Aphasia Detection for Cantonese-Speaking and Mandarin-Speaking Patients Using Pre-Trained Language Models

Ying Qin[1], *Tan Lee*[2], *Anthony Pak Hin Kong*[3], *Feng Lin*[4]

[1]Institute of Information Science, Beijing Jiaotong University, Beijing, China
[1]Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China
[2]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China
[3]Unit of Human Communication, Development, and Information Sciences, The University of Hong Kong, Hong Kong, China
[4]School of Rehabilitation, Nanjing Medical University, Nanjing, China

Abstract

Automatic analysis of aphasic speech based on speech technology has been extensively investigated in recent years, but there has been a few studies on Chinese languages. In this paper, we focus on automatic aphasia detection for Cantonese- and Mandarin-speaking patients using state-of-the-art pre-trained language models that support both traditional and simplified Chinese. Given speech transcriptions of subjects, pre-trained language models are used in two ways: 1) pre-trained language model derived embeddings followed by a classifier; 2) pre-trained language model fine-tuned for aphasia detection task. Both approaches are demonstrated to outperform baseline models using acoustic features and static word embeddings. The best accuracy is obtained with fine-tuned BERT models, achieving 0.98 and 0.94 for Cantonese-speaking and Mandarin-speaking subjects respectively. We also investigate the feasibility of applying the cross-lingual pre-trained language model fine-tuned by aphasia detection task for Cantonese-speaking subjects to Mandarin-speaking subjects with limited data. The promising results will hopefully make it possible to perform detection on those low-resource pathological speech which is difficult to implement a specific detection system.

Index Terms: Aphasia detection, Pre-trained language model, Cantonese, Mandarin

1. Introduction

Aphasia is a type of language impairment resulting from dysfunction in specific brain regions. It is caused typically by stroke or other physical conditions, e.g., head trauma or tumor. Aphasia may impair a person's expressive and/or receptive language skills, including auditory comprehension, verbal expression, reading and writing [1], and negatively impact the patient's daily communication and quality of life. Word finding difficulty and speech disfluency are the primary symptoms of aphasia [2]. There are 15 million stroke cases reported worldwide each year [3], and over 2 million among them come from China [4, 5]. Up to 38% of stroke survivors suffer from aphasia [6]. Due to the high prevalence and shortage of speech-language pathologists (SLPs), many patients do not receive timely assessment and regular follow-up. This motivates us to develop automated solutions to alleviate the shortage of SLPs, making diagnosis and assessment of aphasia more effective and less costly.

Mandarin and Cantonese are two prominent and influential spoken dialects in China. There have been extensive work on automatic detection or assessment of aphasia for English-speaking patients, while studies on Chinese patients relatively remain under-investigated, especially for Mandarin-speaking population. A wide range of acoustic features and text features were investigated to characterize atypical characteristics of pathological speech, including duration features [7, 8, 9], spectral features [7, 10], Goodness of Pronunciation (GOP) scores [7, 11], word frequency based features [8], Part-of-Speech based features [8], word embedding based features [9] etc. Previous studies [8, 9] suggested that text features extracted from speech transcriptions were able to outperform acoustic features in the aphasia assessment. Cross-lingual domain adaptation based on text features was adopted to detect aphasia in Mandarin-speaking subjects [12]. End-to-end approaches based on speech signals or transcriptions without explicit feature extraction were attempted for assessment of Cantonese-speaking aphasia patients [13].

In recent years, pre-trained language models (LMs) fine-tuned for downstream tasks have achieved great success in various natural language processing (NLP) tasks [14] such as question answering, name entity recognition, sentiment classification etc. With speech transcriptions as input, BERT and BERT-like models have been applied for detecting Alzheimer's disease in the challenges of ADReSS [15, 16, 17] and ADReSSo [18, 19, 20], and outperform feature-based approaches on this task [15, 17].

In the present study, pre-trained LMs, namely BERT and RoBERTa, are investigated to tackle the problem of detecting aphasia in Cantonese- and Mandarin-speaking subjects. The approaches of deriving embeddings from pre-trained LM and fine-tuning of pre-trained LM are expected to characterize language impairment as manifested in transcription of speech. The contributions of this study are as follows: (1) pre-trained LMs are adopted to alleviate data shortage problem of impaired speech and the feasibility of applying cross-lingual pre-trained LMs fine-tuned by aphasia detection for Cantonese-speaking subjects to Mandarin-speaking subjects is explored; (2) to the best of our knowledge, this paper presents the first attempt to aphasia detection for Chinese-speaking patients using pre-trained LMs, especially for Mandarin-speaking patients.

2. Datasets

Two databases of aphasic speech, namely Cantonese AphasiaBank (CAB) [21] and Mandarin AphasiaBank (MAB) [22], are used for the aphasia detection task. The former one is developed

979-8-3503-9796-3/22/$31.00

by a joint team of the University of Central Florida and the University of Hong Kong, and the latter is collected by the Nanjing Medical University. Both of them are multi-modal databases, with the goal of supporting fundamental and clinical research on Chinese-speaking aphasia population. Speech recordings in these databases were elicited following the English AphasiaBank protocol [23], with adaptation to the local Chinese culture. Each of aphasia patients (AP) and healthy control (HC) subjects was required to complete several narrative tasks, including picture description, procedure description, story telling and personal monologue. Except personal monologue, each of the remaining 7 narrative tasks has a specific topic (referred to as a "story"), e.g., telling a traditional Chinese fairy tale titled "The boy who cried wolf". The type of speech collected with this protocol is spontaneous. Speech transcriptions were annotated using the CHAT coding system [24].

Detailed information for datasets used in this paper is shown in Table 1. Due to various non-technical reasons, only 92 APs of CAB and 9 APs of MAB are usable in this study. Accordingly 92 HCs and 9 HCs are carefully selected from CAB and MAB respectively to match the age of APs as much as possible to form the CAB-full and MAB-full dataset. The CAB-full is split into a training set (CAB-train) and a test set (CAB-test) with no overlap of speaker. Further data partitioning would lead to even less training data and test data for MAB-full. They are not enough to train pre-trained LMs and give meaningful experimental results. Therefore, we decide to report leave-one-out cross validation results for aphasia detection of Mandarin-speaking subjects. The mean and standard deviation (std) values of the number of characters (#Char) per story spoken from APs and HCs are also listed in Table 1. It is found that HCs tend to produce more speech content than APs.

An overall subjective assessment score named Aphasia Quotient (AQ) was available for all APs. The AQ scores range from 0 to 100, and lower AQ means higher degree of severity.

Table 1: *Basic information about the subjects in CAB-full, CAB-train, CAB-test and MAB-full (M:male and F:female).*

Dataset	AP / HC			
	Age (std)	M	F	#Char (std)
CAB-full	54(9) / 55(12)	60 / 33	32 / 59	94(83) / 180(137)
CAB-train	54(9) / 55(12)	51 / 26	19 / 44	-
CAB-test	53(9) / 55(12)	9 / 7	13 / 15	-
MAB-full	44(13) / 50(7)	6 / 6	3 / 3	74(54) / 176(128)

3. Pre-trained LM for Aphasia Detection

Pre-trained LMs are applied to characterize language impairment manifested in aphasic speech. BERT and RoBERTa based models that support both traditional and simplified Chinese are used. They include *bert-base-chinese*, *bert-base-multilingual-cased* developed by Google [14] and *chinese-bert-wwm*, *chinese-bert-wwm-ext*, *chinese-roberta-wwm-ext*, *chinese-roberta-wwm-ext-large* developed by HFL [25]. The last four models apply Chinese word segmentation and whole word masking (wwm) techniques to the vanilla BERT/RoBERTa model. They have achieved significant performance gain in various Chinese NLP tasks [25]. All pre-trained LMs are obtained from the HuggingFace Transformers Library [26]. As shown in Figure 1, two different approaches are adopted to perform aphasia detection:

1. Pre-trained LM based embeddings: The embeddings de-

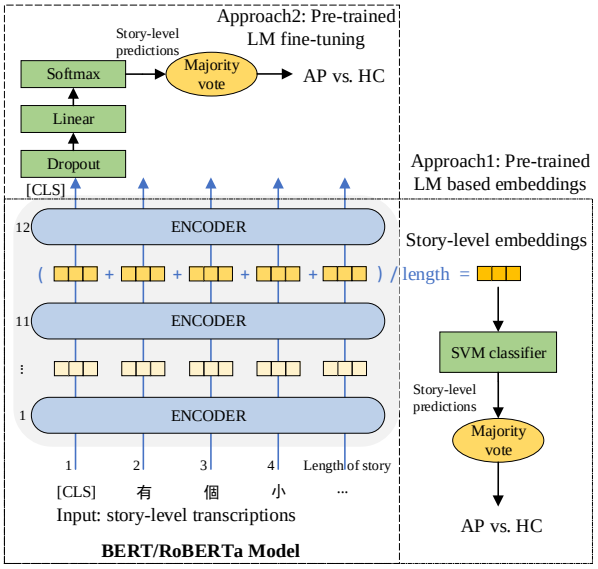

Figure 1: *Two approaches to aphasia detection using pre-trained BERT/RoBERTa models.*

rived from BERT are regarded as contextualized representations of input tokens [14]. To encode a sentence into a fixed-length vector with BERT, it is common to average the contextual embeddings at the last few layers of BERT, or use the output of the first token "[CLS]" as the sentence embedding [27]. In this study, each input token in the speech transcription of a story is encoded into a 768-dimensional contextual embedding using a frozen BERT/RoBERTa model. We perform average pooling on contextual embeddings in the second-to-last layer to represent a whole story produced by each subject. Thus, each subject can obtain 7×768-dimensional (1024-dimensional for the *chinese-roberta-wwm-ext-large* model) story-level embeddings for the 7 narrative tasks. These representations are then used to train a Support Vector Machine (SVM) to differentiate APs from HCs at story level. Majority vote is performed on 7 story-level predictions to generate a speaker-level classification result at the evaluation stage.

2. Pre-trained LM fine-tuning: A pre-trained general-purpose LM can be adapted to various downstream tasks via fine-tuning of model parameters [14]. In this study, BERT and RoBERTa models are fine-tuned on the task of aphasia detection. Specifically, the output of token "[CLS]" from a BERT/RoBERTa model is fed to a dropout layer with probability of 0.3 for the regularization purpose and then fed to a 2-dimensional linear layer with softmax activation function to perform aphasia detection at story level. The parameters of task-specific model are jointly optimized to minimize cross-entropy loss with the Adam optimizer [28]. Speaker-level predictions are obtained with majority vote during the evaluation process.

4. Experimental Setup

4.1. Models for Cantonese-Speaking Subjects

4.1.1. Baselines

1. eGeMAPS features: The eGeMAPS feature set comprises a collection of 88 feature parameters, which cover frequency, energy, spectral, temporal related features. The features were successfully applied to automatic detection of Parkinson's disease

[29] and the Alzheimer's disease [20, 30], which share many similarities with aphasia. For feature extraction, speech recording of each story is first divided into segments of 10 second long. An 88-dimensional feature vector is computed from each speech segment using the OpenSMILE toolkit [31].

2. Word2Vec based embeddings: Word2Vec based embeddings have shown effectiveness in automatic aphasia assessment [13]. The use of Word2Vec embeddings [32] is similar to the pre-trained LM based embeddings as described in Section 3. The speech transcriptions of 118 HCs (with complete transcriptions of all narrative tasks) in CAB, including $182,027$ words, are used to train $2,279$ distinct word vectors (200-dimensional). Note that a word here refers to a Chinese character. For each subject, $7\times$ 200-dimensional story-level embeddings are obtained by averaging all word vectors in accordance to the transcriptions of 7 stories.

With the segment-level eGeMAPS feature vectors or story-level embeddings, SVM is trained to perform classification of segments or stories produced by APs (labeled as 0) and those by HCs (labeled as 1). During the training stage, 5-fold cross validation is performed on CAB-train set to determine the optimal hyperparameter from {kernel: 'rbf', 'poly', 'sigmoid', 'linear'}. At the evaluation stage, the complete CAB-train set is used to train the SVM classifier with the optimal kernel. Segment-/story-level classification results are obtained for the subjects in CAB-test. Majority vote is performed on segment-/story-level results for each test speaker to give the speaker-level results.

4.1.2. Pre-trained LMs

1. Pre-trained LM based embeddings: For the first approach of using pre-trained LM based embeddings followed by a SVM classifier, the training and test procedures are the same as those of using Word2Vec based embeddings.

2. Pre-trained LM fine-tuning: The pre-trained LMs are fine-tuned on CAB-train set to classify story-level transcriptions of APs vs. HCs. Model fine-tuning is carried out with 5-fold cross validation. The following hyperparameters are chosen to maximize the average accuracy across 5 folds: learning rate = e-5, batch size = 8, max input length of 512, epochs = {20 (for *bert-base-chinese* and *chinese-bert-wwm-ext*), 25 (for *bert-base-mulitilingual-cased* and *chinese-bert-wwm*), 30 (for *chinese-bert-wwm-ext* and *chinese-roberta-wwm-ext-large*)}. Cross-validation is run five times with different random seeds, and accuracy reported is averaged across five runs. During the test procedure, distinct story-level classification results are generated with 5 different seeds from each hyperparameter-optimized model trained on the complete CAB-train set. To mitigate the effect of overfitting, the majority vote is performed on the results from 5 seeds to give final story-level classification results. Finally, majority vote is applied across 7 story-level predictions that belong to the same subject to return a single speaker-level classification result.

4.2. Models for Mandarin-Speaking Subjects

Due to the limited amount of MAB data, leave-one-out cross validation is adopted to train the classifier with baseline eGeMAPS features while the same types of pre-trained LMs are used to detect aphasia as for Cantonese-speaking subjects. The hyperparameters are chosen for all pre-trained LMs as follows: learning rate = e-5, batch size = 4, epochs = 20 and max input length of 512. Since there is no unseen test data left, we can only report the cross-validation classification results for 9 APs and 9 HCs at this stage.

The other evaluation arrangement is adopted to alleviate the problem of data scarcity. It is known that Cantonese and Mandarin are both Chinese dialects. They share common characteristics in morphology, semantics and syntax. Also, Cantonese-speaking and Mandarin-speakings APs may have similar symptoms in the linguistic aspect. Inspired by this, we employ pre-trained LMs fine-tuned by CAB-train set to detect aphasia for Mandarin subjects (MAB-full). It is feasible since pre-trained LMs support both simplified and traditional Chinese. The classification performance in such cross-lingual scenario is also reported.

5. Results and Discussion

5.1. Aphasia Detection for Cantonese-Speaking Subjects

First, we compare 5-fold cross validation and evaluation results of aphasia detection for Cantonese-speaking subjects using acoustic features (i.e., eGeMAPS features), story-level embeddings (i.e., Word2Vec and Pre-trained LM based embeddings) and pre-trained LM fine-tuning, as shown in Table 2. Average accuracy with its std value for 5-fold cross validation on CAB-train, and accuracy, macro-averaged precision, recall as well as F1 for evaluation on CAB-test are reported. Note that values of performance metrics are computed based on segment-level (only for eGeMAPS features) or story-level classification results (for other models).

It can be seen that all pre-trained LMs fine-tuned by aphasia detection task achieve better performance than eGeMAPS features and story-level embeddings in both cross-validation and evaluation results. The cross validation results show that our fine-tuned pre-trained LMs have relatively low variance in accuracy. The best evaluation accuracy is 0.925, obtained with the fine-tuned *chinese-bert-wwm* model. It outperforms eGeMAPS features and the best story-level embeddings derived from *chinese-roberta-wwm-ext* by an absolute improvement of 7.2% and 5.8% respectively. It is followed by the fine-tuned *chinese-bert-wwm-ext* and *bert-base-chinese*, achieving an accuracy of 0.912 and 0.906 respectively. We also observe that BERT model does not benefit from multilingual pre-training materials for aphasia detection. The performance of fine-tuned RoBERTa models is worse than that of BERT models although prior work shows the contrary in conventional NLP tasks [25]. For the approaches of using story-level embeddings, pre-trained LMs attain better performance than Word2Vec. This implies that contextual word embeddings derived from pre-trained LMs are more effective in modeling linguistic characteristics of impaired speech than static word embeddings.

The classification result of each subject in CAB-test is computed from segment-/story-level predictions via majority vote. Table 3 summarizes the speaker-level classification results from: (1) two baseline models; (2) the best-performing model (*chinese-roberta-wwm-ext*) using pre-trained LM derived embeddings in Table 2; (3) top three fine-tuned pre-trained LMs in Table 2. Compared to baseline models, two approaches of using pre-trained LMs obviously improve the aphasia detection performance. Results obtained with fine-tuned models of *bert-base-chinese*, *chinese-bert-wwm* and *chinese-bert-wwm-ext* are the same, with the highest accuracy of 0.98. Figure 2 shows the confusion matrices of story-level and speaker-level results given by these three models. Only one AP is mis-classified as HC among 44 test subjects. We notice that the AQ score of this speaker is 99 (highest in APs), meaning that the severity degree

Table 2: *Segment-level (only for eGeMAPS features) or story-level cross-validation and evaluation results of aphasia detection for Cantonese-speaking subjects. Top three highest accuracy (Acc.) scores are in bold and * indicates the best score.*

Type	Model	Cross-validation	Evaluation			
		Average Acc.	Precision	Recall	F1	Acc.
Acoustic features	eGeMAPS features	0.828 ± 0.168	0.852	0.849	0.850	0.853
Story-level embeddings	Word2Vec	0.788 ± 0.103	0.805	0.795	0.794	0.795
	bert-base-chinese	0.826 ± 0.085	0.831	0.831	0.831	0.831
	bert-base-multilingual-cased	0.842 ± 0.074	0.822	0.821	0.821	0.821
	chinese-bert-wwm	0.803 ± 0.114	0.852	0.851	0.850	0.851
	chinese-bert-wwm-ext	0.811 ± 0.067	0.823	0.821	0.821	0.821
	chinese-roberta-wwm-ext	0.816 ± 0.087	0.869	0.867	0.867	0.867
	chinese-roberta-wwm-ext-large	0.807 ± 0.094	0.820	0.818	0.818	0.818
Pre-trained LM fine-tuning	bert-base-chinese	$\mathbf{0.895}^{*} \pm 0.008$	0.907	0.906	0.906	**0.906**
	bert-base-multilingual-cased	0.869 ± 0.011	0.885	0.883	0.883	0.883
	chinese-bert-wwm	$\mathbf{0.892} \pm 0.013$	0.925	0.925	0.925	$\mathbf{0.925}^{*}$
	chinese-bert-wwm-ext	$\mathbf{0.891} \pm 0.012$	0.913	0.912	0.912	**0.912**
	chinese-roberta-wwm-ext	0.870 ± 0.005	0.905	0.903	0.902	0.903
	chinese-roberta-wwm-ext-large	0.879 ± 0.010	0.899	0.896	0.896	0.896

of aphasia is extremely mild, such that this speaker is not surprisingly very hard to be distinguished from HCs. This confirms the effectiveness of our fine-tuned pre-training LMs in aphasia detection.

Table 3: *Speaker-level aphasia detection results of CAB-test.*

Model	Class	Precision	Recall	F1	Acc.
eGeMAPS features	AP	0.86	0.86	0.86	0.86
	HC	0.86	0.86	0.86	
Embedding: Word2Vec	AP	1.00	0.77	0.87	0.89
	HC	0.81	1.00	0.90	
Embedding: chinese-roberta-wwm-ext	AP	0.95	0.91	0.93	0.93
	HC	0.91	0.95	0.93	
Fine-tune: bert-base-chinese	AP	1.00	0.95	0.98	**0.98**
	HC	0.96	1.00	0.98	
Fine-tune: chinese-bert-wwm(-ext)	AP	1.00	0.95	0.98	**0.98**
	HC	0.96	1.00	0.98	

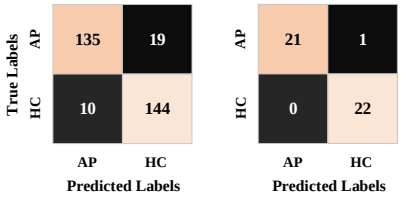

Figure 2: *Confusion matrices of story-level (left) and speaker-level (right) classification results for CAB-test.*

5.2. Aphasia Detection for Mandarin-Speaking Subjects

Due to space limitations, we only report speaker-level aphasia detection results for Mandarin-speaking subjects in Table 4. For the leave-one-out cross validation results, our best classification result attains an accuracy of 0.94 using the fine-tuned *bert-base-chinese* and *chinese-bert-wwm* models, with an absolute gain of 5% compared to the baseline using eGeMAPS features. Other pre-trained LMs perform worse than these models and their performance is omitted in the table. The number of mis-classified subjects is one and two obtained with the best-performing pre-trained LM and eGeMAPS features respectively. The only mis-classified AP given by fine-tuned BERT model is also with the highest AQ score (91.9) among all 9 APs. There is one more HC subject being mis-classified as AP using eGeMAPS features.

For the approach of applying pre-trained LMs fine-tuned with CAB-train (same models as in Table 3) to perform aphasia detection on MAB-full, an accuracy of 0.83 is achieved (three subjects are mis-classified). It performs slightly worse than previous models, but without using data from Mandarin-speaking subjects for model development. This suggests that it is promising to use a well-developed pre-trained LM fine-tuned by relatively resource-rich pathological speech in similar language to enable the detection of same disease from target low-resource pathological speech.

Table 4: *Speaker-level aphasia detection results of MAB-full.*

Leave-one-out Cross Validation Results Based on MAB-full					
Model	Class	Precision	Recall	F1	Acc.
eGeMAPS features	AP	0.89	0.89	0.89	0.89
	HC	0.89	0.89	0.89	
Fine-tune: bert-base-chinese	AP	1.00	0.89	0.94	**0.94**
	HC	0.90	1.00	0.95	
Fine-tune: chinese-bert-wwm	AP	1.00	0.89	0.94	**0.94**
	HC	0.90	1.00	0.95	
Evaluation Results Based on Pre-trained LMs Fine-tuned by CAB-train					
Fine-tune: bert-base-chinese	AP	0.80	0.89	0.84	0.83
	HC	0.88	0.78	0.82	
Fine-tune: chinese-bert-wwm	AP	0.80	0.89	0.84	0.83
	HC	0.88	0.78	0.82	

6. Conclusions

This paper has presented two pre-trained LM based approaches to automatic aphasia detection for Cantonese- and Mandarin-speaking patients. The experimental results demonstrate that BERT models fine-tuned for aphasia detection perform well, outperforming conventional acoustic features and word embedding based models. In addition, our results show the potential of leveraging pre-trained LMs fine-tuned with resource-rich pathological speech in similar language to the same detection task for target subjects with limited amount of data. In the future, we will extent our work to the AQ prediction task using pre-trained LMs and experiment with automated transcriptions instead of manual transcriptions.

7. Acknowledgements

This research is funded by the Fundamental Research Funds for the Central Universities 2021RC244.

8. References

[1] J. C. Rosenbek, L. L. LaPointe, and R. T. Wertz, *Aphasia: A clinical approach*. Pro Ed, 1989.

[2] C. S. Brown and W. L. Cullinan, "Word-retrieval difficulty and disfluent speech in adult anomic speakers," *J. Speech Lang. Hear. R.*, vol. 24, no. 3, pp. 358–365, 1981.

[3] Australian Aphasia Rehabilitation Pathway, "Stroke and aphasias," [Online]. Available: http://www.aphasiapathway.com.au/?name=About-aphasia.

[4] W. Wang, B. Jiang, H. Sun, X. Ru *et al.*, "Prevalence, incidence, and mortality of stroke in China: results from a nationwide population-based survey of 480 687 adults," *circulation*, vol. 135, no. 8, pp. 759–771, 2017.

[5] S. Wu, B. Wu, M. Liu, Z. Chen, W. Wang, C. S. Anderson, P. Sandercock, Y. Wang, Y. Huang, L. Cui, C. Pu *et al.*, "Stroke in China: advances and challenges in epidemiology, prevention, and management," *The Lancet Neurology*, vol. 18, no. 4, pp. 394–405, 2019.

[6] P. M. Pedersen, H. Stig Jørgensen, H. Nakayama, H. O. Raaschou, and T. S. Olsen, "Aphasia in acute stroke: incidence, determinants, and recovery," *Ann. Neurol.*, vol. 38, no. 4, pp. 659–666, 1995.

[7] D. Le and E. M. Provost, "Modeling pronunciation, rhythm, and intonation for automatic assessment of speech quality in aphasia rehabilitation," in *Proc. INTERSPEECH*, 2014, pp. 1563–1567.

[8] D. Le, K. Licata, and E. M. Provost, "Automatic quantitative analysis of spontaneous aphasic speech," *Speech Communication*, vol. 100, pp. 1–12, 2018.

[9] Y. Qin, T. Lee, and A. P. Kong, "Automatic assessment of speech impairment in Cantonese-speaking people with aphasia," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 2, pp. 331–345, 2020.

[10] D. Le, K. Licata, E. Mercado, C. Persad, and E. M. Provost, "Automatic analysis of speech quality for aphasia treatment," in *Proc. ICASSP*, 2014, pp. 4853–4857.

[11] D. Le, K. Licata, and E. M. Provost, "Automatic paraphasia detection from aphasic speech: A preliminary study," in *Proc. INTERSPEECH*, 2017, pp. 294–298.

[12] A. Balagopalan, J. Novikova, M. B. A. McDermott, B. Nestor, T. Naumann, and M. Ghassemi, "Cross-language aphasia detection using optimal transport domain adaptation," in *Proc. Machine Learning for Health Workshop, ML4H@NeurIPS*, vol. 116, 2019, pp. 202–219.

[13] Y. Qin, Y. Wu, T. Lee, and A. P. Kong, "An end-to-end approach to automatic speech assessment for Cantonese-speaking people with aphasia," *J. Signal Process. Syst.*, vol. 92, no. 8, pp. 819–830, 2020.

[14] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[15] 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*, 2020, pp. 2162–2166.

[16] R. Pappagari, J. Cho, L. Moro-Velázquez, and N. Dehak, "Using State of the Art Speaker Recognition and Natural Language Processing Technologies to Detect Alzheimer's Disease and Assess its Severity," in *Proc. INTERSPEECH*, 2020, pp. 2177–2181.

[17] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To BERT or not to BERT: Comparing Speech and Language-Based Approaches for Alzheimer's Disease Detection," in *Proc. INTERSPEECH*, 2020, pp. 2167–2171.

[18] Y. Pan, B. Mirheidari, J. M. Harris, J. C. Thompson *et al.*, "Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic- and BERT-Based Alzheimer's Dementia Detection Through Spontaneous Speech," in *Proc. INTERSPEECH*, 2021, pp. 3810–3814.

[19] 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*, 2021, pp. 3805–3809.

[20] R. Pappagari, J. Cho, S. Joshi, L. Moro-Velázquez, P. Żelasko, J. Villalba, and N. Dehak, "Automatic Detection and Assessment of Alzheimer Disease Using Speech and Language Technologies in Low-Resource Scenarios," in *Proc. INTERSPEECH*, 2021, pp. 3825–3829.

[21] A. P.-H. Kong and S.-P. Law, "Cantonese Aphasiabank: An annotated database of spoken discourse and co-verbal gestures by healthy and language-impaired native Cantonese speakers," *Behav. Res. Methods*, vol. 51, no. 3, pp. 1131–1144, 2019.

[22] Z. Jiang, F. Lin, W. Xiang, Z. Chen, B. Deng *et al.*, "Aphasiabank Mandarin JiangLin PWAs," [Online]. Available: https://aphasia.talkbank.org/access/Mandarin/Aphasia/JiangLin.html.

[23] A. Kertesz, *WAB-R: Western Aphasia Battery-Revised*. PsychCorp, 2007.

[24] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk, Volume II: the Database*, 3rd ed. Psychology Press, 2014.

[25] Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, S. Wang, and G. Hu, "Pre-training with whole word masking for Chinese BERT," *arXiv preprint arXiv:1906.08101*, 2019.

[26] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi *et al.*, "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP*, 2020, pp. 38–45.

[27] B. Li, H. Zhou, J. He, M. Wang, Y. Yang, and L. Li, "On the sentence embeddings from pre-trained language models," in *Proc. EMNLP*, 2020, pp. 9119–9130.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins, and R. H. Ghomi, "Parkinson's disease diagnosis using machine learning and voice," in *Proc. SPMB*, 2018, pp. 1–7.

[30] J. Koo, J. H. Lee, J. Pyo, Y. Jo, and K. Lee, "Exploiting Multi-Modal Features from Pre-Trained Networks for Alzheimer's Dementia Recognition," in *Proc. INTERSPEECH*, 2020, pp. 2217–2221.

[31] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia*, 2010, pp. 1459–1462.

[32] Google Inc., "word2vec," 2013, [Online]. Available: https://code.google.com/archive/p/word2vec/.