Optimizing Pause Context in Fine-Tuning Pre-trained Large Language Models for Dementia Detection

Xiaoquan Ke[1], *Man-Wai Mak*[2], *Helen M. Meng*[1]

[1]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong SAR
[2]Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

xiaoquanke@cuhk.edu.hk, enmwmak@polyu.edu.hk, hmmeng@se.cuhk.edu.hk

Abstract

Speech pauses serve as a valuable and non-invasive biomarker for the early detection of dementia. Our study aims to examine abnormal pauses, specifically their durations, for improving the detection performance. Inspired by the proven performance of the Transformer-based models in dementia detection, we opted for integrating the abnormal pauses into these models. Specifically, we enriched the inputs for the Transformer-based models by fusing between-segment pause context into the automated transcriptions. We performed the experiments on our Cantonese elderly corpus called CU-Marvel. To improve the detection performance, we optimized the pause durations when infusing the pause context into the transcriptions. Our findings suggest that the between-segment pauses could also serve as promising biomarkers. We emphasize the importance of optimizing pause patterns across different languages or datasets. Our findings indicate that various classification tasks prefer distinct patterns of pause infusing.

Index Terms: Dementia detection, text-based embeddings, disfluecnies, pauses

1. Introduction

Dementia is a significant cognitive impairment that can greatly impact the health and daily functioning of those affected. The most common form of dementia is Alzheimer's disease (AD). The disease has a huge impact on the quality of life, not only for individuals with dementia but also for their families and caretakers. Fortunately, with effective detection of early dementia, disease-modifying medications and interventions are possible. Recent advancements in antibody therapy showed promise in slowing the progression of early-stage AD [1, 2]. More recently, the U.S. Food and Drug Administration (FDA) has also approved a new generation of amyloid beta (Aβ) monoclonal antibody, the donanemab, for the treatment of early-stage AD.[1] This shows great promise for managing and potentially slowing down the progression of dementia if early-stage of dementia is detected.

Currently, assessments of dementia can be classified into three categories: (1) genetic analysis, which focuses on identifying genotypes like apolipoprotein E (APoE) [3]; (2) detection of serological biomarkers, such as brain-derived neurotrophic factors [4], Aβ plaques, and neurofibrillary tangles of tau protein [5]; (3) utilization of brain imaging techniques such as magnetic resonance imaging (MRI) [6]. However, most of these assessments are invasive and often not easily accessible to detect early-stage of AD in clinical practices. Some findings suggest that individuals with dementia display language deficits in the pre-clinical stages of the disease, indicating that such deficits may manifest even before the clinical diagnosis is made [7, 8]. Consequently, early detection of dementia can be achieved through speech and language analyses. It can also offer easily accessible biomarkers through neuropsychological assessments for disease prediction and monitoring.

This study focuses on automated assessments of dementia through speech and language analysis. To achieve this, we examine a speech biomarker that is independent of the content of spoken language, specifically the *speech pauses*, for the detection of dementia. Although pauses are common in speech, there is a distinction between normal and abnormal pauses. Our study aims to examine abnormal pauses, specifically their durations, for the detection of dementia. Inspired by the proven performance of Transformer-based models in dementia detection, we opted for integrating the abnormal pauses into these models. Specifically, we enriched the inputs for the Transformer-based models by fusing pause context into the automated transcriptions. We performed the experiments on our Cantonese elderly corpus called CU-Marvel. To improve the detection performance, we optimized the pause durations when infusing the pause context into the transcriptions.

2. Related Work

Recently, Transformer-based models have become increasingly prevalent in dementia detection, demonstrating superior detection performance. In [9], the BERT [10] and ERNIE [11] models were fine-tuned to capture the language characteristics of the speakers in ADReSSo 2021 challenge [12]. Li *et al.* [13] extracted BERT features from both manual and automatic transcriptions. Their results demonstrate the effectiveness of the BERT features for dementia detection.

There have been several studied investigating the speech pauses in dementia detection. Braun *et al.* [14] examined different groups of speech pauses for dementia detection. They also investigated the effect of incorporating pause information from the acoustic context into the text-based assessment using cross-attention. Their findings suggested that the selection of the test should be tailored based on the specific task requirements. Specifically, the verbal fluency test (VFT) was most effective in distinguishing individuals with MCI from HCs when the text-based assessment could learn from acoustic information.

Syed *et al.* [15] compared the efficacy of BERT and its derivatives, including DistilBERT [16] and RoBERTa [17], for capturing the structural and linguistic properties of the transcriptions. They also introduced a special pre-processing step

[1] https://www.fda.gov/drugs/news-events-human-drugs/fda-approves-treatment-adults-alzheimers-disease

that integrates silence durations into the transcriptions. Specifically, when the duration was between 2s and 4s, they added `<uhm>` to the transcriptions. If the silence was between 4s and 6s, they added `<uhm uhm>`. If the silence exceeded 6s, they added `<long silence>`.

Yuan *et al.* [18] applied a special pre-processing step that encodes pauses in the transcriptions for AD detection. More precisely, the pauses were divided into three groups according to their durations: G_1 (pauses between 0.05s and 0.5s), G_2 (pauses between 0.5s and 2s), and G_3 (pauses longer than 2s). Three groups of pauses were encoded using three punctuations <.>, <..>, and <...>, respectively. Finally, the Transformer-based models, BERT [10] and ERNIE [11], were fine-tuned using the pre-processed transcriptions as input.

Dominguez *et al.* [19] investigated speech pauses among three groups of participants, including – healthy older adults (HCs), individuals with mild cognitive impairment (MCI), and AD patients – based on a picture-story description task. They used maximum likelihood estimation (MLE) to fit the pause distributions and divided them into three truncated bins: G_1 (pauses between 0.2s and 0.6s), G_2 (pauses between 0.6s and 1.5s), and G_3 (pauses longer than 1.5s).

The authors in [20] proposed a pause distribution derived from the data-driven analysis of the INTERVIEW dataset [21]. They empirically categorized pauses into six groups based on the distribution: 0.05–0.1s, 0.1–0.3s, 0.3–0.6s, 0.6–1.0s, 1.0–2.0s, and >2.0s. They enriched the transcriptions by representing the respective groupings using two to seven dots, as follows: <..>, <...>, <....>, <.....>, <......>, <.......>. Subsequently, they re-trained BERT and RoBERTa [17] using the pause-enriched transcriptions for the following AD recognition and emotion recognition tasks.

3. Method

This section begins with describing the Cantonese corpus collected for the screening and monitoring of neurocognitive disorders (NCD). Subsequently, we explain the process of fine-tuning the Transformer-based model for the detection of dementia. We then describe the infusing of pause context into the transcriptions, followed by an explanation of how the pause durations were optimized.

3.1. CU-Marvel Cantonese Corpus

Cantonese is one of the major Chinese dialects that has over 80 million native speakers in Southern China. We collected the CU-Marvel corpus for the research on the screening and monitoring of NCD based on spoken language technologies. A series of cognitive tests, including – Boston naming tests, logical memory and recognition tests, Montreal Cognitive Assessment (MoCA) tests, and picture description tests – were given to each participant for assessing the mild cognitive impairment (MCI) and dementia in older adults. According to the assessment results, 553 participants were divided into three groups: 349 healthy older adults (HCs), 167 older adults having minor NCD, and 37 older adults suffering from major NCD.

The corpus was split into a training set and a test set to maintain balance across categories and ages within these two sets, as shown in Table 1. A three-minute rabbit story picture description test was chosen as the narrative speech task for the experiments, as shown in Fig. 1.

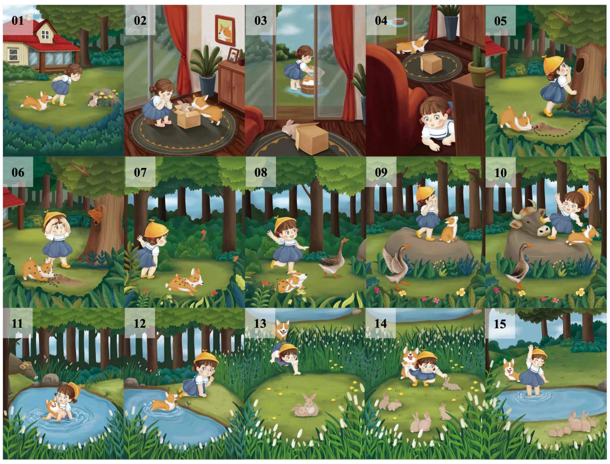


Figure 1: *The cartoon pictures used in the rabbit-story picture description test.*

3.2. Transformer-Based Models Fine-Tuning

As no manual transcriptions were provided, we utilized automatic speech recognition (ASR) techniques to transcribe the speech recordings into automated transcriptions. Specifically, we utilized an wav2vec 2.0 ASR system that was tailored to enhance the accuracy of transcribing elderly Cantonese speech in Hong Kong [22]. The system achieved a character error rate (CER) of 16.27%.

To encode the automatic transcriptions, we selected Bidirectional Encoder Representations from Transformers (BERT) [10] as the pre-trained masked language model for natural language processing. We employed a specific BERT variant, known as bert-base-cantonese[2], which is a pre-trained version of bert-base-chinese[3] on the Cantonese Common Crawl dataset. The transcriptions were input into the BERT model, and the [CLS] embeddings of the last hidden state were fed into the final classification layers. During the training, the whole model was fine-tuned to classify the participants into HCs, minor NCD, or major NCD.

3.3. Pause-Enriched Transcriptions

We analyze the *between-segment* pauses detected by the ASR system. We exclude between-character pauses as their identification requires precise alignment between the transcriptions and speech recordings [18]. However, obtaining accurate character-level alignment is challenging due to errors in the automatic transcriptions. The distributions of between-segment pauses from the CU-Marvel training data are illustrated in Fig. 2. It shows that minor NCD and major NCD tend to have more between-segment pauses than HCs, particularly the long pauses (>2.5s). This indicates that minor NCD and major NCD produce a higher number of short speech segments, with longer pauses between the segments. In contrast, the HCs tend to produce fewer but longer speech segments. It also shows that minor NCD and major NCD exhibit a higher median pause duration compared to HCs.

To infuse speech pauses into the transcriptions, we first categorized and assigned them to different groups (in seconds):

1) *P1* (1 group of pauses): $[0.2, +\infty)$ [14].

[2] https://huggingface.co/indiejoseph/bert-base-cantonese

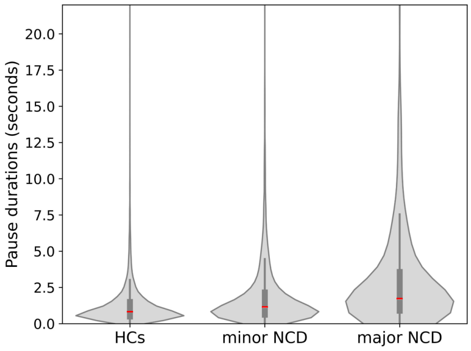
[3] https://huggingface.co/google-bert/bert-base-chinese

Table 1: *Characteristics of the CU-Marvel dataset.*

	Training set			Test set		
	HCs	Minor NCD	Major NCD	HCs	Minor NCD	Major NCD
No. of subjects	275	133	27	74	34	10
Age (years)	70 [65, 74]	75 [69, 81]	81 [78, 85]	70 [66, 74.8]	70 [67, 79.8]	84 [82.2, 86.5]
Educations (years)	9 [6, 11]	7 [4, 11]	6 [2, 9]	8.5 [6, 11]	6 [6, 9]	4 [1.25, 6.75]
MoCA scores	24 [21, 27]	20 [17, 22]	15 [9.5, 17.5]	24 [21, 27]	20 [18, 23]	13.5 [12, 16.8]
Language	Cantonese					
Task	Rabbit-story picture description					
Manual transcriptions	No					

HCs: healthy older adults; NCD: neurocognitive disorders.
The values are presented as *median [interquartile range]*.

Figure 2: *The distributions of between-segment pauses from various categories of the CU-Marvel training data. In each box, the red line represents the median pause duration of the category.*



2) *P2* (3 groups of pauses)
   - $[0.05, 0.5)$, $[0.5, 2.0]$, $(2, +\infty)$ [18]
   - $[0.2, 0.6]$, $(0.6, 1.5)$, $[1.5, +\infty)$ [19]
3) *P3* (6 groups of pauses): $[0.05, 0.1]$, $(0.1, 0.3]$, $(0.3, 0.6]$, $(0.6, 1.0]$, $(1.0, 2.0]$, $(2, +\infty)$ [20].

The P1 grouping is derived from [14], in which pauses lasting 0.2s or longer were encoded as a special token in transcriptions for detecting dementia in a German dementia dataset. The P2 grouping ($[0.05, 0.5)$, $[0.5, 2.0]$, $(2, +\infty)$) [18] was empirically determined from the data analysis of the ADReSS English dataset. They divided pauses as short (between 0.05s and 0.5s), median (between 0.5s and 2s), and long (longer than 2s) pauses and encoded them using three special tokens. The P2 grouping ($[0.2, 0.6]$, $(0.6, 1.5)$, $[1.5, +\infty)$) [19] also divided the pauses into three groups, where the group boundaries were established by using MLE to fit the pause distributions across three participant categories. The P3 grouping [20] used six special tokens to separately encode each group of pauses. The pause grouping was empirically established through a data-driven analysis of the INTERVIEW dataset [21].

During training, the enriched transcriptions were input into the BERT model for classification. The extra special tokens were added to the BERT tokenizer and the model embeddings were adjusted to the new vocabulary length.

3.4. Optimizing Pause Context

For training the BERT models, an Adam optimizer with a learning rate of 0.001 was used to optimize the models' parameters. The batch size was set to 64 and the maximum training epochs was set to 50. To address the class imbalance in the training set, we opted for using the Focal Loss [23] as the loss function with the aim of mitigating the effects of the imbalance. During the training process, early stopping was employed to mitigate the risk of overfitting on the limited training data.

In addition to validating the performance of baseline P1–P3 grouping in our Cantonese dataset, we conducted 5-fold cross-validation (CV) and grid search to optimize the pause groupings. More specifically, the model was trained and evaluated for each of the candidate pause groupings in the search space using the 5-fold CV. The optimal pause grouping was determined if the pause grouping obtained the best mean detection performance during the CV. We optimize the pause groupings using the following search space:

1) P1 (1 group of pauses): $[0.2, +\infty)$; $[0.25, +\infty)$; $[0.5, +\infty)$; $[1.0, +\infty)$; $[1.5, +\infty)$; $[2.0, +\infty)$; $[2.5, +\infty)$; $[3.0, +\infty)$; $[3.5, +\infty)$; $[4.0, +\infty)$.
2) P2 (3 group of pauses):
   - $[0.05, 0.5)$, $[0.5, 2.0]$, $(2, +\infty)$
   - $[0.2, 0.6]$, $(0.6, 1.5)$, $[1.5, +\infty)$
   - $[0.25, 0.5)$, $[0.5, 1.0]$,$(1.0, +\infty)$
   - $[0.25, 0.5)$, $[0.5, 2.0]$,$(2.0, +\infty)$
   - $[0.5, 1.0)$, $[1.0, 2.0]$, $(2.0, +\infty)$
   - $[0.5, 2.0)$, $[2.0, 4.0]$, $(4.0, +\infty)$
   - $[1.0, 2.0)$, $[2.0, 4.0]$, $(4.0, +\infty)$

The search space of P1 was defined to identify abnormal pause durations in speech, and the search space of P2 was established to explore more suitable groupings for our Cantonese dataset.

After optimizing the pause groupings, we applied the optimal pause grouping to train the BERT model and report the results on the CU-Marvel test dataset. The performance metrics include accuracy (ACC) and F_1 scores.

4. Results

Table 2 presents the binary classification results comparing HCs and minor NCD, with the best result obtained from our P1 ($[2.5, +\infty)$), surpassing the BERT models. Additionally, baseline P1 [14] also performs well on the CU-Marvel test data. Comparing baseline P1 [14] and our P1 ($[2.5, +\infty)$) reveals that in the binary classification between HCs and minor NCD, better performance is achieved by encoding only long pauses (>2.5s). If we only encoding the long pauses (>2.5s), the transcriptions of minor NCD exhibit more pause tokens compared to the HCs. This is due to minor NCD have more long pauses than the HCs, as illustrated in Fig. 2. This higher number of pause tokens

Table 2: *The results on the CU-Marvel test data (**HCs vs. minor NCD**).*

Method	Pause grouping	Pause durations	ACC	F1 (mean)
BERT	-	-	0.672	0.597
Baseline	P1 [14]	$[0.2, +\infty)$	0.676	0.637
	P2 [18]	$[0.05, 0.5)$, $[0.5, 2.0]$, $(2, +\infty)$	0.670	0.602
	P2 [19]	$[0.2, 0.6]$, $(0.6, 1.5)$, $[1.5, +\infty)$	0.665	0.595
	P3 [20]	$[0.05, 0.1], \cdots, (2, +\infty)$	0.680	0.602
Ours	P1	$[2.5, +\infty)$	0.698	***0.659***
	P2	$[0.05, 0.5)$, $[0.5, 2.0]$, $(2, +\infty)$	0.670	0.602

Table 3: *The results on the CU-Marvel test data (**HCs vs. major NCD**).*

Method	Pause grouping	Pause durations	ACC	F_1 (mean)
BERT	-	-	0.885	0.790
Baseline	P1 [14]	$[0.2, +\infty)$	0.881	0.755
	P2 [18]	$[0.05, 0.5)$, $[0.5, 2.0]$, $(2, +\infty)$	0.857	0.759
	P2 [19]	$[0.2, 0.6]$, $(0.6, 1.5)$, $[1.5, +\infty)$	0.902	0.809
	P3 [20]	$[0.05, 0.1], \cdots, (2, +\infty)$	0.883	0.789
Ours	P1	$[2.0, +\infty)$	0.904	0.812
	P2	$[1.0, 2.0)$, $[2.0, 4.0]$, $(4.0, +\infty)$	0.924	***0.832***

renders the transcriptions more fragmented, which is different from the HCs, and thus achieving better results. Baseline P2 [18] and P3 [20] exhibit only slightly better performance than the BERT models. Additionally, the performance of baseline P2 [19] exhibits slightly lower performance compared to the BERT models. The findings suggest that enriching transcriptions with pause context may not necessarily improve detection performance. On one hand, baseline P2 [19] utilized between-word pauses, whereas in our dataset, only between-segment pauses were used, making it hard to apply pause grouping directly. On other hand, while baseline P2 [19] determined pause grouping by fitting pause distributions using MLE, their pause distributions might differ from those in our dataset.

The binary classification results between HCs and major NCD are presented in Table 3. It is observed that both our P1 and P2 groupings outperform the BERT models, with our P2 group showing superior performance. This indicates that this binary classification task benefits more from encoding three distinct groups of pauses compared to encoding just one group of pauses. P1 [14], P2 [18], and P3 [20] perform worse than the BERT models. This again suggests the pause groupings from other datasets or languages cannot be directly applied to our Cantonese dataset.

The binary classification results comparing minor NCD and major NCD are shown in Table 4, revealing that the pause grouping adopted from P1 [14] obtained the best performance. P1 grouping [14] encoded nearly all pauses because most of the pauses are longer than 0.2s in our dataset, as shown in Fig. 2. This result suggests that the binary classification between minor NCD and major NCD benefits from encoding almost all the pauses, focusing on positional rather than temporal pause information.

Table 5 shows the three-class classification results among HCs, minor NCD, and major NCD on the CU-Marvel test data. In this classification, our P1 ($[1.5, +\infty)$) achieves the best performance, surpassing the BERT models and the baseline groupings. Some of the baseline grouping perform worse than the BERT models, including P1 [14] and P3 [20].

Table 4: *The results on the CU-Marvel test data (**minor NCD vs. major NCD**).*

Method	Pause grouping	Pause durations	ACC	F_1 (mean)
BERT	-	-	0.718	0.622
Baseline	P1 [14]	$[0.2, +\infty)$	0.795	***0.706***
	P2 [18]	$[0.05, 0.5)$, $[0.5, 2.0]$, $(2, +\infty)$	0.800	0.675
	P2 [19]	$[0.2, 0.6]$, $(0.6, 1.5)$, $[1.5, +\infty)$	0.818	0.668
	P3 [20]	$[0.05, 0.1], \cdots, (2, +\infty)$	0.818	0.700
Ours	P1	$[0.2, +\infty)$	0.795	***0.706***
	P2	$[0.05, 0.5)$, $[0.5, 2.0]$, $(2, +\infty)$	0.800	0.675

Table 5: *The results on the CU-Marvel test data (**HCs vs. minor NCD vs. major NCD**).*

Method	Pause grouping	Pause durations	ACC	F_1 (mean)
BERT	-	-	0.627	0.527
Baseline	P1 [14]	$[0.2, +\infty)$	0.602	0.514
	P2 [18]	$[0.05, 0.5)$, $[0.5, 2.0]$, $(2, +\infty)$	0.614	0.530
	P2 [19]	$[0.2, 0.6]$, $(0.6, 1.5)$, $[1.5, +\infty)$	0.617	0.528
	P3 [20]	$[0.05, 0.1], \cdots, (2, +\infty)$	0.605	0.522
Ours	P1	$[1.5, +\infty)$	0.629	***0.558***
	P2	$[0.05, 0.5)$, $[0.5, 2.0]$, $(2, +\infty)$	0.614	0.530

5. Discussions and Conclusions

In this study, we examine the effectiveness of speech pauses in combination with Transformer-based models for dementia detection using our Cantonese corpus. Although we utilized between-segment pauses, differentiating from previous studies such as [14, 18, 19, 20], which incorporate both between-segment and between-word pauses, our research indicates that between-segment pauses can still serve as a valuable indicator of dementia. Our results demonstrate that encoding between-segment speech pauses yielded the best performance (Table 2, Table 3, and Table 5).

Several studies [14, 18, 19, 20] have indicated that speech pauses serve as valuable indicators of dementia. However, the distinctive abnormal pause patterns observed in other languages cannot be directly applied to our Cantonese dataset. In the binary classification task distinguishing between HCs and major NCD (refer to Table 3), incorporating the baseline P1 and P2 significantly impairs the classification performance compared to the BERT models that without infusing pause context. However, upon optimizing the durations of P1 and P2, the performance surpasses to the BERT models. This emphasizes the importance of optimizing pause patterns to suit the characteristics of the target dataset.

Different classification tasks prefer different pause infusing patterns. In the binary classification between HCs and minor NCD (see Table 2) and the three-class classification scenario (see Table 5), the best performance was achieved by encoding only encoding one group of long pauses. Conversely, in the binary classification of HCs and major NCD (see Table 3), the model showed improved performance when encoding three distinct groups of pauses. In the classification between minor NCD and major NCD (see Table 4), the model benefits from encoding almost all the pauses, focusing on positional rather than temporal pause information.

Overall, the use of between-segment pauses in our Cantonese corpus demonstrates promising effectiveness, providing insights into leveraging such pauses to improve dementia detection performance. Subsequent research could explore more better pause patterns to enhance performance.

6. References

[1] C. H. Van Dyck, C. J. Swanson, P. Aisen, R. J. Bateman, C. Chen, M. Gee, M. Kanekiyo, D. Li, L. Reyderman, S. Cohen *et al.*, "Lecanemab in early alzheimer's disease," *New England Journal of Medicine*, vol. 388, no. 1, pp. 9–21, 2023.

[2] J. R. Sims, J. A. Zimmer, C. D. Evans, M. Lu, P. Ardayfio, J. Sparks, A. M. Wessels, S. Shcherbinin, H. Wang, E. S. M. Nery *et al.*, "Donanemab in early symptomatic alzheimer disease: the trailblazer-alz 2 randomized clinical trial," *Jama*, vol. 330, no. 6, pp. 512–527, 2023.

[3] C.-C. Liu, T. Kanekiyo, H. Xu, and G. Bu, "Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy," *Nat. Rev. Neurol.*, vol. 9, no. 2, pp. 106–118, Jan. 2013.

[4] J.-H. Song, J.-T. Yu, and L. Tan, "Brain-derived neurotrophic factor in Alzheimer's disease: Risk, mechanisms, and therapy," *Mol. Neurobiol.*, vol. 52, no. 3, pp. 1477–1493, Oct. 2014.

[5] S. Amini, B. Hao, L. Zhang, M. Song, A. Gupta, C. Karjadi, V. B. Kolachalama, R. Au, and I. C. Paschalidis, "Automated detection of mild cognitive impairment and dementia from voice recordings: a natural language processing approach," *Alzheimer's & Dementia*, vol. 19, no. 3, pp. 946–955, 2023.

[6] K. A. Johnson, N. C. Fox, R. A. Sperling, and W. E. Klunk, "Brain imaging in Alzheimer disease," *Cold Spring Harb. Perspect. Med.*, vol. 2, no. 4, p. a006213, Jan. 2012.

[7] L. Mickes, J. T. Wixted, C. Fennema-Notestine, D. Galasko, M. W. Bondi, L. J. Thal, and D. P. Salmon, "Progressive impairment on neuropsychological tasks in a longitudinal study of preclinical Alzheimer's disease." *Neuropsychology*, vol. 21, no. 6, pp. 696–705, Nov. 2007.

[8] D. Beltrami, L. Calzà, G. Gagliardi, E. Ghidoni, N. Marcello, R. R. Favretti, and F. Tamburini, "Automatic identification of mild cognitive impairment through the analysis of Italian spontaneous speech productions," in *Proc. Int. Conf. Lang. Resourc. and Eval. (LREC)*, May 2016, pp. 2086–2093.

[9] Y. Qiao, X. Yin, D. Wiechmann, and E. Kerz, "Alzheimer's disease detection from spontaneous speech through combining linguistic complexity and (dis)fluency features with pretrained language models," in *Proc. Interspeech*, Aug. 2021, pp. 3805–3809.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: https://arxiv.org/abs/1810.04805

[11] 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," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, pp. 8968–8975.

[12] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo challenge," in *Proc. Interspeech*, Aug. 2021, pp. 4211–4215.

[13] J. Li, J. Yu, Z. Ye, S. Wong, M. W. Mak, B. Mak, X. Liu, and H. Meng, "A comparative study of acoustic and linguistic features classification for Alzheimer's disease detection," in *Proc. IEEE Int. Conf. Acoust, Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6423–6427.

[14] F. Braun, S. P. Bayerl, F. Hönig, H. Lehfeld, T. Hillemacher, T. Bocklet, and K. Riedhammer, "Infusing acoustic pause context into text-based dementia assessment," in *Proc. Interspeech*. ISCA, Sep. 2024, pp. 1980–1984.

[15] Z. S. Syed, M. S. S. Syed, M. Lech, and E. Pirogova, "Tackling the ADRESSO challenge 2021: The MUET-RMIT system for Alzheimer's dementia recognition from spontaneous speech," in *Proc. Interspeech*, Aug. 2021, pp. 3815–3819.

[16] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," 2018, *arXiv:1910.01108*. [Online]. Available: https://arxiv.org/abs/1910.01108

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*. [Online]. Available: https://arxiv.org/abs/1907.11692

[18] 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 *Proc. Interspeech*, Oct. 2020, pp. 2162–2166.

[19] P. Pastoriza-Dominguez, I. G. Torre, F. Dieguez-Vide, I. Gómez-Ruiz, S. Geladó, J. Bello-López, A. Ávila-Rivera, J. A. Matias-Guiu, V. Pytel, and A. Hernández-Fernández, "Speech pause distribution as an early marker for alzheimer's disease," *Speech Communication*, vol. 136, pp. 107–117, 2022.

[20] J. Yuan, X. Cai, and K. Church, "Pause-encoded language models for recognition of alzheimer's disease and emotion," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2021, pp. 7293–7297.

[21] B. P. Majumder, S. Li, J. Ni, and J. McAuley, "Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 8129–8141.

[22] R. Huang and B. Mak, "Wav2vec 2.0 asr for cantonese-speaking older adults in a clinical setting," in *Proc. Interspeech*. INTERSPEECH, Aug. 2023, pp. 4958–4962.

[23] T.-Y. Ross and G. Dollár, "Focal loss for dense object detection," in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2980–2988.