frontiers in Computer Science

ORIGINAL RESEARCH
published: 29 January 2021
doi: 10.3389/fcomp.2020.624488

Pauses for Detection of Alzheimer's Disease

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

[1]Baidu Research, Sunnyvale, CA, United States, [2]Institute of Neuroscience, Key Laboratory of Primate Neurobiology, Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China

Pauses, 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 with pause-encoded transcripts, we achieved 89.6% accuracy on the test set of the ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) 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.

Keywords: Alzheiemer's disease, pause, BERT, ERNIE, ensemble

OPEN ACCESS

Edited by:
Saturnino Luz,
University of Edinburgh,
United Kingdom

Reviewed by:
Jiri Pribil,
Slovak Academy of Sciences, Slovakia
Shane Sheehan,
University of Edinburgh,
United Kingdom

***Correspondence:**
Jiahong Yuan
jiahongyuan@baidu.com

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *31 October 2020*
Accepted: *11 December 2020*
Published: *29 January 2021*

Citation:
Yuan J, Cai X, Bian Y, Ye Z and Church K (2021) Pauses for Detection of Alzheimer's Disease. Front. Comput. Sci. 2:624488. doi: 10.3389/fcomp.2020.624488

1 INTRODUCTION

Alzheimer's disease (AD) involves a progressive degeneration of brain cells that is irreversible (Mattson, 2004). One of the first signs of the disease is deterioration in language and speech production (Mueller et al., 2017). It is desirable to use language and speech for AD detection (Laske et al., 2015). In this paper, we investigate the use of pauses in speech (both unfilled and filled pauses such as "uh" and "um") for this task.

1.1 Pauses

Unfilled pauses play an important role in speech. The occurrence of pauses is subject to physiological, linguistic, and cognitive constraints (Goldman-Eisler, 1961; Rochester, 1973; Butcher, 1981; Zellner, 1994; Clark, 2006; Ramanarayanan et al., 2013; Hawthorne and Gerken, 2014). How different constraints interact in pause production has been an active research subject for decades. In normal speech, the likelihood of pause occurrence and the duration of pauses are correlated with syntactic and prosodic structure (Brown and Miron, 1971; Grosjean et al., 1971; Krivokapic, 2007). For example, if a sentence has a syntactically complex subject and a syntactically complex object, speakers tend to pause at the subject-verb phrase boundary, and pause duration increases with upcoming complexity (Ferreira, 1991). It has been demonstrated that pauses in speech are used by listeners in sentence parsing (Schepman and Rodway, 2000), and the pause information can benefit automatic parsing (Tran et al., 2018).

Atypical pausing is characteristic of disordered speech such as in Alzheimer's disease, and pauses are often used to measure language and speech problems (Ramig et al., 1995; Yuan et al., 2016; Shea and Leonard, 2019). The difference between typical and atypical pauses is not only on their frequency and duration, but also on where they occur. In this study, we propose a method to encode pauses in transcripts in order to capture the associations between pauses and words through fine-tuning pre-trained language models such as BERT [19] and ERNIE [20], which we describe in **Section 1.2**.

The use of filled pauses may also be different between AD and normal speech. 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 (Clark and Fox Tree, 2002; Corley and Stewart, 2008). From sociolinguistic point of view, women and younger people tend to use more *um* vs. *uh* than men and older people (Tottie, 2011; Wieling et al., 2016). It has also been reported that autistic children use *um* less frequently than normal children (Gorman et al., 2016; Irvine et al., 2016), and that *um* occurs less frequently and is shorter during lying compared to truth-telling (Arciuli et al., 2010).

1.2 Pre-trained LMs and Self-Attention

Modern pre-trained language models such as BERT (Devlin et al., 2018) and ERNIE (Sun et al., 2019) 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.

BERT and BERT-like models are based on the Transformer architecture (Vaswani et al., 2017). 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, including repetitions of words and phrases, use of particular words (and classes of words), as well as pauses. By inserting pauses in word transcripts, we enable BERT-like models to learn the language problems involving pauses.

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. Dodge et al. (2020) 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 ensemble method to address this concern.

1.3 Automatic Detection of AD

There is a considerable literature on AD detection from continuous speech (Filiou et al., 2019; Pulido et al., 2020). This literature considers a wide variety of features and machine learning techniques. Fraser et al. (2016) used 370 acoustic and linguistic features to train logistic regression models for classifying AD and normal speech. Gosztolya et al. (2019) 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 (de Ipiña et al., 2017; Fritsch et al., 2019; Palo and Parde, 2019), 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. 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" (Luz et al., 2020). This paper stems from our effort for the shared task.

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 (Goodglass et al., 2001). Transcripts were annotated using the CHAT coding system (MacWhinney, 2000). 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 (MacWhinney, 2000). 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 (MacWhinney, 2000).

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 (Yuan and Liberman, 2008). The aligner used a special model "sp" to

[1] https://gluebenchmark.com

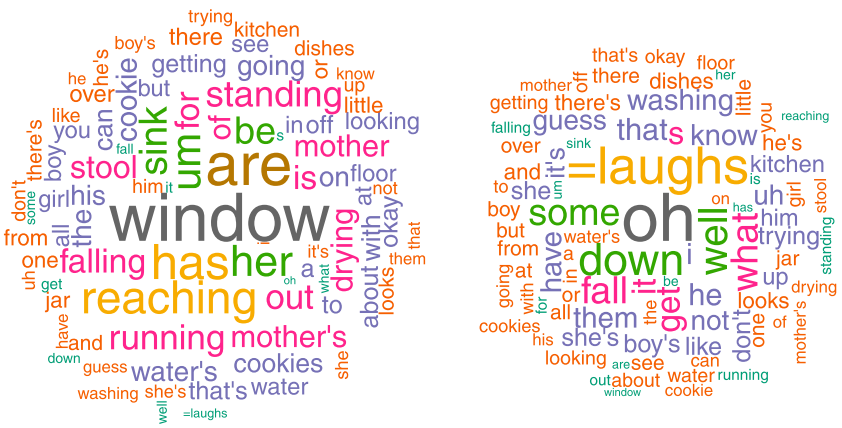

FIGURE 1 | The word cloud on the left highlights words that are more common among control subjects than AD; the word cloud on the right highlights words that are more common among AD than control.

TABLE 1 | Subjects with AD say uh more often, and um less often.

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

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 findings in the literature. **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

Duration of pauses was 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.

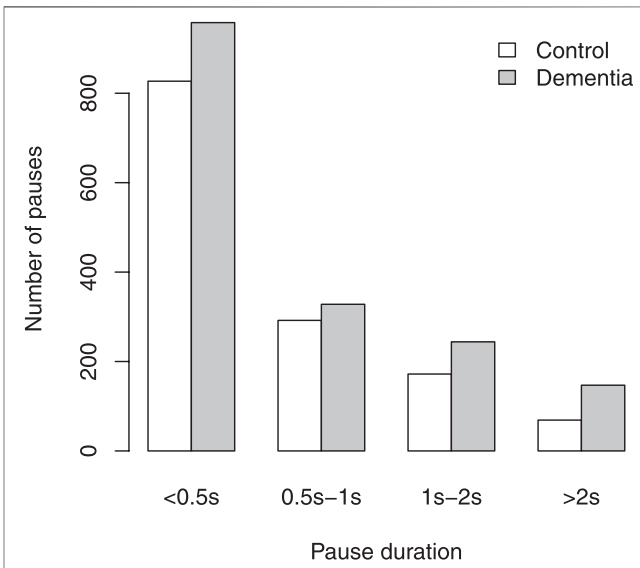

FIGURE 2 | Subjects with AD have more pauses (in all duration bins).

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

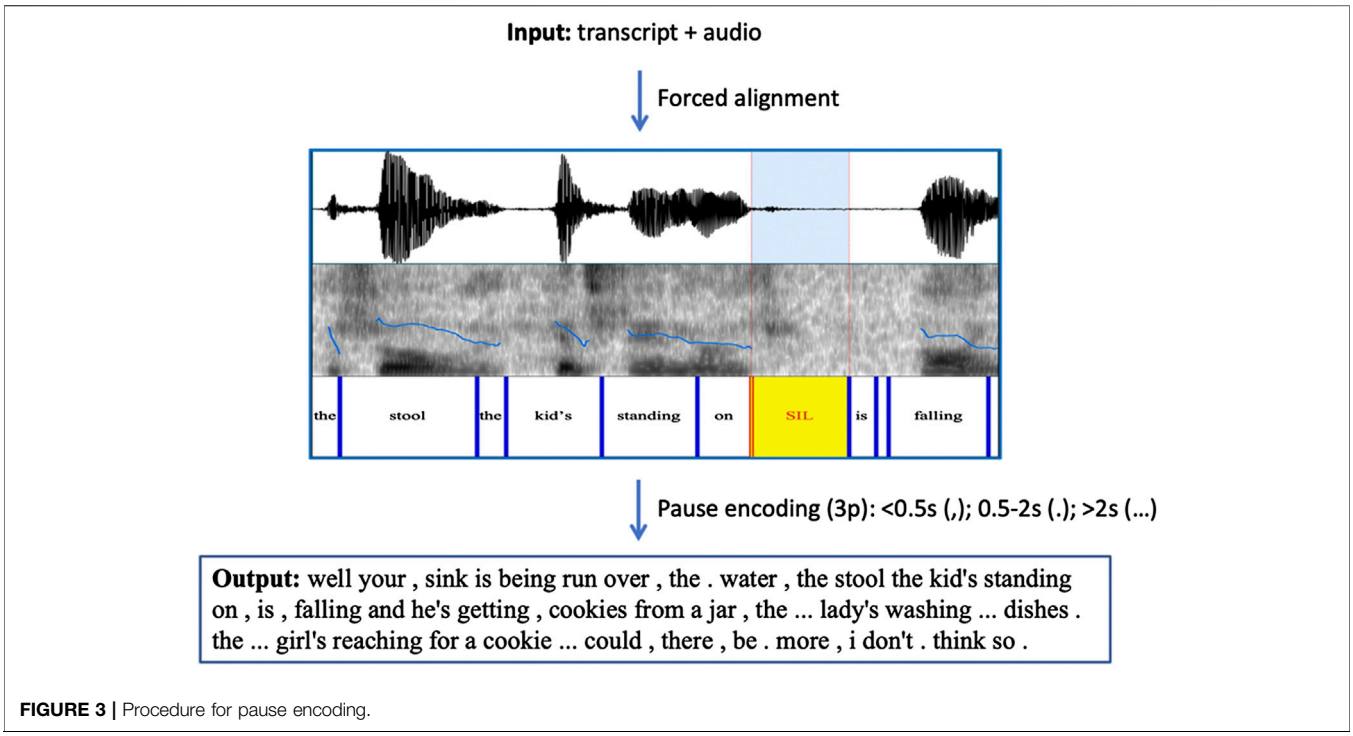

FIGURE 3 | Procedure for pause encoding.

model consists 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 one for AD.

We also encoded pauses in the input word sequence. We grouped pauses into three bins: short (under 0.5 s); medium (0.5–2 s); and long (over 2 s). 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. The procedure is illustrated in **Figure 3**.

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.

The fine-tuning process is illustrated in **Figure 4**.

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

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. We computed 50 runs of LOO cross-validation. The hyperparameter setting was the same across runs except for random seeds. The results are shown in the last row ($N = 1$) of **Tables 2** and **3**. Over the 50 runs, LOO accuracy ranged from 0.75 to 0.86 for BERT with three pauses, from 0.78 to 0.87 for ERNIE with three pauses, and from 0.77 to 0.85 for ERNIE with no Pauses. 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).

To address this brittleness, we introduced the following ensemble procedure. From the results of LOO cross-validation, we calculated the majority vote over N runs for each of the 108 subjects, and used the majority vote to return a single label for each subject. 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 of LOO cross-validation. The results are shown in **Table 2** and **3**. In the tables, the first row summarizes 100 draws of $N = 5$ runs. The second row is similar, except $N = 15$. All of the ensemble rows have better means and less variance than the last row, which summarizes the 50 individual runs of LOO cross-validation without ensemble ($N = 1$). **Figure 5** illustrates **Table 2** and **3**. In **Figure 5** the black lines represent accuracy of individual runs whereas the purple lines represent ensemble accuracy of $N = 35$. We can see that there is a wide variance in individual runs (black).

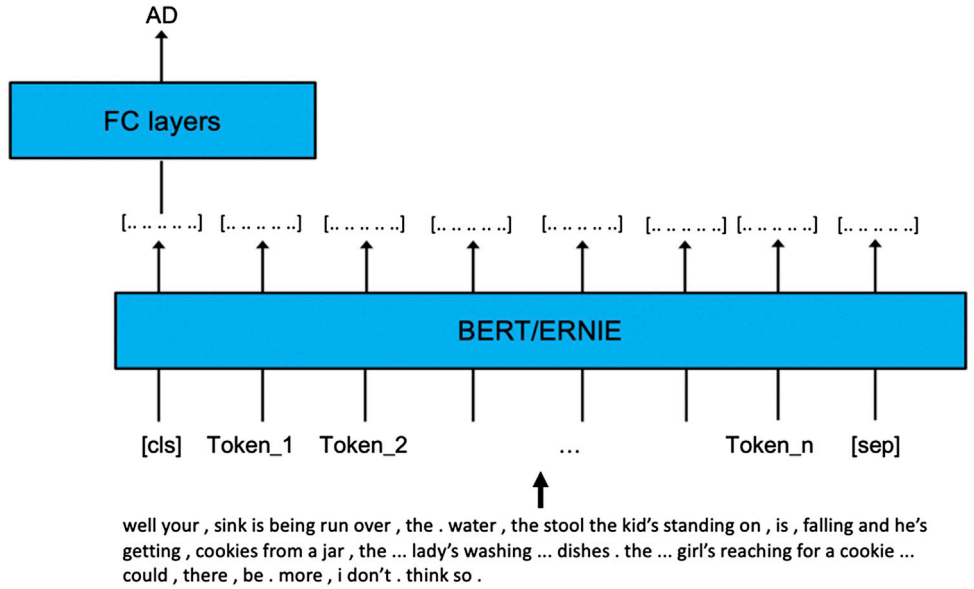

FIGURE 4 | Procedure for fine-tuning.

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
1	**0.819 ± 0.023**	**0.750–0.861**

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
1	**0.827 ± 0.020**	**0.778–0.870**	**0.816 ± 0.023**	**0.769–0.852**

The proposed ensemble method (purple) improves the mean and reduces variance over estimates based on a single run.

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. 0p indicates that no pause was encoded, and 3p and 6p indicate, respectively, that three and six lengths of pauses were encoded. To compare with three pauses, 6p represents six bins of pauses, encoded as: "," (under 0.5 s), "." (0.5–1 s); ".." (1–2 s), ". . ." (2–3 s), ". . . ." (3–4 s), ". . . . ." (over than 4 s). The 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% (Luz et al., 2020).

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 as cognitive load, arousal, and working memory (Daneman, 1991; Arciuli et al., 2010). 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

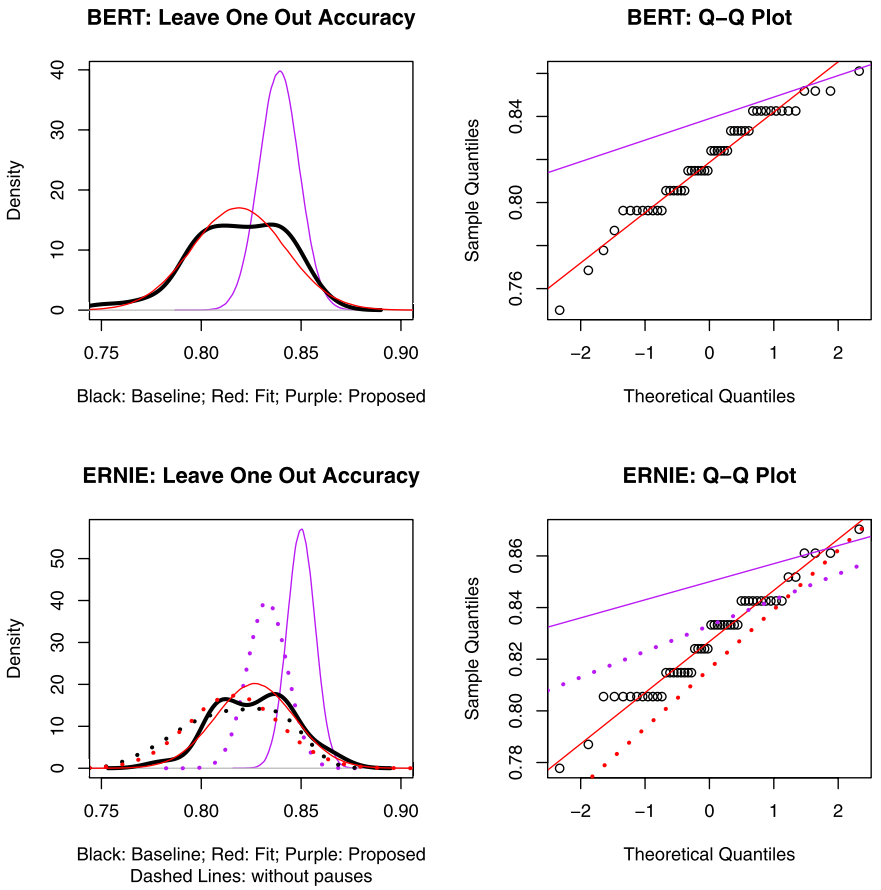

FIGURE 5 | Individual and ensemble Leave-one-out (LOO) accuracy for BERT with pauses (top) and ERNIE with and without pauses (bottom). Black lines represent accuracy of individual runs; purple lines represent ensemble accuracy of $N = 35$.

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 ()	0.700	0.830	0.870	0.620	0.780	0.710	0.750
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**

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 (Gorman et al., 2016; Irvine et al., 2016), and that *um* was used less frequently during lying compared to truth-telling (Benus et al., 2006; Arciuli et al., 2010). 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.

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. There was no difference between three pauses and six pauses in terms of improvement in accuracy. More studies are needed to investigate the categories of pause length and determine the optimal number of pauses to be encoded for AD classification.

ERNIE was designed to learn language representation enhanced by knowledge masking strategies, including entity-level masking and phrase-level masking. Through these

strategies, ERNIE "implicitly learned the information about knowledge and longer semantic dependency, such as the relationship between entities, the property of a entity and the type of a event". (Sun et al., 2019) We think this may be why ERNIE performs better on recognition of Alzheimer's speech, in which memory loss causes not only language problems but also difficulties of recognizing entities and events.

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 CONCLUSION

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.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JY, Principal investigator, corresponding author; XC, Running ERNIE experiments; YB, Help running BERT experiments; ZY, Consultation on Alzheimer's disease, paper editing and proofreading; KC, Visualization of LOO experiment results, paper editing and proofreading.

ACKNOWLEDGMENTS

We thank Julia Li and Hao Tian for their suggestion and help with ERNIE. This paper is an extended version of our paper presented at Interspeech 2020 (Yuan et al., 2020).

REFERENCES

Arciuli, J., Mallard, D., and Villar, G. (2010). "Um, i can tell you're lying": linguistic markers of deception versus truth-telling in speech. *Appl. Psycholinguist.* 31, 397–411. 10.1017/S0142716410000044

Benus, S., Enos, F., Hirschberg, J., and Shriberg, E. (2006). "Pauses in deceptive speech," in Speech prosody 2006, Dresden, Germany, May 2–5, 2006.

Brown, E., and Miron, M. (1971). Lexical and syntactic predictors of the distribution of pause time in reading. *J. Verb. Learn. Verb. Behav.* 10, 658–667. doi:10.1016/S0022-5371(71)80072-5

Butcher, A. (1981). *Aspects of the speech pause: phonetic correlates and communicative functions*. Kiel, Germany: Institut fur Phonetik der Universitat Kiel.

Clark, H. H., and Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition* 84, 73–111. doi:10.1016/s0010-0277(02)00017-3

Clark, H. H. (2006). *Pauses and hesitations: psycholinguistic approach*. Encyclopedia of Language & Linguistics, 244–248.

Corley, M., and Stewart, O. (2008). Hesitation disfluencies in spontaneous speech: the meaning of um. *Language and Linguistics Compass* 2, 589–602. 10.1111/j.1749-818X.2008.00068.x

Daneman, M. (1991). Working memory as a predictor of verbal fluency. *J. Psycholinguist. Res.* 20, 445–464. 10.1007/BF01067637

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint. Available at: https://arxiv.org/abs/1810.04805 (Accessed October 11, 2018).

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., and Smith, N. (2020). Fine-tuning pretrained language models: weight initializations, data orders, and early stopping. arXiv preprint. Available at: https://arxiv.org/abs/2002.06305 (Accessed February 15, 2020).

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *J. Mem. Lang.* 30, 210–233.

Filiou, R.-P., Bier, N., Slegers, A., Houzé, B., Belchior, P., and Brambati, S. M. (2019). Connected speech assessment in the early detection of alzheimer's disease and mild cognitive impairment: a scoping review. *Aphasiology.* 34, 1–33. 10.1080/02687038.2019.1608502

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify alzheimer's disease in narrative speech. *J Alzheimers Dis.* 49 (2), 407–422. doi:10.3233/JAD-150520

Fritsch, J., Wankerl, S., and Nöth, E. (2019). "Automatic diagnosis of alzheimer's disease using neural network language models," in ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing, Brighton, United Kingdom, May 12, 2020 (ICASSP IEEE), 5841–5845.

Goldman-Eisler, F. (1961). The distribution of pause durations in speech. *Lang. Speech* 4, 232–237. doi:10.1177/002383096100400405

Goodglass, H., Kaplan, E., and Barresi, B. (2001). *Boston diagnostic Aphasia examination*. 3rd Edition. Philadelphia: Lippincott Williams & Wilkins.

Gorman, K., Olson, L., Hill, A., Lunsford, R., Heeman, P., and van Santen, J. (2016). Uh and um in children with autism spectrum disorders or language impairment. *Autism Res.* 9, 854–865. doi:10.1002/aur.1578

Gosztolya, G., Vincze, V., Toth, L., Pakaski, M., Kalman, J., and Hoffmann, I. (2019). Identifying mild cognitive impairment and mild alzheimer's diseasebased on spontaneous speech using asr and linguistic features. *Comput. Speech Lang.* 53, 181–197. 10.1016/j.csl.2018.07.007

Grosjean, F., Grosjean, L., and Lane, H. (1971). The patterns of silence: performance structures in sentence production. *Cognit. Psychol.* 11, 58–81. doi:10.1016/0010-0285(79)90004-5

Hawthorne, K., and Gerken, L. (2014). From pauses to clauses: prosody facilitates learning of syntactic constituency. *Cognition* 133, 420–428. doi:10.1016/j.cognition.2014.07.013

Irvine, C. A., Eigsti, I. M., and Fein, D. (2016). Uh, um, and autism: filler disfluencies as pragmatic markers in adolescents with optimal outcomes from autism spectrum disorder. *J. Autism Dev. Disord.* 46, 1061–1070. doi:10.1007/s10803-015-2651-y

Krivokapic, J. (2007). Prosodic planning: effects of phrasal length and complexity on pause duration. *J. Phonetics* 35, 162–179. doi:10.1016/j.wocn.2006.04.001

Laske, C., Sohrabi, H. R., Frost, S., López-de-Ipiña, K., Garrard, P., Buscema, M., et al. 2015). Innovative diagnostic tools for early detection of Alzheimer's disease. *Alzheimers Dement* 11, 561–578. doi:10.1016/j.jalz.2014.06.004

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). Alzheimer's dementia recognition through spontaneous speech: the ADReSS Challenge," in Proceedings of INTERSPEECH 2020, Shanghai, China, October 25–29, 2020.

MacWhinney, B. (2000). *The CHILDES project: tools for analyzing talk*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Mattson, M. P. (2004). Pathways towards and away from Alzheimer's disease. *Nature* 430, 631–639. doi:10.1038/nature02621

Mueller, K. D., Koscik, R. L., Hermann, B., Johnson, S. C., and Turkstra, L. S. (2017). Declines in connected language are associated with very early mild cognitive impairment: results from the Wisconsin registry for alzheimer's prevention. *Front. Aging Neurosci.* 9, 437. doi:10.3389/fnagi.2017.00437

Palo, F. D., and Parde, N. (2019). "Enriching neural models with targeted features for dementia detection," in Proceedings of the 57th annual Meeting of the Association for computational linguistics (ACL), Florence, Italy, July 2019.

Pulido, M. L. B., Hernández, J. B. A., Ballester, M. A. F., González, C., Mekyska, J., and Smékal, Z. (2020). Alzheimer's disease and automatic speech analysis: a review. *Expert Syst. Appl.* 150, 113213. doi:10.1016/j.eswa.2020.113213

Ramanarayanan, V., Goldstein, L., Byrd, D., and Narayanan, S. (2013). An investigation of articulatory setting using real-time magnetic resonance imaging. *J. Acoust. Soc. Am.* 134, 510–519. doi:10.1121/1.4807639

Ramig, L., Countryman, S., Thompson, L., and Horii, Y. (1995). Comparison of two forms of intensive speech treatment for Parkinson disease. *J. Speech Hear. Res.* 38, 1232–1251. doi:10.1044/jshr.3806.1232

Rochester, S. R. (1973). The significance of pauses in spontaneous speech. *J. Psycholinguist. Res.* 2, 51–81. doi:10.1007/BF01067111

Schepman, A., and Rodway, P. (2000). Prosody and parsing in coordination structures. *Q. J. Exp. Psychol.* 53, 377–396. doi:10.1080/713755895

Shea, C., and Leonard, K. (2019). Evaluating measures of pausing for second language fluency research. *Can. Mod. Lang. Rev.* 75, 1–20. 10.3138/cmlr.2018-0258

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., et al. (2019). Ernie 2.0: a continual pre-training framework for language understanding. arXiv preprint. Available at: https://arxiv.org/abs/1907.12412 (Accessed July 29, 2019).

Tottie, G. (2011). Uh and um as sociolinguistic markers in british English. *Int. J. Corpus Linguist.* 16, 173–197. 10.1075/ijcl.16.2.02tot

Tran, T., Toshniwal, S., Bansal, M., Gimpel, K., Livescu, K., and Ostendorf, M. (2018). Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information. arXiv preprint. Available at: https://arxiv.org/abs/1704.07287 (Accessed April 24, 2017).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in neural information processing systems*, 5998–6008.

Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., and Liberman, M. (2016). Variation and change in the use of hesitation markers in germanic languages. *Lang. Dynam. Change* 6, 199–234. 10.1163/22105832-00602001

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., and Church, K. (2020). "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease," in Proceedings of INTERSPEECH 2020, Shanghai, China, October 25–29, 2020.

Yuan, J., and Liberman, M. (2008). Speaker identification on the scotus corpus. *J. Acoust. Soc. Am.* 123, 3878. doi:10.1121/1.2935783

Yuan, J., Xu, X., Lai, W., and Liberman, M. (2016). Pauses and pause fillers in Mandarin monologue speech: the effects of sex and proficiency. *Proc. Speech Prosody* 2016, 1167–1170. doi:10.21437/SpeechProsody.2016-240

Zellner, B. (1994). "Pauses and the temporal structure of speech," in *Fundamentals of speech synthesis and speech recognition*. Editor E. Keller (Chichester: John Wiley), 41–62.

Conflict of Interest: Authors JY, XC, YB and KC were employed by company Baidu USA Inc.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.