



Whisper-Based Multilingual Alzheimer’s Disease Detection and Improvements for Low-Resource Language

Kaichen Jia^{1,}, Jinpeng Li^{1,*}, Ke Li², Wei-Qiang Zhang^{1,#}*

¹Department of Electronic Engineering, Tsinghua University, China

²Beijing Haitian Ruisheng Science Technology Ltd., China

`jk21@mails.tsinghua.edu.cn`, `wqzhang@tsinghua.edu.cn`

Abstract

Alzheimer’s Disease (AD) poses a growing global health challenge due to population aging. Using spontaneous speech for the early diagnosis of AD has emerged as a notable area of research. In response to the global trend of AD, our study proposes a speech-based multilingual AD detection method. In our study, we utilize Whisper for transfer learning to build a multilingual pre-trained AD diagnostic model that achieves 81.38% accuracy on a test set comprising multiple languages. To enhance low-resource language performance, we fine-tune the pre-trained model with multilingual data and full transcripts as prompts, achieving a 4-7% accuracy improvement. Additionally, we incorporate the speaker’s background information, enhancing the accuracy of low-resource languages by 11-13%. The results demonstrate the validity of our work in multilingual Alzheimer’s detection tasks and also illustrate the feasibility of our approach in addressing the global need for Alzheimer’s detection.

Index Terms: multilingual, Whisper, transfer learning, Alzheimer’s disease classification, low-resource language

1. Introduction

Alzheimer’s disease is a prevalent neurodegenerative disorder and constitutes the primary form of dementia among the elderly population. Current estimates suggest that approximately 55 million individuals globally are affected by Alzheimer’s disease or other types of dementia. As a major global health challenge, the incidence of Alzheimer’s disease is rising in tandem with aging populations across various regions [1, 2]. Clinical research indicates that early intervention and preventive measures can substantially enhance the condition of patients with Alzheimer’s disease. Consequently, there is an increased focus on the diagnostic detection of potential cases of the disease [3].

Speech-based classification methods for detecting Alzheimer’s disease have been recognized as effective for early diagnosis and prevention, attracting considerable attention in recent studies. Nonetheless, existing research predominantly addresses single-language tasks related to Alzheimer’s disease. Zhu et al. [4] proposed a generalizable framework: leveraging text generated by an ASR model (wav2vec) as input to a BERT-based feature extractor, followed by an inference layer for downstream Alzheimer’s disease classification tasks. Wang et al. [5] investigates the use of prompt-based fine-tuning of pre-trained language models, which further optimized AD

classification performance by incorporating disfluency features such as pauses into prompt. Pan et al. [6] explored the performance of acoustic information and semantic features extracted from recordings in AD detection by using two ASR paradigms, wav2vec2.0 and time delay neural networks. Pérez-Toro et al. [7] achieved good results in AD classification tasks by extracting acoustic features such as x-vectors, preludes, and emotion embeddings as well as linguistic features such as complexity and word embeddings. Although these studies approach the topic from different perspectives, they all use English as the research language. Most studies in this field focus on tasks involving English, although research has also been conducted in other languages, including Chinese, Spanish, and some minority languages like Hungarian [8–10]. Notably, some studies have identified shared audio characteristics across various AD language tasks and have explored the potential for transfer effects in cross-language AD recognition [11]. For example, Pérez-Toro et al. [12] explored the effects of transfer learning from English to Spanish by applying a classifier trained on the English Alzheimer’s corpus to Spanish data. Luz et al. [13] proposed the ADReSS-M challenge to investigate the extent to which models primarily trained on English data can generalize to Greek-language Alzheimer’s disease detection tasks. These cross-linguistic studies provide a theoretical foundation for the development of multilingual Alzheimer’s disease detection models.

The Whisper-based transfer learning approach aims to leverage the original Whisper model for fine-tuning in Alzheimer’s disease classification tasks [14]. Whisper [15], a Transformer-based multilingual speech recognition model, is trained on extensive and diverse speech datasets, enabling effective processing of audio inputs across languages and accents. Renowned for its robustness in noisy environments and high accuracy in speech-to-text transcription, Whisper is widely applicable to automatic speech recognition (ASR), speech translation, and speech analysis. In the transfer learning process, AD audio data is segmented and truncated to meet Whisper’s input constraints while being fed into the model [16, 17]. A classification prefix is introduced as textual input to guide the model in generating diagnostic labels (“Normal” or “Alzheimer”). To mitigate information loss from audio truncation and ensure global context comprehension, full audio transcripts are provided as reference prompts during classification. This method achieves efficient AD classification without altering the original model architecture. Although initial validation focused on English data, Whisper’s inherent multilingual capabilities suggest strong potential for cross-linguistic AD detection.

In our study, we extend the Whisper transfer learning framework to multilingual settings, constructing a multilingual speech classification model for Alzheimer’s disease. For low-

*These authors contributed equally.

#Corresponding author.

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

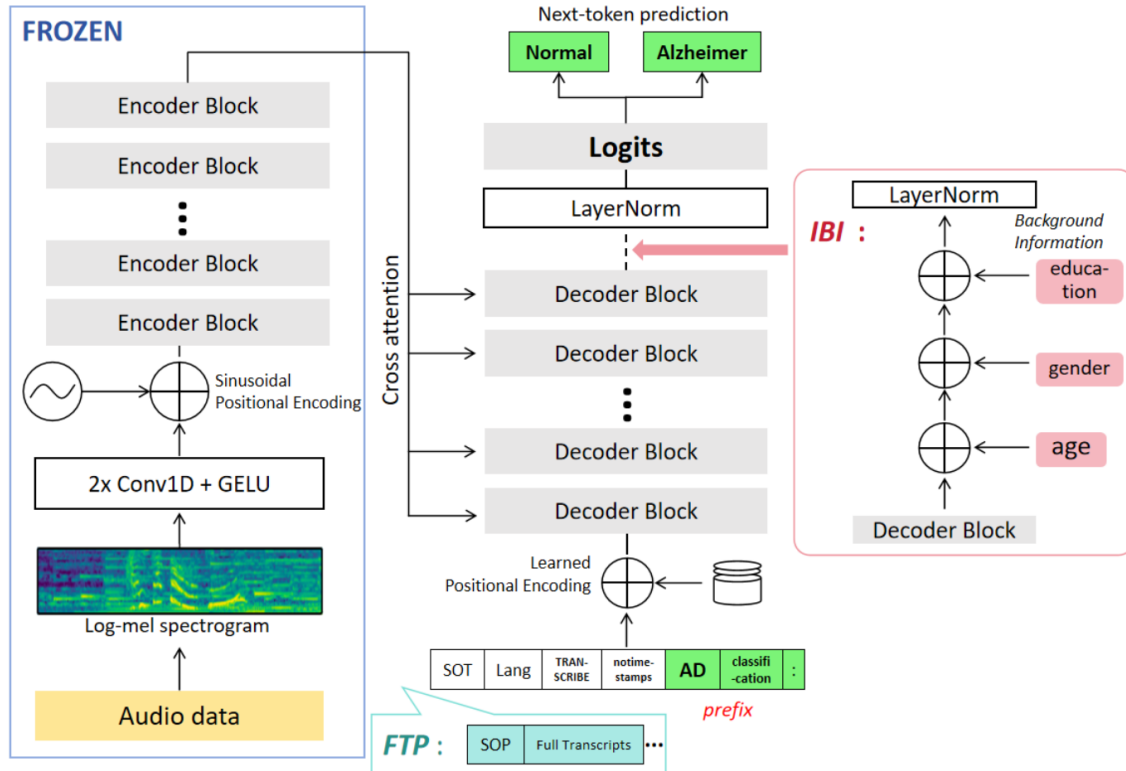


Figure 1: The overall framework of our research methodology. The overall structure is built on the basis of Whisper. The “FTP” part is to give the full transcripts as a prompt to the model sequence input. The “IBI” part is the specific process of integrating the background information into the decoder.

resource language scenarios, we further enhance performance by incorporating full audio transcripts and integrating speaker background information (e.g., age, gender, education level) into the decoding process. These innovations effectively address data scarcity challenges while improving diagnostic accuracy in linguistically diverse contexts.

2. Methods

2.1. Multilingual Joint Pre-Training (MJT)

Building on the principles of transfer learning with the Whisper model, we propose a multilingual joint training approach that leverages common audio characteristics across Alzheimer’s patients in different languages. This approach aims to enable the model to learn these shared features and effectively classify Alzheimer’s disease across various languages.

In this study, we utilize different Alzheimer’s disease audio datasets in three languages: English, Chinese, and Spanish. We apply Whisper transfer learning to develop a multilingual pre-trained model capable of efficiently processing Alzheimer’s data in these languages. We use audio data from different datasets as input and fine-tune the decoder while freezing the encoder parameters. Given the texts prefixed with “AD classification:”, the model is guided to ultimately generate English discriminative results of “Normal” or “Alzheimer”. This method is called Whisper-MJT. The model obtained by Whisper-MJT is termed to the multilingual pre-trained model and is used in subsequent experiments.

2.2. Low-Resource Adaptation (MJT-FT)

Existing multi-resource Alzheimer’s disease datasets are frequently limited to a few languages, resulting in sparse or absent AD audio data for many other languages. Traditional single-language AD detection methods often fall short in performance when applied to these low-resource datasets. To address this issue, we utilize the multilingual pre-trained model to aid in the detection of AD in languages with previously unavailable or limited data. By leveraging the general characteristics learned by the pre-trained model, we enhance diagnostic capabilities in low-resource languages.

For languages with minimal AD data, we perform data replication to match the amount of original training data. We fine-tune the pre-trained model by incorporating the available limited data. This augmentation process helps the model learn language-specific features more effectively. To mitigate overfitting risks associated with data repetition, we combine the augmented data with the original training data, aiming to capture both common and language-specific features.

Additionally, we observe that incorporating full transcripts as prompts (FTP) improves AD detection performance during single-language training on low-resource data. To further enhance the cross-lingual recognition capabilities of the hybrid model, we also include full transcripts as prompts in the training process. This method, which enhances AD detection in low-resource languages through additional fine-tuning, is designated as Whisper-MJT-FT. For scenarios involving zero-resource languages, employing the mixed pre-trained model di-

rectly demonstrates an effective improvement in detection performance.

2.3. Incorporating Background Information (IBI)

In the diagnostic process for Alzheimer’s Disease, background information about subjects is an important reference factor. Statistical data reveal that older individuals are more likely to develop the disease, and those with higher educational attainment are also at increased risk under comparable audio performance conditions. To improve diagnostic accuracy in analyzing AD audio data, we propose incorporating background information into the model. Building on the Whisper transfer learning framework, we integrate background information into the decoder input. During the decoding phase, this information is incorporated following the model’s self-attention mechanism. By introducing a linear layer approach, we sequentially give variables such as age, gender, and education level to the model decoder, combining this information with the output of the self-attention module, which together adjusts the logits distribution and ultimately produces a normalized decoded output. In this study, the method that integrates background information is referred to as Whisper-MJT-IBI. This approach is particularly beneficial for low-resource languages, where the lack of reference data can be addressed, thereby significantly enhancing AD detection performance.

3. Experiment

3.1. Datasets

For our multilingual Alzheimer’s disease classification task, we employ three audio datasets: ADReSSo, NCMMSC, and Ivanova. The ADReSSo dataset, provided by the ADReSSo Challenge [18], includes English recordings of both patients and healthy individuals describing a specific image. The NCMMSC dataset, sourced from the 2021 NCMMSC Alzheimer’s Disease Recognition Challenge¹, consists of extended Chinese audio recordings involving descriptions of an image [19]. The Ivanova dataset features Spanish-language recordings from patients with various cognitive states, primarily consisting of verbal renditions of excerpts from Don Quixote [20]. For low-resource language tasks, we utilized the ADReSS-M dataset from the ICASSP 2023 SPGC Challenge [13], which includes a dev set and a test set of Greek Alzheimer’s disease recordings.

3.2. Audio Preprocessing

In the AD datasets, most audio recordings exceed 30 seconds in length and do not meet the input limits of Whisper. Consequently, preprocessing is required to segment the audio into smaller fragments. Building upon the method outlined in [14], we retain the longer final segment (greater than 15 seconds) rather than discarding it entirely, to increase the number of training samples. Table 1 shows the information of each dataset, where “Training Samples” is the number of training samples before and after pre-processing. For ADReSS-M, which is a real low-resource case, we only include 14 audios from the dev set after processing for training. During the training process, in order to equalize the distribution of data from different languages, we repeat the data 40 times to achieve a sufficient dataset size. To enhance binary classification performance, all instances of Mild Cognitive Impairment (MCI) are reclassified as Alzheimer’s Disease (AD). All audio data are uniformly

¹<https://github.com/THUsatlab/AD2021>

downsampled to 16 kHz to meet input requirements.

Table 1: Summary of the datasets used in the study

Dataset	Language	Training Samples	Main Content
ADReSSo	English	237 → 571	Picture description
NCMMSC	Chinese	280 → 668	Picture description
Ivanova	Spanish	361 → 559	Story retelling
ADReSS-M	Greek	8 → 14 → 560	Picture description

3.3. Experimental Setup

In our experiment, we adhere to the settings of the Whisper transfer learning methodology. We utilize the Whisper-medium model as the base model. During training, we freeze the encoder and modify only the decoder. We use the cross-entropy loss function, which is optimized with the AdamW optimizer. The remaining hyperparameters are configured as follows: (epoch: 5, batch size: 1, learning rate: 0.0001, weight decay: 0.01, and Adam epsilon: 1.0e-8). To ensure the reliability and reproducibility of the results, we use a random seed method [21, 22]. During the data presentation phase, we aggregate results from 10 random seeds to determine the vote result and select the best result from the top-performing seeds in individual tests.

When we use full transcripts as a text prompt, due to length constraints imposed by the tokenizer, the prompt is truncated to retain only the final 335 tokens from the fully encoded transcript sequence. In the Whisper-MJT-IBI method, we incorporate background information as numerical input into the decoder. Specifically, age is input as a direct numeric value, gender is encoded with Male as 1 and Female as -1, and educational level is provided as a numeric input. For the small fraction of NA entries, we assume a placeholder value of 6.

All experiments are performed using a single NVIDIA GeForce RTX 3090 GPU (24GB memory) for both model training and evaluation. Following adequate training iterations until convergence is achieved, we employ the final converged checkpoint for comprehensive performance assessment.

4. Results

Table 2 presents the performance test results of the pre-trained model, obtained through joint training on the NCMMSC, ADReSSo, and Ivanova datasets across various language test sets. We display both the voting results and the best performance in different languages, as well as the detection results on mixed test sets. The results demonstrate that the model exhibits strong diagnostic capability across AD data in three different languages.

Table 3 presents the results of our experiments on AD detection using the low-resource ADReSS-M (Greek) dataset. We explore leveraging transfer learning with Whisper for the ADReSS-M data as a set of baseline controls. Introducing full transcripts as prompts in English leads to a 4.35% improvement in diagnostic accuracy. These results indicate that achieving accurate AD detection with a single-language low-resource dataset is challenging. When applying the multilingual pre-trained model directly to Greek AD diagnosis, the accuracy is equivalent to that achieved with the English FTP, showing a similar 4.35% improvement. Further fine-tuning the pre-trained model with the ADReSS-M dataset, augmented with English FTP, resulted in an overall accuracy increase of 8.72%. As illustrated in figure 2, this method not only significantly en-

Table 2: The performance of the multilingual pre-trained model on different test sets, the definitions of accuracy, precision, recall and F1 score can be found at the baseline paper [18]. Where “AD” stands for “Alzheimer” and “CN” stands for “Normal”.

Test Set		Accuracy(%)	Precision (%)		Recall(%)		F1 score (%)	
			AD	CN	AD	CN	AD	CN
ADReSSo	vote	74.65	71.79	78.12	80.00	69.44	75.67	73.52
	best	76.06	76.47	75.68	74.29	77.78	75.36	76.72
NCMMSC	vote	87.39	86.42	89.47	94.59	75.56	90.32	81.93
	best	88.24	87.50	89.74	94.59	77.78	90.91	83.33
Ivanova	vote	75.44	75.00	75.76	69.23	80.65	72.00	78.13
	best	80.70	80.00	81.25	76.92	83.87	78.43	82.54
mix-dataset	vote	80.57	80.42	80.77	85.19	75.00	82.74	77.78
	best	81.38	82.48	80.00	83.70	78.57	83.09	79.28

hances the pre-trained model’s diagnostic capability for low-resource AD data but also preserves the model’s inherent multilingual AD detection abilities. The final results demonstrate that integrating cross-linguistic features learned from the pre-trained model significantly aids in low-resource AD detection. .

Table 3: Comparison of diagnostic results of different methods on the ADReSS-M Greece dataset. Where “Whisper-TL” is an acronym for Whisper transfer learning

Method		Accuracy(%)	F1 score (%)	
			AD	CN
Whisper-TL	vote	52.17	66.67	15.38
	best	54.35	67.69	22.22
Whisper-TL (with FTP)	vote	56.52	54.55	58.33
	best	60.87	59.09	62.50
Whisper-MJT	vote	56.52	47.37	62.96
	best	58.70	57.78	59.57
Whisper-MJT-FT	vote	60.87	64.00	57.14
	best	67.39	65.12	69.39

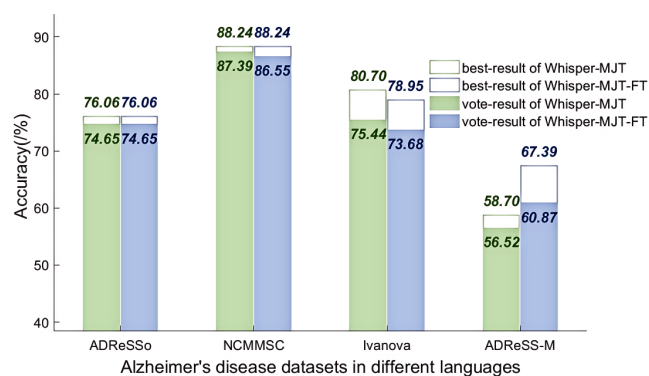


Figure 2: Performance comparison between the multilingual pre-trained model (Whisper-MJT) and its fine-tuned variant (Whisper-MJT-FT) cross four different language test sets.

We explore the enhancement of low-resource AD detection by incorporating background information into our pre-trained model. As demonstrated in Table 4, due to the limited data in

the ADReSS-M dataset, we integrate background information from the ADReSSo dataset and perform joint fine-tuning using English full transcripts as prompts.

Table 4: Comparison of results in the process of progressively incorporating background information using the Whisper-MJT-IBI method, with Whisper-MJT serving as the initial baseline data.

Method	+ AGE	+ GEN.	+ EDU.	Acc (%)	
				vote	best
Whisper-MJT	✗	✗	✗	56.52	58.70
Whisper-MJT-IBI	✗	✗	✗	54.35	63.04
	✓	✗	✗	58.70	67.39
	✓	✓	✗	65.22	69.57
	✓	✓	✓	67.39	71.74

We sequentially introduce background information as follows: incorporating age information alone results in a 2.18% improvement in accuracy; adding both age and gender information leads to an 8.70% increase; and including age, gender, and educational background information achieves a 13.04% enhancement in accuracy. These results indicate that integrating contextual information during the model decoding process significantly enhances diagnostic performance for low-resource language AD detection.

5. Conclusion

In this paper, we conduct an Alzheimer’s disease detection task based on the transfer learning of the multilingual speech recognition large model, Whisper. By guiding the model to capture the common audio features of spontaneous speech in Alzheimer’s patients across different languages, we have developed a multilingual pre-trained model capable of detecting Alzheimer’s disease in English, Chinese, and Spanish. For low-resource languages like Greek, we fine-tune the pre-trained model using a multilingual dataset that incorporates Greek data and combine full transcripts as prompts. This approach results in an accuracy improvement of 4-7% for Greek Alzheimer’s disease detection. Additionally, by integrating the background information of the subjects, we have increased the AD detection accuracy for Greek by 11-13%. We expect that this method can provide insights and assistance for the future establishment of robust multilingual Alzheimer’s disease detection systems.

6. References

- [1] S. Gauthier, C. Webster, S. Servaes, J. Morais, and P. Rosa-Neto, "World Alzheimer report 2022: Life after diagnosis: Navigating treatment, care and support," Alzheimer's Disease International, London, England, Tech. Rep., 2022.
- [2] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arighi, "Forecasting the global burden of Alzheimer's disease," *Alzheimer's & Dementia*, vol. 3, no. 3, pp. 186–191, 2007.
- [3] A. P. Porsteinsson, R. Isaacson, S. Knox, M. N. Sabbagh, and I. Rubino, "Diagnosis of early Alzheimer's disease: Clinical practice in 2021," *The Journal of Prevention of Alzheimer's Disease*, vol. 8, pp. 371–386, 2021.
- [4] 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.
- [5] Y. Wang, J. Deng, T. Wang, B. Zheng, S. Hu, X. Liu, and H. Meng, "Exploiting prompt learning with pre-trained language models for Alzheimer's disease detection," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023.
- [6] 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.
- [7] P. A. Pérez-Toro, S. P. Bayerl, T. Arias-Vergara, J. C. Vásquez-Correa, P. Klumpp, M. Schuster, E. Nöth, J. R. Orozco-Arroyave, and K. Riedhammer, "Influence of the interviewer on the automatic assessment of Alzheimer's disease in the context of the ADReSSo challenge," in *Proc. INTERSPEECH*, 2021, pp. 3785–3789.
- [8] Y.-W. Chien, S.-Y. Hong, W.-T. Cheah, L.-H. Yao, Y.-L. Chang, and L.-C. Fu, "An automatic assessment system for Alzheimer's disease based on speech using feature sequence generator and recurrent neural network," *Scientific Reports*, vol. 9, no. 1, 2019, art. no. 19597.
- [9] C. Sanz, F. Carrillo, A. Slachevsky, G. Forno, M. L. Gorno Tempini, R. Villagra, A. Ibáñez, E. Tagliazucchi, and A. M. García, "Automated text-level semantic markers of Alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 14, no. 1, 2022, art. no. e12276.
- [10] G. Gosztolya, V. Vincze, L. Tóth, M. Pákási, J. Kálmán, and I. Hoffmann, "Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181–197, 2019.
- [11] X. Chen, Y. Pu, J. Li, and W.-Q. Zhang, "Cross-lingual Alzheimer's disease detection based on paralinguistic and pre-trained features," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023.
- [12] P. A. Pérez-Toro, P. Klumpp, A. Hernandez, T. Arias, P. Lillo, A. Slachevsky, A. M. García, M. Schuster, A. K. Maier, E. Noeth *et al.*, "Alzheimer's detection from English to Spanish using acoustic and linguistic embeddings," in *Proc. INTERSPEECH*, 2022, pp. 2483–2487.
- [13] S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. MacWhinney, "Multilingual Alzheimer's dementia recognition through spontaneous speech: A signal processing grand challenge," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023.
- [14] 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.*, 2024, pp. 11 211–11 215.
- [15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Machine Learning*, vol. 202, 2023, pp. 28 492–28 518.
- [16] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-AT: Noise-robust automatic speech recognizers are also strong general audio event taggers," in *Proc. INTERSPEECH*, 2023, pp. 2798–2802.
- [17] M. Wang, Y. Li, J. Guo, X. Qiao, Z. Li, H. Shang, D. Wei, S. Tao, M. Zhang, and H. Yang, "WhiSLU: End-to-end spoken language understanding with Whisper," in *Proc. INTERSPEECH*, 2023, pp. 770–774.
- [18] 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*, 2021, pp. 3780–3784.
- [19] 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.
- [20] 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 & Language*, vol. 73, 2022, art. no. 101341.
- [21] 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.
- [22] 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.