Disfluencies and Fine-Tuning Pre-trained Language Models for Detection of Alzheimer's Disease

Jiahong Yuan[1], *Yuchen Bian*[1], *Xingyu Cai*[1], *Jiaji Huang*[1], *Zheng Ye*[2], *Kenneth Church*[1]

[1]Baidu Research, USA
[2]CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, China

{jiahongyuan, yuchenbian, xingyucai, huangjiaji, kennethchurch}@baidu.com, yez@ion.ac.cn

Abstract

Disfluencies and language problems in Alzheimer's Disease can be naturally modeled by fine-tuning Transformer-based pre-trained language models such as BERT and ERNIE. Using this method, we achieved 89.6% accuracy on the test set of the ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) Challenge, a considerable improvement over the baseline of 75.0%, established by the organizers of the challenge. The best accuracy was obtained with ERNIE, plus an encoding of pauses. Robustness is a challenge for large models and small training sets. Ensemble over many runs of BERT/ERNIE fine-tuning reduced variance and improved accuracy. We found that *um* was used much less frequently in Alzheimer's speech, compared to *uh*. We discussed this interesting finding from linguistic and cognitive perspectives.

Index Terms: Alzheimer's disease, disfluency, BERT, ERNIE, ensemble

1. Introduction

Alzheimer's disease (AD) involves a progressive degeneration of brain cells that is irreversible [1]. Therefore, early diagnosis and intervention is essential. One of the first signs of the disease is deterioration in language and speech production [2]. Case studies of the writings of the British Novelist Iris Murdoch indicated that lexical and syntactic changes occurred in the early stage of her AD [3]. Similarly, a study of President Ronald Regan's non-scripted news conferences found decreases in unique words and increases in conversational fillers and non-specific nouns well before his diagnosis of AD [4].

It is desirable to use language and speech for AD detection [5]. The ADReSS challenge of INTERSPEECH 2020 is "to define a shared task through which different approaches to AD detection, based on spontaneous speech, could be compared" [6]. This paper describes our effort for the shared task.

1.1. Studies of speech and language in AD and AD detection

There is an extensive literature on the characteristics of language and speech production in people with AD at various stages of the disease. Summaries of the studies can be found in [7, 8, 9]. Language impairment in AD is most evident in lexical, semantic, and pragmatic aspects. For example, people with AD often produce semantically "empty" words (e.g., *thing*, *stuff*) [10], use fewer information-bearing nouns and especially verbs [11], and their discourse appears to be disorganized [12]. Other aspects (syntax, phonology, and articulation) are believed to be relatively well preserved until late stages of the disease [13], though this conclusion is controversial [14, 15].

Many language problems cause disfluency in connected speech. Disfluencies are also common in normal spontaneous speech [16]. There are various types of disfluencies such as repetitions, false starts, repairs, filled and unfilled pauses. The phonetic consequence of speech disfluency has been well studied [17]. English has two common filled pauses, *uh* and *um*. There is a debate in the literature as to whether *uh* and *um* are intentionally produced by speakers [18, 19]. From sociolinguistic point of view, women and younger people tend to use more *um* vs. *uh* than men and older people [20, 21]. It has also been reported that autistic children use *um* less frequently than normal children [22, 23], and that *um* occurs less frequently and is shorter during lying compared to truth-telling [24]. It will be interesting to examine whether the use of *uh* and *um* in AD speech is different from normal speech. We did a preliminary investigation on this question, which is reported in Section 2. Although disfluencies are a part of normal speech, there is a boundary between normal and abnormal disfluencies. The boundary resides in a high dimensional space, determined by many interrelated factors such as pauses, repetitions, linguistic errors, discourse incoherence, etc. Classification of AD and normal speech requires a model that can capture these factors.

There is a considerable literature on AD detection from continuous speech [25, 26]. This literature considers a wide variety of features and machine learning techniques. [27] used 370 acoustic and linguistic features to train logistic regression models for classifying AD and normal speech. [28] found that acoustic and linguistic features were about equally effective for AD classification, but the combination of the two performed better than either by itself. Neural network models such as Convolutional Neural Networks and Long Short-Term Memory (LSTM) have also been employed for the task [29, 30, 31], and very promising results have been reported. However, it is difficult to compare these different approaches, because of the lack of standardized training and test data sets. One objective of the ADReSS challenge is to overcome this obstacle [6].

1.2. Pre-trained LMs and Self-attention

Modern pre-trained language models such as BERT [32] and ERNIE [33] were trained on extremely large corpora. These models appear to capture a wide range of linguistic facts including lexical knowledge, phonology, syntax, semantics and pragmatics. Recent literature is reporting considerable success on a variety of benchmark tasks with BERT and BERT-like models.[1] We expect that the language characteristics of AD can also be captured by the pre-trained language models when fine-tuned to the task of AD classification.

[1] https://gluebenchmark.com

BERT and BERT-like models are based on the Transformer architecture [34]. These models use self-attention to capture associations among words. Each attention head operates on the elements in a sequence (e.g., words in the transcript for a subject), and computes a new sequence of the weighed sum of (transformed) input elements. There are various versions of BERT and ERNIE. There is a base model with 12 layers and 12 attention heads for each layer, as well as a larger model with 24 layers and 16 attention heads for each layer. Conceptually the self-attention mechanism can naturally model many language problems in AD mentioned in Section 1.1, including repetitions of words and phrases, use of particular words (and classes of words), as well as pauses. We proposed a method to encode pauses in a word sequence to enable BERT-like models to take advantage of disfluencies involving pauses, described in Section 3.1.

Previous studies have found that when fine tuning BERT for downstream tasks with a small data set, the model has a high variance in performance. Even with the same hyperparameter values, distinct random seeds can lead to substantially different results. [35] conducted a large-scale study on this issue. They fine-tuned BERT hundreds of times while varying only the random seeds, and found that the best-found model significantly outperformed previous reported results using the same model. In this situation, using just one final model for prediction is risky given the variance in performance during training. We propose an ensembling method to address this concern.

2. Data and analysis

2.1. Data

The data consists of speech recordings and transcripts of descriptions of the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [36]. Transcripts were annotated using the CHAT coding system [37]. We only used word transcripts, the morphological and syntactic annotations in the transcripts were not used in our experiments.

The training set contains 108 speakers, and the test set contains 48 speakers. In each data set, half of the speakers are people with AD and half are non-AD (healthy control subjects). Both data sets were provided by the challenge. The organizers also provided speech segments extracted from the recordings using a simple voice detection algorithm, but no transcripts were available for the speech segments. We didn't use these speech segments. Our experiments were based on the entire recordings and transcripts.

2.2. Processing transcripts and forced alignment

The transcripts in the data sets were annotated in the CHAT format, which can be conveniently created and analyzed using CLAN [37]. For example: "the [x 3] bench [: stool]." In this example, [x 3] indicates that the word 'the' was repeated three times, [: stool] indicates that the preceding word, "bench" (which was actually produced), refers to stool. Details of the transcription format can be found in [37].

For the purpose of forced alignment and fine tuning, we converted the transcripts into words and tokens that represent what were actually produced in speech. 'w [x n]' were replaced by repetitions of w for n times, punctuation marks and various comments annotated between '[]' were removed. Symbols such as (.), (..), (...), <, >, / and xxx were also removed.

The processed transcripts were forced aligned with speech recordings using the Penn Phonetics Lab Forced Aligner [38].

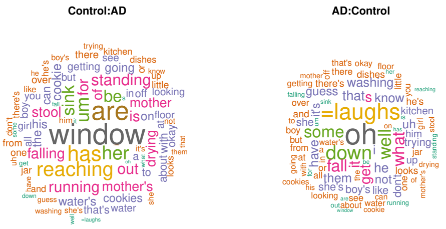

Figure 1: *The word cloud on the left highlights words that are more common among control subjects than AD; the word cloud on the right highlights words that are more common among AD than control.*

Table 1: *Subjects with AD say uh more often, and um less often.*

	uh	*um*
Control (non-AD)	130	51
Dementia (AD)	183	20

The aligner used a special model 'sp' to identify between-word pauses. After forced alignment, the speech segments that belong to the interviewer were excluded. The pauses at the beginning and the end of the recordings were also excluded. Only the subjects' speech, including pauses in turn-taking between the interviewer and the subject, were used.

2.3. Word frequency and *uh/um*

From the training data set, we calculated word frequencies for the Control and AD groups respectively. Words that appear 10 or more times in both groups are shown in the word clouds in Figure 1. The following words are at least two times more frequent in AD than in Control: *oh* (4.33), *=laughs* (laughter, 3.18), *down* (2.66), *well* (2.42), *some* (2.2), *what* (2.16), *fall* (2.15). And the words that are at least two times more frequent in Control than in AD are: *window* (4.4), *are* (3.83), *has* (3.0), *reaching* (2.8), *her* (2.62), *um* (2.55), *sink* (2.3), *be* (2.21), *standing* (2.06).

Compared to controls, subjects with AD used relatively more laughter and semantically "empty" words such as *oh*, *well*, and *some*, and fewer present particles (*-ing* verbs). This is consistent with the literature as discussed in Section 1.1. Table 1 shows an interesting difference for filled pauses. The subjects with AD used more *uh* than the control subjects, but their use of *um* was much less frequent.

2.4. Unfilled pauses

Durations of pauses were calculated from forced alignment. Pauses under 50 ms were excluded, as well as pauses in the interviewer's speech. We binned the remaining pauses by duration as shown in Figure 2. Subjects with AD have more pauses in every group, but the difference between subjects with AD and non-AD is particularly noticeable for longer pauses.

3. BERT and ERNIE Fine-tuning

3.1. Input and Hyperparameters

Pre-trained BERT and ERNIE models were fine-turned for the AD classification task. Each of the $N = 108$ training speakers is considered a data point. The input to the model consists

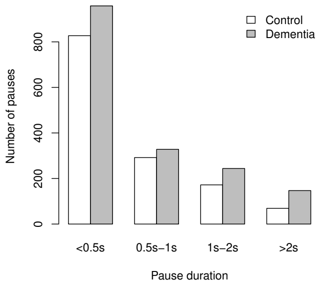

Figure 2: *Subjects with AD have more pauses (in all duration bins).*

of a sequence of words from the processed transcript for every speaker (as described in Section 2.2). The output is the class of the speaker, 0 for Control and 1 for AD.

We also encoded pauses in the input word sequence. We grouped pauses into three bins: short (under 0.5 sec); medium (0.5-2 sec); and long (over 2 sec). The three bins of pauses are coded using three punctuations ",", ".", and "…", respectively. Because all punctuations were removed from the processed transcripts, these inserted punctuations only represent pauses. Two examples of the input text are given below:

1. S136 (AD): *well your , sink is being run over , the . water , the stool the kid's standing on , is , falling and he's getting , cookies from a jar , the ... lady's washing ... dishes . the ... girl's reaching for a cookie ... could , there , be . more , i don't . think so .*
2. S062 (non-AD): *well there's a kid , stealing cookies from the cookie jar and his stool's about to topple over his , his sister's . asking for one the ... cookie jar is open of course the cupboard's open . the , mother's drying dishes the sink is overflowing . there are some , dishes on the side board . window's open i don't ... know , what else you want , there are curtains in the window i don't know if there's any .*

We used Bert-for-Sequence-Classification[2] for fine tuning. We tried both "bert-base-uncased" and "bert-large-uncased," and found slightly better performance with the larger model. The following hyperparameters (slightly tuned) were chosen: learning rate = 2e-5, batch size = 4, epochs = 8, max input length of 256 (sufficient to cover most cases). The standard default tokenizer was used (with an instruction not to split "..."). Two special tokens, [CLS] and [SEP], were added to the beginning and the end of each input.

ERNIE fine-tuning started with the "ERNIE-large" pre-trained model (24 layers with 16 attention heads per layer). We used the default tokenizer, and the following hyperparameters: learning rate = 2e-5, batch size = 8, epochs = 20 and max input length of 256.

3.2. Ensemble Reduces Variance in LOO Accuracy

When conducting LOO (leave-one-out) cross-validation on the training set, large differences in accuracy across runs were observed, as illustrated in Figure 3. The black lines in Figure 3 were computed over 50 runs of BERT3p (top) and 50 runs of ERNIE0p and 50 runs of ERNIE3p (bottom). 0p indicates that no pause was encoded, and 3p indicates that three lengths of pauses were encoded. Each run reports a leave-one-out (LOO) accuracy. Everything was the same across runs except for random seeds. Over the 50 runs, LOO accuracy ranged from 0.75 to 0.86 for BERT3p, from 0.78 to 0.87 for ERNIE3p, and from 0.77 to 0.85 for ERNIE0p. The large variance suggests performance on unseen data is likely to be brittle. Such brittleness is to be expected given the large size of the BERT and ERNIE models and the small size of the training set (108 subjects).

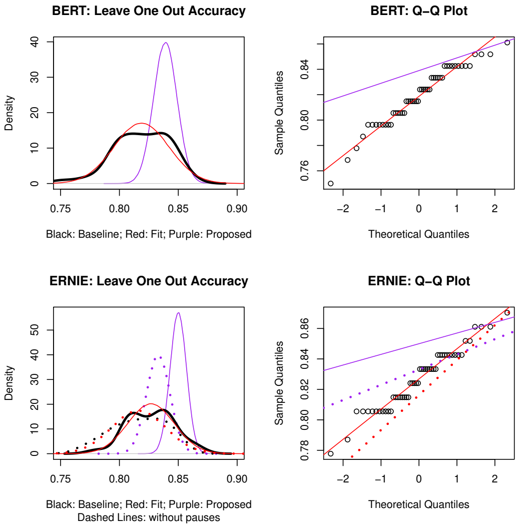

Figure 3: *We computed 50 estimates of leave-one-out (LOO) accuracy for BERT with pauses (top) and ERNIE with and without pauses (bottom). There is a wide variance in both cases (black). The proposed ensemble method (purple) improves the mean and reduces variance. Pauses are useful. Solid lines (with pauses) are better than dashed lines (without pauses).*

To address this brittleness, we introduced the following ensemble procedure. From the results of LOO cross validation, we calculated the majority vote over 50 runs for each of the $N = 108$ subjects, and used the majority vote to return a single label for each subject. Tables 2-3 and Figure 3 show that this ensemble procedure improves the mean and reduces the standard deviation over estimates based on a single run.

To make sure that the ensemble estimates would generalize to unseen data, we tested the method by selecting $N = 5$, $N = 15$, ..., runs from the 50 runs reported in Figure 3. The results in the first row of Table 2 summarize 100 draws of $N = 5$ runs. The second row is similar, except $N = 15$. All of the rows in Table 2 have better means and less variance than the black line in Figure 3. Table 3 is like Table 2, except the means are even better with ERNIE than BERT. From Table 3 and Figure 3, we can also see that results with pauses are better than results without pauses.

4. Evaluation

Under the rules of the challenge, each team is allowed to submit results of five attempts for evaluation. Predictions on the test set from the following five models were submitted for evaluation: BERT0p, BERT3p, BERT6p, ERNIE0p, and ERNIE3p. To compare with three pauses, 6p represents six bins of pauses, encoded as: "," (under 0.5 sec), "." (.5-1 sec); ".." (1-2 sec), ". . ." (2-3 sec), ". . . ." (3-4 sec), ". . . . ." (over than 4 sec). The

[2] https://github.com/huggingface/transformers

Table 2: *Ensemble improves LOO (leave-one-out) estimates of accuracy; better means with less variance.*

	BERT with Three Pauses	
N	mean ± sd	min - max
5	0.837 ± 0.010	0.815 - 0.861
15	0.840 ± 0.011	0.815 - 0.861
25	0.839 ± 0.011	0.815 - 0.870
35	0.838 ± 0.010	0.824 - 0.861
45	0.839 ± 0.011	0.824 - 0.861

Table 3: *Ensemble also improves LOO for ERNIE (with and without pauses). LOO results are better with pauses than without, and better with ERNIE than BERT.*

	ERNIE with Three Pauses		ERNIE with No Pauses	
N	Mean ± Std	Min - Max	Mean ± Std	Min - Max
5	0.845 ± 0.013	0.806 - 0.880	0.828 ± 0.016	0.796 - 0.870
15	0.851 ± 0.008	0.833 - 0.870	0.831 ± 0.012	0.796 - 0.861
25	0.853 ± 0.007	0.833 - 0.870	0.833 ± 0.010	0.815 - 0.861
35	0.854 ± 0.007	0.824 - 0.861	0.836 ± 0.009	0.815 - 0.852
45	0.854 ± 0.007	0.833 - 0.861	0.834 ± 0.008	0.815 - 0.861

dots are separated from each other, as different tokens.

Following the method proposed in Section 3.2, we made 35 runs of training for each of the five models, with 35 random seeds. The classification of each sample in the test set was based on the majority vote of 35 predictions. Table 4 lists the evaluation scores received from the organizers.

The best accuracy was 89.6%, obtained with ERNIE and three pauses. It is a nearly 15% increase from the baseline of 75.0% [6].

ERNIE outperformed BERT by 4% on input of both three pauses and no pause. Encoding pauses improved the accuracy for both BERT and ERNIE. There was no difference between three pauses and six pauses in terms of improvement in accuracy.

5. Discussion

The group with AD used more *uh* but less *um* than the control group. In speech production, disfluencies such as hesitations and speech errors are correlated with cognitive functions such cognitive load, arousal, and working memory [24, 39]. Hesitations and disfluencies increase with increased cognitive load and arousal as well as impaired working memory. This may explain why the group with AD used more *uh*, as a filled pause and hesitation marker. More interestingly, they used less *um* than the control group. This indicates that unlike *uh*, *um* is more than a hesitation marker. Previous studies have also reported that children with autism spectrum disorder produced *um* less frequently than typically developed children [22, 23], and that *um* was used less frequently during lying compared to truth-telling [24, 40]. All these results seem to suggest that *um* carries a lexical status and is retrieved in speech production. One possibility is that people with AD or autism have difficulty in retrieving the word *um* whereas people who are lying try not to use this word. More research is needed to test this hypothesis.

Table 4: *Evaluation results: Best accuracy (acc) with ERNIE and three pauses (3p). Pauses are helpful: three pauses (3p) and six pauses (6p) have better accuracy than no pauses (0p).*

	Precision		Recall		F1		Acc
	non-AD	AD	non-AD	AD	non-AD	AD	
Baseline[6]	0.670	0.600	0.500	0.750	0.570	0.670	0.625
BERT0p	0.742	0.941	0.958	0.667	0.836	0.781	0.813
BERT3p	0.793	0.947	0.958	0.750	0.868	0.837	0.854
BERT6p	0.793	0.947	0.958	0.750	0.868	0.837	0.854
ERNIE0p	0.793	0.947	0.958	0.750	0.868	0.837	0.854
ERNIE3p	0.852	0.952	0.958	0.833	0.902	0.889	**0.896**

From our results, encoding pauses in the input was helpful for both BERT and ERINE fine-tuning for the task of AD classification. Pauses are ubiquitous in spoken language. They are distributed differently in fluent, normally disfluent, and abnormally disfluent speech. As we can see from Figure 2, the group with AD used more pauses and especially more long pauses than the control group. With pauses present in the text, the self-attention mechanism in BERT and ERNIE may learn how the pauses are correlated with other words, for example, whether there is a long pause between the determiner *the* and the following noun, which occurs more frequently in AD speech. We think this is part of the reason why encoding pauses improved the accuracy. Both BERT and ERNIE were pre-trained on text corpora, with no pause information. Our study suggests that it may be useful to pre-train a language model using speech transcripts (either solely or combined with text corpora) that include pause information.

6. Conclusions

Accuracy of 89.6% was achieved on the test set of the ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) Challenge, with ERNIE fine-tuning, plus an encoding of pauses. There is a high variance in BERT and ERNIE fine-tuning on a small training set. Our proposed ensemble method improves the accuracy and reduces variance in model performance. Pauses are useful in BERT and ERNIE fine-tuning for AD classification. *um* was used much less frequently in AD, suggesting that it may have a lexical status.

7. Acknowledgements

We thank Julia Li and Hao Tian for their suggestion and help with ERNIE.

8. References

[1] M. P. Mattson, "Pathways towards and away from alzheimer's disease," *Nature*, vol. 430, pp. 631–639, 2004.

[2] K. D. Mueller, R. L. Koscik, B. Hermann, S. C. Johnson, and L. S. Turkstra, "Declines in connected language are associated with very early mild cognitive impairment: Results from the wisconsin registry for alzheimer's prevention," *Frontiers in Aging Neuroscience*, vol. 9, 2018.

[3] P. Garrard, L. Maloney, J. R. Hodges, and K. Patterson, "The effects of very early alzheimer's disease on the characteristics of writing by a renowned author." *Brain : a journal of neurology*, vol. 128 Pt 2, pp. 250–60, 2005.

[4] V. Berisha, S. Wang, A. LaCross, and J. M. 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 : JAD*, vol. 45 3, pp. 959–63, 2015.

[5] C. Laske, H. R. Sohrabi, S. Frost, K. L. de Ipiña, and S. E. O'Bryant, "Innovative diagnostic tools for early detection of alzheimer's disease," *Alzheimer's & Dementia*, vol. 11, pp. 561–578, 2015.

[6] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020. [Online]. Available: https://arxiv.org/abs/2004.06833

[7] V. O. B. Emery, "Language impairment in dementia of the alzheimer type: a hierarchical decline?" *International journal of psychiatry in medicine*, vol. 30 2, pp. 145–64, 2000.

[8] G. Szatlóczki, I. Hoffmann, V. Vincze, J. Kálmán, and M. Pákáski, "Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease," *Frontiers in Aging Neuroscience*, vol. 7, 2015.

[9] A. Slegers, R.-P. Filiou, M. Montembeault, and S. M. Brambati, "Connected speech features from picture description in alzheimer's disease: A systematic review," *Journal of Alzheimer's disease*, vol. 65 2, pp. 519–542, 2018.

[10] D. Kempler, "Language changes in dementia of the alzheimer type," in *Dementia and Communication*, R. Lubinski, Ed. Philadelphia: B. C. Decker, 1991, ch. 7, pp. 98–113.

[11] M. M. Kim and C. K. Thompson, "Verb deficits in alzheimer's disease and agrammatism: Implications for lexical organization," *Brain and Language*, vol. 88, pp. 1–20, 2004.

[12] M. Mentis, J. Briggs-Whittaker, and G. D. Gramigna, "Discourse topic management in senile dementia of the alzheimer's type," *Journal of speech and hearing research*, vol. 38 5, pp. 1054–66, 1995.

[13] D. Kempler, S. Curtiss, and C. Jackson, "Syntactic preservation in alzheimer's disease." *Journal of speech and hearing research*, vol. 30 3, pp. 343–50, 1987.

[14] K. Croot, J. R. Hodges, J. H. Xuereb, and K. Patterson, "Phonological and articulatory impairment in alzheimer's disease: A case series," *Brain and Language*, vol. 75, pp. 277–309, 2000.

[15] L. Altmann, D. Kempler, and E. Andersen, "Speech errors in alzheimer's disease : reevaluating morphosyntactic preservation," *Journal of Speech Language and Hearing Research*, vol. 44, pp. 1069–82, 2001.

[16] E. Shriberg, "Preliminaries to a theory of speech disfluencies," phd, University of California, Berkeley, 1994. [Online]. Available: http://www.speech.sri.com/people/ees/publications.html

[17] ——, "To 'errrr' is human: Ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association*, vol. 31, pp. 153 – 169, 06 2001.

[18] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, pp. 73–111, 2002.

[19] M. Corley and O. Stewart, "Hesitation disfluencies in spontaneous speech: The meaning of um," *Language and Linguistics Compass*, vol. 2, pp. 589–602, 07 2008.

[20] G. Tottie, "Uh and um as sociolinguistic markers in british english," *International Journal of Corpus Linguistics*, vol. 16, pp. 173–197, 2011.

[21] M. Wieling, J. Grieve, G. Bouma, J. Fruehwald, J. Coleman, and M. Liberman, "Variation and change in the use of hesitation markers in germanic languages," *Language Dynamics and Change*, vol. 6, no. 2, pp. 199–234, 2016.

[22] C. A. Irvine, I.-M. Eigsti, and D. Fein, "Uh, um, and autism: Filler disfluencies as pragmatic markers in adolescents with optimal outcomes from autism spectrum disorder," *Journal of Autism and Developmental Disorders*, vol. 46, pp. 1061–1070, 2016.

[23] K. Gorman, L. Olson, A. Hill, R. Lunsford, P. Heeman, and J. Santen, "Uh and um in children with autism spectrum disorders or language impairment," *Autism research : official journal of the International Society for Autism Research*, vol. 9, pp. 854–865, 2016.

[24] J. Arciuli, D. MALLARD, and G. Villar, ""um, i can tell you're lying": Linguistic markers of deception versus truth-telling in speech," *Applied Psycholinguistics*, vol. 31, pp. 397 – 411, 07 2010.

[25] R.-P. Filiou, N. Bier, A. Slegers, B. Houzé, P. Belchior, and S. M. Brambati, "Connected speech assessment in the early detection of alzheimer's disease and mild cognitive impairment: a scoping review," *Aphasiology*, pp. 1–33, 2019.

[26] M. L. B. Pulido, J. B. A. Hernández, M. A. F. Ballester, C. González, J. Mekyska, and Z. Smékal, "Alzheimer's disease and automatic speech analysis: A review," *Expert Systems With Applications*, vol. 150, p. 113213, 2020.

[27] 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 2, pp. 407–22, 2016.

[28] G. Gosztolya, V. Vincze, L. Toth, M. Pakaski, J. Kalman, and I. Hoffmann, "Identifying mild cognitive impairment and mild alzheimer's diseasebased on spontaneous speech using asr and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181–197, 2019.

[29] J. Fritsch, S. Wankerl, and E. Nöth, "Automatic diagnosis of alzheimer's disease using neural network language models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5841–5845.

[30] F. D. Palo and N. Parde, "Enriching neural models with targeted features for dementia detection," in *ACL*, 2019.

[31] K. L. de Ipiña, U. M. de Lizarduy, P. M. Calvo, B. Beitia, J. Garcia-Melero, M. Ecay-Torres, A. Estanga, and M. Faúndez-Zanuy, "Analysis of disfluencies for automatic detection of mild cognitive impartment: a deep learning approach," *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI)*, pp. 1–4, 2017.

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

[33] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "Ernie 2.0: A continual pre-training framework for language understanding," *arXiv preprint arXiv:1907.12412*, 2019.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[35] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith, "Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping," *arXiv preprint arXiv:2002.06305*, 2020.

[36] E. K. H. Goodglass and B. Barresi, *Boston Diagnostic Aphasia Examination – Third Edition*. Philadelphia: Lippincott Williams & Wilkins, 2001.

[37] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[38] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," *The Journal of the Acoustical Society of America*, vol. 123, p. 3878, 2008.

[39] M. Daneman, "Working memory as a predictor of verbal fluency," *Journal of Psycholinguistic Research*, vol. 20, pp. 445–464, 1991.

[40] S. Benus, F. Enos, J. Hirschberg, and E. Shriberg, "Pauses in deceptive speech," in *Speech Prosody 2006*, 2006.