Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

PPGs-BERT: Leveraging Phoneme Sequence and BERT for Alzheimer's Disease Detection from Spontaneous Speech

Qi Sun[1,*], *Ziyue Qiu*[2,*], *Yu Pu*[1], *Jinpeng Li*[1], *Xuchu Chen*[1], *Wei-Qiang Zhang*[1,#]

[1]Department of Electronic Engineering, Tsinghua University, China
[2]Ira A. Fulton Schools of Engineering, Arizona State University, United States
sun-q24@mails.tsinghua.edu.cn, zqiu30@asu.edu, wqzhang@tsinghua.edu.cn

Abstract

Alzheimer's Disease (AD) is a neurodegenerative condition characterized by linguistic impairments. While ASR and LLMs show promise in AD detection, ASR often normalizes key AD-related speech patterns and faces cross-lingual challenges due to language dependencies. Besides, ASR training demands extensive matched data. In our paper, however, we employ a phoneme recognizer as a frontend tokenizer. Provided it has comprehensive phoneme coverage, a multitude of linguistic phenomena can be represented via phoneme sequences, including hesitations, repetitions, pauses, mispronunciations, and even distinctions between different language identities that are crucial for AD detection. Furthermore, the BERT model is employed to extract high-dimensional features from the Phonetic PosteriorGrams (PPGs), which are ultimately used to diagnose Alzheimer's disease. Our approach offers cross-lingual applicability, achieves competitive accuracy, and maintains computational efficiency.

Index Terms: phonetic posteriorGrams, Alzheimer's disease detection, cross-lingual, lightweight

1. Introduction

Alzheimer's Disease (AD), a progressive neurodegenerative disorder, profoundly affects spontaneous speech, leading to disfluencies such as increased hesitations, prolonged pauses, and difficulties in lexical retrieval [1]. These speech anomalies serve as critical indicators of cognitive decline, making speech-based assessment a promising avenue for early AD detection. Recent studies have underscored the importance of preserving these subtle speech features, particularly through phoneme-based models, to enhance the accuracy and robustness of AD classification [2].

In addition, the application of deep learning models has shown significant promise in the domain of AD detection. Zolnoori et al. leveraged DistilBERT for AD classification, demonstrating its clinical potential due to its efficient architecture and reliable performance [3]. Similarly, Gómez-Zaragozá et al. highlighted the utility of pauses and punctuation in speech, employing state-of-the-art machine learning methods to enhance AD detection accuracy [4]. These studies reinforce the significance of subtle speech patterns in cognitive assessment and highlight the growing reliance on advanced language models for clinical applications.

Despite advancements in Automatic Speech Recognition (ASR)-based approaches, significant challenges persist [5–13]. ASR systems inherently normalize speech by correcting pronunciation and sentence structure, often leading to the loss of critical disfluencies such as hesitations, repetitions, and pauses—features that are highly indicative of AD [14, 15]. Although prior works, such as those by Pan et al. [16] and Pappagari et al. [17], successfully integrated ASR outputs with BERT for AD classification, their reliance on language-specific transcriptions limits cross-lingual generalization. This challenge is further compounded when applying models trained on one language to another, due to variations in phonetic structures and linguistic patterns.

Nie et al. addressed these limitations by extending the BERT framework to accept PPGs derived from a front-end phoneme recognizer, preserving phoneme-level speech features and demonstrating enhanced performance in cross-lingual speech tasks [18]. PPGs, which retain silent syllables, repeated syllables, and filler sounds like "um" and "uh," offer a robust representation of speech that is consistent across languages. While several studies have leveraged PPGs for detecting speech anomalies such as pauses and mispronunciations [19, 20], their application to AD detection remains underexplored.

In this paper, we propose the PPGs-BERT model, which removes the potential for ASR-induced modifications to AD-related features such as hesitations and pauses. In addition, phonemes show stronger cross-linguistic generalization than text, allowing our model to be applied across languages. Furthermore, the backend BERT model enables the recognition of phoneme sequences to complete the detection process. PPGs-BERT provides a lightweight, efficient, and cross-lingual alternative to traditional systems, preserving speech features essential for accurate AD detection.

2. Methods

2.1. Model Overview

The PPGs-BERT model consists primarily of three components: a phoneme recognizer, BERT model, and a binary classification network, as illustrated in Figure 1. The model primarily utilizes the BERT model to extract high-dimensional representations from phonetic posteriorGrams (PPGs) for Alzheimer's disease detection.

Alzheimer's disease detection is achieved through two stages of training. In the first stage, a phoneme recognizer is employed to extract features from PPGs. These features, along with positional encodings and temporally masked data, serve as input to the BERT model, which is then trained using Masked Language Modeling (MLM). In the second stage, during Alzheimer's disease detection, the parameters of the PPGs-BERT model are frozen. The PPGs-BERT model is utilized to extract high-dimensional representations from speech, followed

[*]These authors contributed equally.
[#]Corresponding author.

This work was supported by the National Natural Science Foundation of China under Grant No. 62276153.

10.21437/Interspeech.2025-489

by temporal average pooling. Subsequently, a binary classification network comprising two fully connected layers is trained.

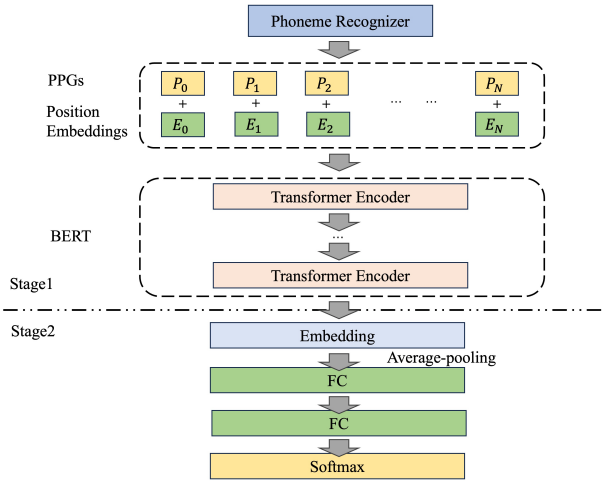


Figure 1: *PPGs-BERT model structure.*

2.2. Phonetic Posterior-Based Feature Extraction

Phonetic PosteriorGrams (PPGs) represent the posterior probability of each phoneme class corresponding to each specific time frame in the speech [21]. Assuming the phoneme set of speech is $\mathcal{P}$, a phoneme recognizer $g_p(x)$ can be trained using labeled data, and the posterior probability of $\boldsymbol{x}$ for each phoneme $\boldsymbol{p}$ can be calculated given the acoustic feature vector $\boldsymbol{x}$.

In this study, we extracted phoneme posterior probabilities from speech using a pre-trained CRNN phoneme recognizer. During training, the phoneme recognizer first extracts Mel-spectrogram features from the speech and converts them to a decibel scale. Subsequently, the extracted features are fed into a model comprising convolutional neural networks and bidirectional long short-term memory networks (BiLSTM) for phoneme classification.

To pre-train the model, we utilized the TIMIT speech corpus, which comprises a phonetically labeled collection of utterances from 630 speakers across eight distinct American English dialect regions. This pure English dataset is sampled at 16 kHz and includes 61 phonemes, including 48 consonant-vowel combinations, 12 vowels, and a silence symbol [22]. Therefore, a phoneme recognizer can be used to extract PPG features from speech that are independent of the speaker.

2.3. BERT

In the original BERT model, the input comprises a combination of word embeddings, positional embeddings, and segment embeddings. However, in the PPGs-BERT model, the input features are PPGs obtained by converting speech into phonemes using a phoneme recognizer. The phoneme posterior probabilities at each frame are continuous distributions, making it infeasible to generate corresponding word embeddings as in the original BERT model. To address this challenge, this study modifies the input layer of the BERT model by removing the token embedding layer and the segment embedding layer. Instead, the PPGs features are directly added with positional encodings and used as the input to the BERT model. During training, 15% of the PPGs features in a mini-batch are randomly selected and processed according to the masking scheme of the original BERT model. The modified BERT model performs only the MLM task, which is used to reconstruct the masked data. The reconstruction loss is calculated using the mean squared error. This modified BERT model is capable of extracting high-dimensional features from the PPGs.

3. Experienment

3.1. Monolingual Experiment

To assess the viability of the PPGs-BERT model for AD detection tasks, we conduct a monolingual AD detection experiment using the English ADReSSo [23] dataset. Initially, we train the PPGs-BERT model on samples derived from healthy individuals within the dataset, to capture and understand phoneme patterns inherent in normal human speech. Subsequently, we train downstream linear classification models on this dataset. Our experimental design employed a binary classification task, specifically discerning between healthy individuals and AD patients, aligning with the ADReSSo challenge.

Our approach utilizes PPGs as input to a BERT-based classifier, ensuring that speech features are retained before textual transformation. We hypothesize that pauses and hesitations are critical features that should not be ignored.To validate our hypothesis, we perform an ablation study on the TIMIT-61 dataset by removing the following phonemes:

- **pau** (pause)
- **epi** (epenthetic vowels, e.g., inserted schwa sounds)

To evaluate the impact of hesitation-related phonemes on AD detection, we train two variations of the PPGs-BERT model. The Baseline Model is trained on the full TIMIT-61 phoneme set, while the Filtered Model is trained on a modified version of the dataset in which the phonemes ”pau” and ”epi” are removed.

Additionally, we explore the structural aspects of PPGs-BERT, specifically experimenting with varying numbers of layers of self-attention blocks. Our investigation involves evaluating model performance using 2, 4, 6, and 8 layers of self-attention blocks, respectively, aiming to determine the most optimal model structure for AD detection.

3.2. Cross-Lingual Adaption Experiment

We utilize three datasets in English, Chinese, and Spanish following the defined datasplits for cross-lingual AD detection. Since mild cognitive impairment (MCI) is an early AD stage, we merge MCI into AD, making the task binary classification between healthy controls (HC) and AD patients.

Table 1: *Summary of Datasets Used in This Study*

Dataset	Language	Groups	Total
ADReSSo [23]	English	AD: 122, HC: 115	237
NCMMSC [24]	Chinese	AD: 26, MCI: 54, HC: 44	280
Ivanova [25]	Spanish	AD: 74, MCI: 91, HC: 197	362

To examine the cross-lingual generalization capabilities of PPGs, we conduct a cross-lingual adaptation experiment employing the phoneme recognizer and PPGs-BERT trained on one language data, consistent with our monolingual AD detection experiment, we freeze the parameters of PPGs-BERT. Then we adapt the model to the the other two languages domain

by solely training the parameters of downstream classification models on the other two language datasets, shown in Table 4.

3.3. Experimental Settings

Since the audio file lengths in the dataset are not uniform, and the maximum input length of the BERT model is 512, we used a time window of 10s and a window shift of 6s to segment the training and test sets. For audio files with fewer segments after segmentation in the training set, we used data augmentation with noise to obtain a balanced training set. The training data was then split into a training set and a validation set in a ratio of 4:1. During testing, each segment of the test set was tested separately, and the classification probabilities obtained were normalized. The classification probabilities of each segment of an audio file were then added together to obtain the result for that audio file.

Due to the utilization of a phoneme recognizer to extract 61-dimensional features from phoneme posterior probability graphs, the feature dimensionality of the BERT model is accordingly set to 61. The BERT model comprises primarily 6 layers of Transformer Encoder, with each layer containing 12 attention heads. During the initial training phase of the PPGs-BERT model, Adam optimizer is employed with a learning rate of 0.0002 for 30 iterations. Subsequently, in the second training phase, the parameters of the PPGs-BERT model are fixed. This model is then utilized to extract high-dimensional representations of each segment, followed by averaging pooling of the BERT outputs. Consequently, each speech segment is transformed into a 61-dimensional vector. Subsequent classification involves passing this vector through two fully connected layers with ReLU activation functions, with 32 and 2 neurons, respectively. Training of the classification network employs stochastic gradient descent with momentum, utilizing a learning rate of 0.0001 and a momentum of 0.9 for 20 iterations.

As for the control experiments using MobileNetV2 and ResNet, the same data processing procedures are followed, Mel-spectrogram features of the audio are extracted, logarithmically transformed, and then fed into the respective neural networks. Training is conducted using Adam optimizer with a learning rate of 0.0001 for 40 iterations.

4. Result

4.1. Monolingual Experiment

Table 2 presents the classification performance on ADReSSo. Removing hesitation-related phonemes led to a significant drop in accuracy (-14.0%) and F1-score (-11.9%), indicating their importance in AD detection. The filtered model which removes "pau" and "epi" misclassified more AD cases as healthy controls, emphasizing the role of hesitation cues.

Table 2: *Effect of Removing Hesitation-Related Phonemes on AD Detection*

Model	Accuracy (%)	Recall (%)	Precision (%)	F1 (%)
Unfilitered	73.2	74.3	72.2	73.2
Filtered	59.2	65.7	57.5	61.3

Table 3 compares PPGs-BERT-6 with existing AD detection models. PPGs-BERT-6 achieves the highest accuracy (73.2%), surpassing SVM-based methods (eGeMAPs, Acoustic features) and CNN-based models (MobileNetV2, ResNet), which struggle with acoustic variability. ASR-based methods (Wav2vec + BERT, ASpIRE + BERT) show lower generalization, as text-based models fail to retain phoneme-level disfluencies critical for AD detection.

Table 3: *Comparison of PPGs-BERT with Other Works*

Model	Accuracy	Recall	Precision	F1
eGeMAPs + SVM [26]	64.7	–	–	–
Acoustic features + SVM [27]	64.8	31.4	91.7	46.8
MobileNetV2 [28]	64.8	62.9	64.7	63.8
ResNet [29]	63.4	80.0	59.6	68.3
Wav2vec + BERT [30]	73.2	73.2	73.4	73.2
ASpIRE + BERT [31]	63.1	–	–	–
DeepSpeech + BERT [32]	70.5	–	–	–
LogMel + BERT (ablation)	54.9	97.1	13.9	68.0
PPGs-BERT-6	**73.2**	**74.3**	**72.2**	**73.2**
PPGs-BERT-8	66.2	51.4	72.0	60.0
PPGs-BERT-4	63.4	77.1	60.0	67.5

To assess the impact of Transformer depth, we evaluate PPGs-BERT with different layer configurations (4, 6, 8 layers). PPGs-BERT-6 achieves the best balance, while PPGs-BERT-4 underfits due to limited capacity, and PPGs-BERT-8 overfits, favoring a specific class. Figure 2 presents the confusion matrices, showing classification biases in deeper models due to data limitations.

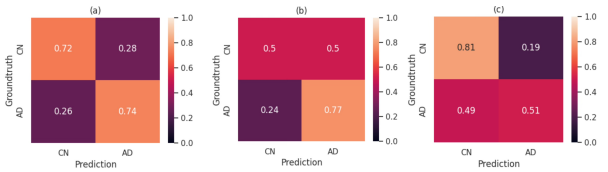


Figure 2: *The confusion matrices for the PPGs-BERT-6, PPGs-BERT-8, and PPGs-BERT-4.*

The ***eGeMAPS+SVM*** model [26], the official baseline from the INTERSPEECH ADReSSo challenge , utilizes the eGeMAPS paralinguistic feature set with an SVM classifier. Similarly, in the ***Acoustic Features+SVM*** model [27], features such as mean, variance, and peak of Mel frequency cepstrum coefficients were extracted to form a 168-dimensional feature set, which was then classified using support vector machines. These two methods have certain detection effects, possibly due to the fact that the paralinguistic feature set can reflect information such as speech speed and pauses, while there are certain differences between patients and healthy individuals.

Two CNN-based models, ***MobileNetV2*** [28] and ***ResNet*** [29] both extract logarithmic Mel spectral features from audio and input them into their respective neural networks for classification. ***MobileNetV2*** uses Mel-spectrogram features as input and employs neural networks for classification. ***ResNet***, on the other hand, processes Mel-spectrogram images using a convolutional residual network, classifying speech samples into categories of Alzheimer's disease patients and healthy individuals. The effectiveness of traditional neural network methods in detecting Alzheimer's disease is limited, possibly due to the difficulty of conventional convolutional neural networks in extracting long-distance features.

The ASR-based models, ***Wav2vec+BERT*** [30], ***ASpIRE+BERT*** [31], and ***DeepSpeech+BERT*** [32], follow a

Table 4: *Cross-lingual Performance of PPGs-BERT-6 (Accuracy %)*

Model	**Training: ADReSSo**			**Training: NCMMSC**			**Training: Ivanova**		
Testing	ADReSSo	NCMMSC	Ivanova	ADReSSo	NCMMSC	Ivanova	ADReSSo	NCMMSC	Ivanova
Acoustic + SVM	64.8	62.3	60.5	52.1	81.1	58.1	63.4	62.3	72.1
PPGs-BERT-6 (Ours)	**73.2**	**86.5**	**73.7**	**63.0**	**82.4**	**61.4**	**69.0**	**55.5**	**66.7**

speech-to-text pipeline. ***Wav2vec+BERT*** extracts text representations from Wav2vec transcriptions. This study has achieved comparable performance, indicating that high-dimensional representation extracted from PPGs can reflect the sound change of patients to a certain extent. ***ASpIRE+BERT*** used a pre-trained ASpIRE recipe in Kaldi. ASpIRE was a far-field speech recognition challenge held by IARPA. ***DeepSpeech+BERT*** generate transcripts from spontaneous speech using a pre-trained English DeepSpeech model from Mozilla [33]. Although these approaches achieve strong performance on monolingual datasets, they suffer from feature loss, as ASR models normalize speech patterns, removing essential disfluencies such as hesitations and pauses—key markers for AD detection. Additionally, these models perform poorly in cross-lingual and multilingual settings, where phoneme-level cues are more consistent than text-based transcriptions.

To further assess the contribution of PPGs, we conducted an ablation experiment by replacing PPG inputs with logarithmic Mel-spectral features (LogMel+BERT). Both models share the same BERT architecture and training parameters, ensuring a controlled comparison. The results indicate a significant performance drop when using LogMel features, underscoring the importance of phoneme-level information. Unlike Mel-spectral features, which emphasize low-level acoustic patterns, PPGs capture critical disfluencies such as fillers, repeated syllables, and pauses, which are strong indicators of cognitive impairment.

This comprehensive comparison highlights that PPGs-BERT-6 outperforms most baselines by effectively preserving speech characteristics that are often lost in ASR-based and CNN-based models. The ablation study further confirms that PPGs are pivotal for robust AD detection, as they retain language-agnostic phoneme structures essential for cross-lingual adaptability. These findings emphasize the superiority of PPG-based modeling over traditional acoustic and text-based methods, making PPGs-BERT-6 a good and scalable solution for multilingual AD diagnosis.

4.2. Cross-Lingual Performance on PPGs-BERT

Table 4 presents the cross-lingual performance of PPGs-BERT-6 and the Acoustic+SVM model. Overall, PPGs-BERT-6 demonstrates superior performance when the training and testing datasets are the same. Moreover, under cross-lingual settings, it exhibits a smaller performance drop compared to Acoustic+SVM, indicating stronger cross-lingual capability. For instance, when trained on the English ADReSSo dataset, it attains 73.2% accuracy on ADReSSo, 86.5% on the Chinese NCMMSC dataset, and 73.7% on the Spanish Ivanova dataset. This significant performance gain over Acoustic+SVM, especially in cross-lingual settings (24.2% higher on NCMMSC and 13.2% higher on Ivanova), highlights the advantage of modeling phoneme-level features that are more consistent across languages compared to language-specific acoustic patterns.

The performance drop observed when PPGs-BERT-6 is trained on non-English datasets and tested on ADReSSo (63.0% and 69.0%, respectively) due to cross-linguistic setting. However, PPGs-BERT-6 still outperforms Acoustic+SVM by a substantial margin, emphasizing its robustness and cross-lingual adaptability. This resilience can be attributed to the use of PPGs, which capture universal speech characteristics such as pauses, hesitations, and articulation patterns that are indicative of cognitive decline, regardless of the language spoken.

Furthermore, the 82.4% accuracy achieved when training and testing on the NCMMSC dataset demonstrates that PPGs-BERT-6 can effectively model languages with fewer similarities to English, such as Chinese, while maintaining strong within-language performance. Similarly, the 66.7% accuracy on the Ivanova dataset highlights the model's ability to identify AD-related phonetic patterns in Spanish, reinforcing its effectiveness across typologically diverse languages.

PPGs-BERT-6 excels by retaining phoneme-level speech cues, including disfluencies, hesitations, and pauses, which are essential AD markers. In contrast, ASR-based models normalize speech, removing critical hesitation patterns and limiting their cross-lingual performance. CNN-based models, relying on low-level acoustic features, are sensitive to pitch and accent variations and fail to capture high-denominational phoneme feature necessary for AD detection. Unlike ASR and CNN approaches, PPGs-BERT-6 models phoneme sequences in a language-agnostic manner, ensuring greater generalization.

In summary, the generally superior performance of PPGs-BERT-6 can be attributed to two key factors. First, its phoneme recognizer extracts universal phonetic patterns that are less sensitive to linguistic variations, enabling it to capture language-independent speech characteristics associated with AD. Second, our model leverages BERT's sequential modeling to capture high-denominational feature while preserving speech disfluencies crucial for AD detection. This combination allows PPGs-BERT-6 to bridge the gap between different languages, achieving stable performance without requiring language-specific retraining.

5. Conclusion

PPGs-BERT demonstrates reliable recognition performance with consistent phoneme coverage, making it well-suited for limited AD datasets. Notably, it does not require diverse matched data during training, thereby enhancing its adaptability to resource-constrained settings. Unlike ASR-based approaches that normalize speech, the model leverages phonetic posteriorgrams and preserves critical speech disfluencies such as hesitations and pauses. The proposed model outperforms baseline methods, achieving an 8.41% improvement over the ADReSSo challenge baseline, while demonstrating superior cross-lingual adaptability across English, Chinese, and Spanish datasets. It offers a cross-lingual, lightweight, and effective substitute for conventional ASR-BERT systems while maintaining speech characteristics necessary for precise AD detection.

6. References

[1] I. Hoffmann, D. Nemeth, C. D. Dye, M. Pákáski, T. Irinyi, and J. Kálmán, "Temporal parameters of spontaneous speech in alzheimer's disease," *International Journal of Speech-Language Pathology*, vol. 12, no. 1, pp. 29–34, 2010.

[2] S. Saeedi, S. Hetjens, M. O. W. Grimm, and B. Barsties v. Latoszek, "Acoustic speech analysis in Alzheimer's disease: A systematic review and meta-analysis," *The Journal of Prevention of Alzheimer's Disease*, vol. 11, pp. 1789–1797, 2024.

[3] M. Zolnoori, A. Zolnour, and M. Topaz, "Adscreen: A speech processing-based screening system for automatic identification of patients with Alzheimer's disease and related dementia," *Artificial Intelligence in Medicine*, vol. 143, no. C, 2023, art. no. 102624.

[4] L. Gómez-Zaragozá, S. Wills, C. Tejedor-Garcia, J. Marín-Morales, M. Alcañiz, and H. Strik, "Alzheimer disease classification through ASR-based transcriptions: Exploring the impact of punctuation and pauses," in *Proc. Interspeech*, 2023, pp. 2403–2407.

[5] V. T. Bui, T. C. Luong, and O. T. Tran, "Transformer-based joint learning approach for text normalization in Vietnamese automatic speech recognition systems," *Cybernetics and Systems*, vol. 55, no. 7, pp. 1614–1630, 2024.

[6] H. Wang, Z. Jin, M. Geng, S. Hu, G. Li, T. Wang, H. Xu, and X. Liu, "Enhancing pre-trained ASR system fine-tuning for dysarthric speech recognition using adversarial data augmentation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024, pp. 12 311–12 315.

[7] W. Yu, C. Tang, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Connecting speech encoder and large language model for ASR," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024, pp. 12 637–12 641.

[8] Y. Hono, K. Mitsuda, T. Zhao, K. Mitsui, T. Wakatsuki, and K. Sawada, "Integrating pre-trained speech and language models for end-to-end speech recognition," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 13 289–13 305.

[9] T. Udagawa, M. Suzuki, G. Kurata, M. Muraoka, and G. Saon, "Multiple representation transfer from large language models to end-to-end ASR systems," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024, pp. 10 176–10 180.

[10] Z. Shah, S.-A. Qi, F. Wang *et al.*, "Exploring language-agnostic speech representations using domain knowledge for detecting Alzheimer's dementia," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023, pp. 1–2.

[11] J. Li and W.-Q. Zhang, "Whisper-based transfer learning for Alzheimer disease classification: Leveraging speech segments with full transcripts as prompts," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024, pp. 11 211–11 215.

[12] Z. Cui, W. Wu, C. Zhang, W.-Q. Zhang, and J. Wu, "Transferring speech-generic and depression-specific knowledge for Alzheimer's disease detection," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2023, pp. 1–8.

[13] X. Chen, Y. Pu, and W.-Q. Zhang, "Detection of Alzheimer's disease based on dVAE-BERT model," *Acta Electronica Sinica*, vol. 52, no. 9, 2024, art. no. 2971-2978.

[14] X. Qi, Q. Zhou, J. Dong, and W. Bao, "Noninvasive automatic detection of alzheimer's disease from spontaneous speech: a review," *Frontiers in Aging Neuroscience*, vol. 15, 2023, art. no. 1224723.

[15] Y. Nie, J. Zhao, Z. Qiu, J. Bai, and W.-Q. Zhang, "Leveraging bert to improve spoken language identification of code-switching speech," *International Journal of Asian Language Processing*, vol. 34, no. 1, 2024, art. no. 2450003.

[16] Y. Pan, B. Mirheidari, J. M. Harris, J. C. Thompson, M. Jones, J. S. Snowden, D. Blackburn, and H. Christensen, "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.

[17] 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.

[18] Y. Nie, J. Zhao, W.-Q. Zhang, and J. Bai, "Bert-lid: Leveraging bert to improve spoken language identification," in *Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2022, pp. 384–388.

[19] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[20] Y. Qin, T. Lee, and A. P. H. Kong, "Combining phone posteriorgrams from strong and weak recognizers for automatic speech assessment of people with aphasia," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019, pp. 6420–6424.

[21] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2009, pp. 421–426.

[22] C. Lopes and F. Perdigao, "Phoneme recognition on the timit database," in *Speech Technologies*, I. Ipsic, Ed. Rijeka: IntechOpen, 2011, ch. 14.

[23] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Proc. of Interspeech*, Shanghai, China, 2020.

[24] X.-C. Chen, W.-Q. Zhang, and Y. Ma, "Raw waveform-based end-to-end Alzheimer's disease detection method," *Acta Electron. Sin.*, vol. 51, no. 12, pp. 3582–3590, 2023.

[25] O. Ivanova, J. J. G. Meilán, F. Martínez-Sánchez, I. Martínez-Nicolás, T. E. Llorente, and N. C. González, "Discriminating speech traits of Alzheimer's disease assessed through a corpus of reading task for Spanish language," *Computer Speech and Language*, vol. 73, 2022, art. no. 101341.

[26] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo challenge," *arXiv preprint arXiv:2104.09356*, 2021.

[27] A. Balagopalan and J. Novikova, "Comparing acoustic-based approaches for Alzheimer's disease detection," in *Proc. Interspeech*, 2021, pp. 3800–3804.

[28] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[30] Y. Zhu, A. Obyat, X. Liang, J. A. Batsis, and R. M. Roth, "WavBERT: Exploiting semantic and non-semantic speech using Wav2vec and BERT for dementia detection," in *Proc. Interspeech*, 2021, pp. 3790–3794.

[31] 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.

[32] J. Chen, J. Ye, F. Tang, and J. Zhou, "Automatic detection of Alzheimer's disease using spontaneous speech only," in *Proc. Interspeech*, 2021, pp. 3830–3834.

[33] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.