidate classifiers were used: k nearest neighbors (kNN) with $k = 3$, random forest (RF) with a max. depth of 2, decision tree (DT), support vector classifier (SVC) with a radial basis function

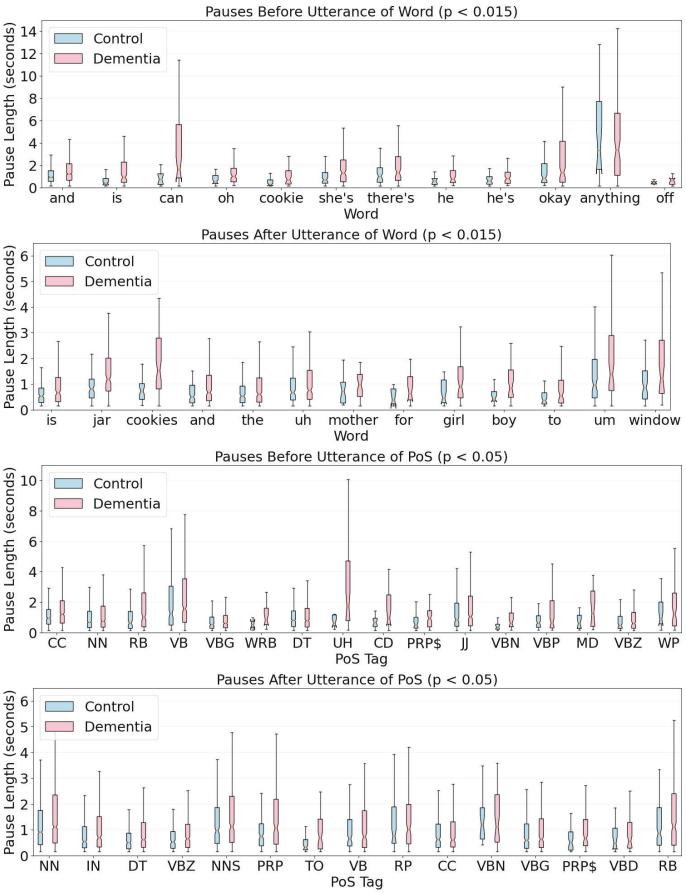

Fig. 2. Boxplots of pauses before/after common words/parts-of-speech (most significant differences only). For better visualization, very short pauses ($<$ 150 ms) were not included in the plots.

(RBF) kernel, linear SVC with L1 penalty, stochastic gradient descent (SGD) classifier, multilayer perceptron (MLP) with a 100-unit hidden layer and ReLU activation, gradient boosting, and XGBoost with a max. depth of 2.

For model selection, our goal is to find not just a good classifier but also good values for parameters such as W and b. Thus, multiple combinations of parameters were experimented with to identify good model candidates. Because the number of all possible combinations is very large, this was done in an iterative fashion by fixing some parameters while varying others. Each combination of parameters was tested with all 9 classifiers. For each combination, only the classifier with the best average accuracy score during cross-validation was retained. Then, to evaluate a particular choice of a parameter, the parameter was kept fixed and aggregate scores (mean, standard deviation [SD] and max. of classifier scores) were computed over variations of other parameters. For example, to evaluate word-based categories, other parameters such as feature type were varied, and then aggregates scores were computed over these variations.

Table I summarizes the results from these experiments. In experiment (1), transcript type, feature type and categorization type were varied while other parameters were fixed ($W = 75$,

TABLE I

RESULTS FROM THE MODEL SELECTION EXPERIMENTS DESCRIBED IN SECTION III-B. MEAN, SD AND MAX. ARE COMPUTED FROM THE BEST PERFORMING CLASSIFIERS FOR DIFFERENT COMBINATIONS OF PARAMETERS. (1), (2) AND (3) DENOTE EXPERIMENT NUMBERS.

	Mean Acc.	***SD of Acc.***	***Max. Acc.***
(1) Word-based cat.	**73.77%**	4.12%	78.87%
PoS tag-based cat.	72.25%	2.9%	74.88%
(2) Histogram-based feat.	**74.78%**	2.46%	77.59%
Quantile-based feat.	71.35%	1.88%	74.93%
(2) Ground truth trans.	72.90%	3.00%	76.12%
ASR transcripts	**73.24%**	2.68%	77.59%
(3) kNN	64.85%	4.43%	73.75%
Random forest	68.12%	1.86%	72.83%
Decision tree	65.09%	2.89%	72.42%
SVC with RBF kernel	72.61%	3.88%	80.19%
Linear SVC	72.77%	2.69%	79.04%
SGD	70.43%	2.94%	76.30%
MLP	**75.44%**	3.11%	80.69%
Gradient boosting	71.91%	2.40%	77.61%
XGBoost	71.09%	2.03%	76.78%

$q = 4$, and histogram-based features with $b = 20$ and log scale for bins). Based on this, word-based categorization was found to work better on average across different combinations of parameters. In experiment (2), transcript type, feature type and W (between 50 and 200) were varied while other parameters were fixed (word-based categorization, $q = 4$ and $b = 20$ with log scale for histogram-based features). Based on this, on average, histogram-based features were found to work better, and ASR-generated transcripts were found to work *slightly* better. In both experiments (1) and (2), feature selection was not performed. In the final experiment (3), the scale of histogram bins was varied between linear and log, W was varied between 100 and 150, b was varied between 8 and 20, N was varied between 200 and 2000 (along with a variation with no feature selection), and ASR-generated transcripts were used. On average, the MLP and SVC were found to work best.

The best performing model that generalized well was MLP with ASR-generated transcripts, histogram-based features, word-based categories, $W = 150$, $b = 8$, linear scale for histogram bins and $N = 1500$ selected features. When this model was evaluated on the testing set, the classification accuracy was 81.03%, F1 score was 0.86, precision was 0.85 and recall was 0.87. During cross-validation, the mean accuracy was 79.78%, mean F1 score was 0.85, mean precision was 0.83 and mean recall was 0.88. Prior to feature selection, the number of features was 3000. Feature selection reduced this to $N = 1500$ features.

With the same set of parameters but using XGBoost instead of MLP, the feature importance scores were noted. The only features with non-zero importance scores were "max pause after 'open"', "bin #4 of pauses before 'window"', "bin #4 of pauses after 'action"', "max pause before 'anything"', and "bin #4 of pauses before 'cookie"'. The testing accuracy was 70.69% and mean cross-validation accuracy was 71.58%.

The performance differences between ground truth and ASR transcripts in experiment (2) were minor, so we repeated experiment (3) but with ground truth transcripts instead of ASR transcripts. The performances were similar. The best model had a testing set accuracy of 81.03% and a mean cross-validation accuracy of 82.25%. It used $W = 200$, $b = 8$, linear scale for histogram bins and $N = 500$ selected features.

TABLE II

COMPARISON OF SELECTED MODEL WITH OTHER MODELS THAT USE THE PITT CORPUS. "GT" REFERS TO USE OF GROUND TRUTH TRANSCRIPTS.

	Accuracy	***F1 Score***	***Precision***	***Recall***
Proposed method (ASR)	81.0%	0.86	0.85	0.87
Proposed method (GT)	81.0%	0.84	0.94	0.76
Pauses without context	57.9%	0.72	0.60	0.91
Haider et al. 2019 [30]	78.7%	0.78	0.80	0.77
Klumpp et al. 2018 [31]	84.4%	-	-	-

When all pause lengths were used without considering the context, the best accuracy was 57.94% using a random forest. This indicates that taking into account the context behind pauses, even something as simple as adjacent words, can significantly improve performance.

Table II shows a comparison of the proposed method with other methods that also used the Pitt corpus. Note that the methodologies differ between the proposed method and other authors' methods. Haider et al. [30] trained decision tree classifiers on different sets of acoustic features (eGeMAPS, ComParE 2013 and MRCG) and fused the results with a vote among the classifiers. By contrast, our method uses a simpler set of features based on pause lengths, making it more amenable to interpretability. They used leave-one-subject-out cross-validation. Klumpp et al. [31] trained a multilayer perceptron with one hidden layer on 546 word frequency features. They used leave-one-subject-out cross-validation as well. Like our method, their method does not preserve word order, but it uses a larger set of words compared to our method. It also reduces words to their stems, which can lead to loss of potentially useful [32] information about word forms.

IV. DISCUSSION

In the proposed method, the number of features is relatively small, with the best-performing model using 1500 features. Moreover, the number of features does not increase with audio length. This keeps computational requirements in check during training/inference. The features also largely preserve privacy. They do not retain any sensitive words or word ordering. Note that it may be possible to recover some word ordering by observing the features corresponding to minimum and maximum pause values, but that would only give away a small number of short combinations of common words at worst. Thus, these features can be included as part of a larger set of privacy-preserving features. A secure local device (e.g., a wearable) can perform transcription, extract such features from the transcript, and send the features to a remote server. The server can then use a larger model to perform inference without having direct access to the audio or transcript.

The results from the tests of statistical significance hint at the following.

- People with AD tend to have longer unvoiced pauses after filler words like "um" and "uh". This might show

that they need more time to think of words or what to say than the time afforded to them by saying "um" or "uh". Neurotypical people use these fillers to buy them more time when they are formulating what to say, and usually the amount of time they take is enough to keep conversing.
- Those with AD tend to have longer pauses before and after the word "is", which might indicate trouble retrieving adjectives and verbs.
- Those with AD tend to pause longer after saying "the", which might indicate trouble retrieving a specific noun.
- Those with AD tend to have longer pauses after words such as "mother", "girl" and "boy". These are relatively high frequency words, but after they have said those words, speakers must explain something about what these characters are doing and how they relate to each other. This might be taxing for word retrieval.

It should be noted that the specific words associated with pauses apply only to the *Cookie Theft* task. Other topics should be evaluated as well to identify topic-specific norms as well as general patterns or principles.

V. Conclusion

In this paper, we proposed a new Alzheimer's disease (AD) detection method. It takes the raw speech audio as input, performs ASR and forced alignment for transcription with word-level timestamps, extracts pause distributions with context information, and finally performs classification using a multilayer perceptron (MLP). The proposed model preserves privacy, uses interpretable features, and is small enough for potential use in local devices like wearables.

References

[1] H. D. Larkin, "Lecanemab gains fda approval for early alzheimer disease," *Jama*, vol. 329, no. 5, pp. 363–363, 2023.

[2] R. A. Sluis, D. Angus, J. Wiles, A. Back, T. A. Gibson, J. Liddle, P. Worthy, D. Copland, and A. J. Angwin, "An Automated Approach to Examining Pausing in the Speech of People With Dementia," *American Journal of Alzheimer's Disease & Other Dementiasr*, vol. 35, p. 153331752093977, Jan. 2020.

[3] A. Pistono, M. Jucla, E. J. Barbeau, L. Saint-Aubert, B. Lemesle, B. Calvet, B. Köpke, M. Puel, and J. Pariente, "Pauses During Autobiographical Discourse Reflect Episodic Memory Processes in Early Alzheimer's Disease," *Journal of Alzheimer's Disease*, vol. 50, no. 3, pp. 687–698, Feb. 2016.

[4] S. Singh, R. S. Bucks, and J. M. Cuerden, "Evaluation of an objective technique for analysing temporal variables in DAT spontaneous speech," *Aphasiology*, vol. 15, no. 6, pp. 571–583, Jun. 2001.

[5] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 2162–2166.

[6] S. Team, "Silero models: pre-trained enterprise-grade stt / tts models and benchmarks," https://github.com/snakers4/silero-models, 2021.

[7] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Oct. 2020, arXiv:2006.11477 [cs, eess].

[8] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, Jul. 2023, pp. 28 492–28 518.

[10] M. Martinc, F. Haider, S. Pollak, and S. Luz, "Temporal Integration of Text Transcripts and Acoustic Features for Alzheimer's Diagnosis Based on Spontaneous Speech," *Frontiers in Aging Neuroscience*, vol. 13, p. 642647, Jun. 2021.

[11] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken Language Derived Measures for Detecting Mild Cognitive Impairment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2081–2090, Sep. 2011.

[12] S. S. H. Zargarbashi and B. Babaali, "A Multi-Modal Feature Embedding Approach to Diagnose Alzheimer Disease from Spoken Language," Oct. 2019, arXiv:1910.00330 [cs, eess, stat].

[13] M. Rohanian, J. Hough, and M. Purver, "Multi-Modal Fusion with Gating Using Audio, Lexical and Disfluency Features for Alzheimer's Dementia Recognition from Spontaneous Speech." ISCA, Oct. 2020, pp. 2187–2191.

[14] A. Roshanzamir, H. Aghajan, and M. Soleymani Baghshah, "Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, p. 92, Dec. 2021.

[15] N. Shivhare, S. Rathod, and M. R. Khan, "Automatic Speech Analysis of Conversations for Dementia Detection Using LSTM and GRU," in *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)*. Nagpur, India: IEEE, Nov. 2021, pp. 1–7.

[16] J. Li, J. Yu, Z. Ye, S. Wong, M. Mak, B. Mak, X. Liu, and H. Meng, "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)*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 6423–6427.

[17] A. Meghanani, A. C. S., and A. G. Ramakrishnan, "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)*. Shenzhen, China: IEEE, Jan. 2021, pp. 670–677.

[18] K. Chlasta and K. Wołk, "Towards Computer-Based Automated Screening of Dementia Through Spontaneous Speech," *Frontiers in Psychology*, vol. 11, p. 623237, Feb. 2021.

[19] L. Ilias and D. Askounis, "Multimodal Deep Learning Models for Detecting Dementia From Speech and Transcripts," *Frontiers in Aging Neuroscience*, vol. 14, p. 830943, Mar. 2022.

[20] J. Chen, J. Ye, F. Tang, and J. Zhou, "Automatic Detection of Alzheimer's Disease Using Spontaneous Speech Only," in *Interspeech 2021*. ISCA, Aug. 2021, pp. 3830–3834.

[21] S. Farzana, A. Deshpande, and N. Parde, "How You Say It Matters: Measuring the Impact of Verbal Disfluency Tags on Automated Dementia Detection," in *Proceedings of the 21st Workshop on Biomedical Language Processing*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 37–48.

[22] Z. Liu, E. J. Paek, S. O. Yoon, D. Casenhiser, W. Zhou, and X. Zhao, "Detecting Alzheimer's Disease Using Natural Language Processing of Referential Communication Task Transcripts," *Journal of Alzheimer's Disease*, vol. 86, no. 3, pp. 1385–1398, Apr. 2022.

[23] J. Yuan, X. Cai, Y. Bian, Z. Ye, and K. Church, "Pauses for Detection of Alzheimer's Disease," *Frontiers in Computer Science*, vol. 2, p. 624488, Jan. 2021.

[24] F. Agbavor and H. Liang, "Predicting dementia from spontaneous speech using large language models," *PLOS Digital Health*, vol. 1, no. 12, p. e0000168, Dec. 2022.

[25] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 06 1994.

[26] A. M. Lanzi, A. K. Saylor, D. Fromm, H. Liu, B. MacWhinney, and M. L. Cohen, "DementiaBank: Theoretical Rationale, Protocol, and Illustrative Analyses," *American Journal of Speech-Language Pathology*, vol. 32, no. 2, pp. 426–438, Mar. 2023.

[27] NLTK, "Source code for nltk.tag.perceptron," https://www.nltk.org/_modules/nltk/tag/perceptron.html, 2015.

[28] M. Honnibal, "A good part-of-speech tagger in about 200 lines of python," https://explosion.ai/blog/part-of-speech-pos-tagger-in-python, 2013.
[29] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.
[30] F. Haider, S. De La Fuente, and S. Luz, "An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, Feb. 2020.
[31] P. Klumpp, J. Fritsch, and E. Noeth, "ANN-based Alzheimer's disease classification from bag of words," in *Speech Communication; 13th ITG-Symposium*, Oct. 2018, pp. 1–4.
[32] L. Ilias and D. Askounis, "Explainable Identification of Dementia From Transcripts Using Transformer Networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4153–4164, Aug. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9769980/