# Alzheimer's Disease Detection using Pretrained Acoustic Representations: A Comparison Study

Jinchao Li, Xixin Wu, Xunying Liu, and Helen Meng

The Chinese University of Hong Kong, Hong Kong SAR, China
{jcli,wuxx,xyliu,hmmeng}@se.cuhk.edu.hk

**Abstract.** Alzheimer's disease (AD) is a progressive neurodegenerative disease that causes cognitive and physical impairment. Effective AD detection assists immensely in timely intervention and progression deceleration. Recently research on AD detection has made progress using pretrained acoustic representations. However, there is a lack of comparison of different representations. This paper conducts a comparison study on AD detection using several latest well-performed pretrained representations. We also investigate the performance variation by using different intermediate layers of the pretrained models. This study confirms the effectiveness of the pretrained representations for the AD detection task, and also has two interesting observations in the experiments: (i) higher layers of the pretrained model are more suitable for AD detection, and (ii) using one single layer outperforms a weighted sum of all layers. These observations appeal to further investigation into pretrained models specifically trained for AD detection.

**Keywords:** Alzheimer's Disease Detection · Acoustic Features · Self-supervised Learning.

## 1 Introduction

Alzheimer's disease (AD), characterized by a clear decline of cognitive functioning, including memory, language, thinking and behavior, is a major kind of neurocognitive disease (also called dementia) [1]. AD can intrigue irreversible deterioration of cognitive functioning, which doesn't have effective therapy nowadays. As estimated in 2019, AD affects over 50 million people globally [2]. Hence, the effective and early detection of AD is essential for timely intervention and decelerating disease progression.

Speech and language impairments are considered one of the most characteristic symptoms of AD at a very early stage, such as temporal disfluency, word finding and word retrieval difficulties [3–6]. This lays the theoretical foundation for using this acoustic and linguistic information to more robustly and more holistically detect and understand the Alzheimer's Disease, and it is attracting increasing research attention. Compared to conventional diagnosis methods, e.g., Magnetic Resonance Imaging and Fluid Analysis [7], spoken language-based diagnosis methods are cheaper, more convenient, and have a high potential for large-scale screening. Research efforts have been devoted to investigating AD detection using speech (the acoustic signal) and language (words and sentences) features as biomarkers [4–6]. Since obtaining the text from speech requires either costly human annotation or an automatic speech recognition (ASR) system, which may yield recognition errors that affect the final detection results, high-accuracy detection purely based on acoustic features is desired. Previous research has explored the utilization of conventional acoustic features, such as ComParE [8], eGeMAPS [9] and disfluency measures [10–12].

Recently, pretrained representations for speech have achieved significant successes in various spoken language processing areas [13, 14]. One representative approach is self-supervised learning (SSL). The idea behind SSL is to learn a representation of input text or speech data based on a self-supervised framework, e.g., predicting masked input tokens, using a large amount of data. The pretrained SSL models can then be transferred to downstream tasks, e.g., ASR [14], speaker verification [15], to boost the performance as the application of learnt representation implicitly augments the data for the downstream tasks. Another pioneering work by Radford *et al.* used a weakly-supervised approach to learn the prediction of audio transcripts based on a large amount of multilingual and multitask data (680,000 hours) on the internet [16]. This weakly-supervised framework also provides another option for extracting pretrained representations from speech.

The pretrained representations have also been successfully applied to audio-based AD detection [17–19]. For example, Koo *et al.* [17] used the averaged VGGish [20] features, which is trained for audio classification, and achieved better performance than traditional acoustic features. Balagopalan *et al.* [18] combined the conventional MFCC features with the pre-trained Wav2Vec 2.0. Syed *et al.* [19] used an ensemble method by fusing the embeddings of OpenL3-Music and Environment variants [21].

Though in the previous research [18, 20], self-supervised learned acoustic representations (e.g., VGGish and Wav2Vec 2.0) demonstrate superior performance than conventional acoustic features (e.g., the ComParE feature sets [10]), there is a lack of performance comparison of the various pretrained models in AD detection. Also, which layer(s) of the pretrained models is(are) more suitable for AD detection is unclear, though previous research suggests that higher layers contain more word and semantic information [22], which may be beneficial to the AD detection task [17–19]. In this paper, we try to investigate the effects of pretrained acoustic representations, including both SSL and weakly-supervised, by conducting a range of experiments using several advanced representations, including Wav2vec 2.0 [14], HuBERT [23], WavLM [24] and Whisper [16]. Through the experiments, we observe that the four SSL representations achieve comparable and good results, compared to the baselines. We further investigate the contribution of different layers from the pretrained models and found that higher layers generally provide better performance than lower layers, which coincides with previous research findings. Another interesting observation is that AD detection systems with one single upper layer outperform the systems with a weighted sum of all layers.

## 2 Methodology

We compare four SSL representations as introduced as follows. A neural model is built upon these representations to predict the disease as two classes, i.e., Alzheimer's disease (AD) and healthy control (HC).

### 2.1 Investigated Pretrained Models

**Wav2vec 2.0.** Baevski *et al.* proposed Wav2vec 2.0 to jointly learn contextualized speech representations and an inventory of discretized speech units [14]. The speech inputs are masked in the latent space and the model is trained with a contrastive loss defined over the discrete units. Compared to Wav2vec [25], where the prediction is the next step of a speech signal, Wav2vec 2.0 achieves significantly better performance with jointly learnt discretized speech units.
**HuBERT.** HuBERT introduces a prior lexicon based on offline clustering to provide labels for speech units [23]. The model is trained to predict the cluster assignments from the input speech units, which encourages the model to learn a combined acoustic and language model. The training and the clustering can be iterated to improve performance.
**WavLM.** In order to solve full-stack downstream speech tasks, WavLM jointly learns masked speech prediction and denoising, by using some simulated noisy or overlapped speech data [24]. The gated relative position bias is utilized to better capture the sequence ordering of input speech. With these improvements, WavLM is effective for not only the ASR task, but also speaker-related tasks, e.g., speaker diarization [26].
**Whisper.** Radford *et al.* [16] proposed a robust ASR system, named Whisper, trained in a weak supervision fashion, i.e., the multilingual and multitask training data contains noisy transcriptions. However, they vastly scaled the data to 680,000 hours. Whisper is competitive with other fully supervised ASR systems on several common benchmarks and obtains stable performances in zero-shot settings without any fine-tuning, demonstrating its strong robustness. We propose to use the intermediate layers of Whisper as pretrained representations for AD detection and compare them to other SSL representations.

### 2.2 Alzheimer's Disease Detection Model

As illustrated in Fig. 1, we proposed a neural architecture to extract features based on the pretrained representations and elaborately use them for the AD detection task. All of the aforementioned pretrained models are generally composed of a feature encoder, projector, and several stacked Transformer encoders (TE). The embeddings from different TE layers usually contain different information, i.e., lower layers encode local acoustic features, followed by phonetic,

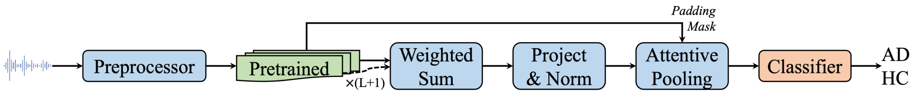


**Fig. 1.** Proposed model architecture.

word identity, and word meaning information [22, 27]. To avoid information loss due to only using the last layer of the pretrained TEs, we utilized all layers plus the input layer using a weighted sum of these layers with trainable weights. The weighted-summed embedding is then projected into a lower-dimension vector and normalized to reduce redundancy while retaining the intra-class variability. The projection and normalization is followed by an attentive time-axis pooling layer [28] to compress the sequence with variant time lengths into a fixed-length vector, and to focus on important time steps for the target tasks. Finally, we used a simple linear layer with the softmax activation to predict the diagnosis results of AD or HC.

## 3 Experiments

### 3.1 Dataset

We used the Alzheimer's Dementia Recognition Through Spontaneous Speech (ADReSS) Challenge 2020 dataset [10], which is a selected part of Pitt Corpus in the DementiaBank database [29]. The dataset consists of 156 speech samples and associated transcripts from non-AD (35 male, 43 female) and AD (35 male, 43 female) English-speaking participants for the Cookie Theft picture description task, and is divided into a standard train (108 participants, about 2 hours) and test (48 participants, about 1 hour) sets with balanced distributions of age, gender and disease conditions.

### 3.2 Experimental Setup

In the experiments, we first enhanced the audios with the FullSubNet toolkit [30]. Then, we sliced each audio into 30-second segments with a hop ratio of 0.25, and averaged the predicted detection probabilities for the same speaker. The models are evaluated on the unseen ADReSS test data with the metrics of classification accuracy and F1 (macro) scores.

The variants of Wav2vec 2.0, Hubert, and WavLM used in this work are the "large" ones, which output 1024-dimensional embeddings. We used the Whisper variant of "small.en", which outputs 768-dimensional embeddings. Upon the extracted pretrained representations, we used two stacked 8-dimensional linear layers with layer normalization each as the projector and a fully-connected linear layer as the classifier. We inserted dropout layers with a rate of 0.25 before the projector and the classifier, respectively, for regularization. We explored two ways of aggregating information from the layers of pretrained models, i.e., "weighted sum (WS)" and "maximum single (MS)". WS obtains a weighted sum of all the pretrained layers, and MS selects one single layer with the best performance. We also compared the attentive temporal pooling with mean values across time steps (denoted as "Mean" in Table 1). Moreover, we compared our proposed models with previous literature on the same ADReSS dataset, including models based on conventional acoustic features [10], VGGish embeddings [17] and fusion of OpenL3 embeddings [19].

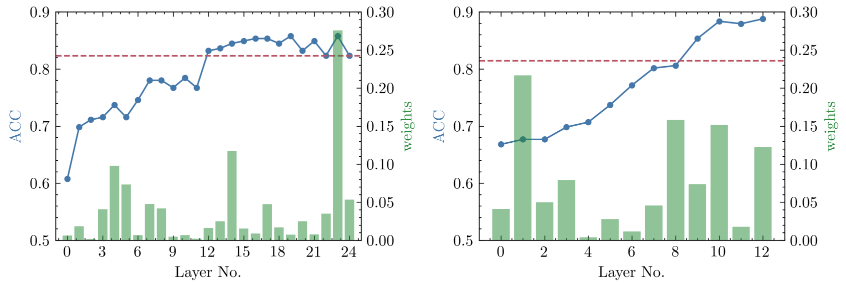
We adopted a cross-entropy loss as the training objective. The optimizer we used is the AdamW [31] with a weight decay of $1e-5$. In the training process, we froze the pretrained models and only trained the subsequent modules with a learning rate of $1e-4$ and a batch size of 16 for 50 epochs.

**Table 1.** Comparison of AD detection performance based on acoustic features and various representations. "AGG" denotes "aggregation", "WS" denotes "weighted sum", and "MS" denotes "maximum single".

| Feature | Layer AGG | Time AGG | Accuracy(%) | F1-score(%) |
|---|---|---|---|---|
| ComParE [10] | N/A | Mean | 62 | 62 |
| VGGish [17] | Top | Mean | 72.92 | 72.62 |
| OpenL3 [19] | Top | Mean | 81.25 | 81.20 |
| Wav2vec 2.0 | WS | Mean | 77.50 | 76.69 |
| HuBERT | WS | Mean | 78.88 | 78.79 |
| WavLM | WS | Mean | 79.74 | 79.66 |
| Whisper | WS | Mean | 79.31 | 79.30 |
| WavLM | WS | Attention | 82.33 | 82.33 |
| Whisper | WS | Attention | 81.47 | 81.46 |
| WavLM | MS | Attention | 85.78 | 85.78 |
| Whisper | MS | Attention | **88.79** | **88.79** |

### 3.3 Experimental Results

We compared acoustic representations from various pretrained models, including SSL models of Wav2vec 2.0, HuBERT, WavLM and the latest weakly-supervised model Whisper, as shown in Table 1. The performances are benchmarked with the baseline systems. It can be found that the systems that use weighted sum and mean pooling based on pretrained acoustic representations achieve much better performances than the baseline system with the ComParE features [10]. The four pretrained representations obtain comparable performance, with WavLM and Whisper achieving slightly better performances. Compared with the mean pooling strategy, the Attentive Pooling boosts the performances of systems based on WavLM and Whisper by a significant margin. These experimental results confirm the effectiveness of pretrained representations for the AD detection task.



(a) WavLM

(b) Whisper

**Fig. 2.** Effectiveness of various layers of pretrained models, (a) WavLM and (b) Whisper. The blue line shows the performance of the systems using only one single layer of the pretrained models. The green bars represent the attention weights assigned to each pretrained layer in the systems using all pretrained layers. The red dashed line denotes the performances of the systems using a weighted sum of all pretrained layers.

We chose WavLM and Whisper, as representatives for the two frameworks of SSL and weakly-supervised, to further investigate the effectiveness of each layer of the pretrained models. We built systems using only one single layer from the pretrained models. The performances are shown as the blue lines in Fig. 2. The systems using a weighted sum of all layers are presented as a dashed red line, and the weights are shown as green bars. It can be observed that, (i) systems using higher single layers generally achieve better performances than those using lower layers; (ii) some systems using one single upper layer can outperform the system using a weighted sum of all layers; and (iii) the weight distributions do not match the performances of systems using one single layer. The first observation coincides with previous research that higher layers capture more word and semantic information [22], which are crucial for AD detection. This also supports the way of using the topmost layer for AD detection that is widely adopted in previous research [17–19]. Observations (ii) and (iii) together lead to an inference that due to data sparsity of the AD detection task, the attention weights are not sufficiently trained, therefore the systems using all layers are inferior to some well-performing systems using one single upper layer.

## 4 Conclusion

Alzheimer's disease detection is attracting increasing attention, and many efforts have been devoted to making use of pretrained acoustic representations for better detection results. However, there is a lack of comparison of various pretrained representations. Also, which layer(s) of pretrained models is(are) more suitable for AD detection is unclear. In this work, we conducted a comparison study on the latest advanced pretrained representations, including Wav2vec 2.0, HuBERT, WavLM and Whisper, for AD detection. Through the experiments, we found that the four examined representations provide comparable results, with WavLM and

Whisper obtaining slightly better results. We also found that higher layers of the pretrained models are more suitable for AD detection. In the future, we will investigate models specifically pretrained for AD detection.

## 5 Acknowledgements

This project is partially supported by the HKSARG Research Grants Council's Theme-based Research Grant Scheme (Project No. T45-407/19N).

## References

1. P. S. Sachdev, D. Blacker, D. G. Blazer, et al., "Classifying neurocognitive disorders: the dsm-5 approach," *Nature Reviews Neurology*, vol. 10, no. 11, pp. 634–642, 2014.
2. C. Lynch, "World alzheimer report 2019: Attitudes to dementia, a global survey: Public health: Engaging people in adrd research," *Alzheimer's & Dementia*, vol. 16, pp. e038255, 2020.
3. G. Szatloczki, I. Hoffmann, V. Vincze, et al., "Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease," *Frontiers in aging neuroscience*, vol. 7, pp. 195, 2015.
4. K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.
5. J. Weiner, C. Frankenberg, J. Schröder, and T. Schultz, "Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 674–681.
6. M. L. B. Pulido, J. B. A. Hernández, M. Á. F. Ballester, et al., "Alzheimer's disease and automatic speech analysis: a review," *Expert systems with applications*, vol. 150, pp. 113213, 2020.
7. G. Gainotti, D. Quaranta, M. G. Vita, and C. Marra, "Neuropsychological predictors of conversion from mild cognitive impairment to alzheimer's disease," *Journal of Alzheimer's disease*, vol. 38, no. 3, pp. 481–495, 2014.
8. B. Schuller, S. Steidl, A. Batliner, et al., "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *INTERSPEECH*, 2013.
9. F. Eyben, K. R. Scherer, B. W. Schuller, et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.
10. S. Luz, F. Haider, S. de la Fuente, et al., "Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge," *Proc. Interspeech 2020*, pp. 2172–2176, 2020.
11. S. Luz, F. Haider, S. de la Fuente, et al., "Detecting cognitive decline using speech only: The adresso challenge," in *INTERSPEECH 2021*. ISCA, 2021.
12. 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 *INTERSPEECH*, 2021, vol. 22, pp. 3805–3809.

13. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
14. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.
15. Z. Chen, S. Chen, Y. Wu, et al., “Large-scale self-supervised speech representation learning for automatic speaker verification,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6147–6151.
16. A. Radford, J. W. Kim, T. Xu, et al., “Robust speech recognition via large-scale weak supervision,” Tech. Rep., Technical report, OpenAI, 2022. URL https://cdn. openai. com/papers/whisper. pdf, 2022.
17. J. Koo, J. H. Lee, J. Pyo, et al., “Exploiting multi-modal features from pre-trained networks for alzheimer’s dementia recognition,” in *INTERSPEECH*, 2020.
18. A. Balagopalan and J. Novikova, “Comparing acoustic-based approaches for alzheimer’s disease detection,” *arXiv preprint arXiv:2106.01555*, 2021.
19. Z. S. Syed, M. S. S. Syed, M. Lech, and E. Pirogova, “Automated recognition of alzheimer’s dementia using bag-of-deep-features and model ensembling,” *IEEE Access*, vol. 9, pp. 88377–88390, 2021.
20. S. Hershey, S. Chaudhuri, D. P. Ellis, et al., “CNN architectures for large-scale audio classification,” in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.
21. J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, “Look, listen, and learn more: Design choices for deep audio embeddings,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.
22. A. Pasad, J.-C. Chou, and K. Livescu, “Layer-wise analysis of a self-supervised speech representation model,” in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.
23. W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, et al., “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
24. S. Chen, C. Wang, Z. Chen, et al., “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, 2022.
25. S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition.,” in *INTERSPEECH*, 2019.
26. S. Horiguchi, Y. Fujita, S. Watanabe, et al., “End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors,” *Proc. Interspeech 2020*, pp. 269–273, 2020.
27. H.-S. Choi, J. Lee, W. Kim, et al., “Neural analysis and synthesis: Reconstructing speech from self-supervised representations,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 16251–16265, 2021.
28. C. d. Santos, M. Tan, B. Xiang, and B. Zhou, “Attentive pooling networks,” *arXiv preprint arXiv:1602.03609*, 2016.
29. J. T. Becker, F. Boiler, O. L. Lopez, et al., “The natural history of alzheimer’s disease: description of study cohort and accuracy of diagnosis,” *Archives of neurology*, vol. 51, no. 6, pp. 585–594, 1994.

30. X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6633–6637.
31. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.