LEVERAGING PRETRAINED REPRESENTATIONS WITH TASK-RELATED KEYWORDS FOR ALZHEIMER'S DISEASE DETECTION

Jinchao Li[1], *Kaitao Song*[2], *Junan Li*[1], *Bo Zheng*[1], *Dongsheng Li*[2], *Xixin Wu*[1], *Xunying Liu*[1], *Helen Meng*[1]

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Microsoft Research Asia, Shanghai, China
[1]{jcli,jli,bzheng,wuxx,xyliu,hmmeng}@se.cuhk.edu.hk, [2]{kaitaosong,Dongsheng.Li}@microsoft.com

ABSTRACT

With the global population aging rapidly, Alzheimer's disease (AD) is particularly prominent in older adults, which has an insidious onset and leads to a gradual, irreversible deterioration in cognitive domains (memory, communication, etc.). Speech-based AD detection opens up the possibility of widespread screening and timely disease intervention. Recent advances in pre-trained models motivate AD detection modeling to shift from low-level features to high-level representations. This paper presents several efficient methods to extract better AD-related cues from high-level acoustic and linguistic features. Based on these features, the paper also proposes a novel task-oriented approach by modeling the relationship between the participants' description and the cognitive task. Experiments are carried out on the ADReSS dataset in a binary classification setup, and models are evaluated on the unseen test set. Results and comparison with recent literature demonstrate the efficiency and superior performance of proposed acoustic, linguistic and task-oriented methods. The findings also show the importance of semantic and syntactic information, and feasibility of automation and generalization with the promising audio-only and task-oriented methods for the AD detection task.

Index Terms— Alzheimer's disease, task-oriented, pretrained embeddings, transfer learning, multimodality

1. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disease that causes gradual, irreversible deterioration in cognitive domains (memory, communication, etc.). Since there is no effective treatment for AD currently, early detection of this disease is particularly crucial for timely intervention and better disease control [1,2]. Previous studies [3–6] have shown that symptoms of AD may be observable in spoken language at a very early stage, such as temporal disfluency, and difficulties in word finding and retrieval. These studies lay the theoretical foundation for using acoustic and linguistic information to screen for AD, which is attracting increasing interest from international research community.

The use of spoken language to screen for AD offers the advantages of affordability, accessibility and hence scalability, compared to conventional methods such as brain scans, blood tests and face-to-face neuropsychological assessments [7]. Active research efforts are being devoted to finding spoken language features (both audio and linguistic features) as biomarkers of AD [8–11]. For example, Weiner and Frankenberg *et al.* [9, 11] explored many traditional acoustic and linguistic features with a nested forward feature selection method, and found that the features on parts-of-speech, word categories and pauses are highly related to AD. Hence, possible approaches to design advanced algorithms to extract powerful acoustic and linguistic features to diagnose Alzheimer's disease have become an emerging topic.

Inspired by the success of pretrained models, especially in speech (e.g., VGGish [12], Wav2Vec 2.0 [13], OpenL3 [14]) and text (e.g., BERT [15], ERNIE [16], Glove [17]), the development of AD detection is shifting from low-level features to higher-level representations in pretrained models. Although the BERT-like models have achieved promising performance on the AD detection task [15, 16, 18, 19], we note that these works mainly used higher-level representations of pretrained models, while features from intermediate layers may not have been devoted sufficient attention. It is shown that intermediate layers encode rich hierarchical information for various features, e.g., surface features at the bottom, syntactic features in the middle and semantic features at the top [20]. Therefore, exploring and leveraging different levels of pretrained representations for better AD detection task is worthy of investigation.

The acoustic and linguistic features are typically used to distinguish the people with Alzheimer's disease from healthy controls, modeling the richness or disorder of participants' utterances to some extent. In addition to these features, it is also important to model the correctness and pertinence of participants' utterances for the cognitive tasks. For example, in the widely used Cookie Theft Picture Description task, where people are asked to describe everything happening in a picture, Laura *et al.* proposed information coverage measured by the statistics of the text and predefined referent [21]. This motivates us to propose features for modeling both the characteristics of disorder and the pertinence to the cognitive tasks.

In this work, we propose several efficient strategies to extract AD-related cues from embeddings of pretrained models, including aggregation along the layer and time dimension. We also use these extracted embeddings to measure task-related pertinence by correlation operation. To validate the effectiveness of our proposed method, we conduct experiments on the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) dataset [22] with a binary classification setup. The proposed models are evaluated with accuracy and F1 scores, and obtain comparable results over other state-of-the-art methods in recent literature. It's also exciting to find that the acoustic representations are comparable to the linguistic features, which could be more robust and generalizable in multilingual tasks and more helpful for fully automatic AD detection tasks.

The rest of the paper is organized as follows. Section 2 introduces the proposed methods for modeling with acoustic, linguistic and visual embeddings. Then, Section 3 describes the experimental setups and results, as well as further analysis of the proposed meth-

[1]The work was done during the author's internship with Microsoft.

ods and comparison with recent literature. Finally, Section 4 concludes the paper and presents possible future research directions.

2. METHODOLOGY

This section presents our methodology. Firstly, we describe the high-level embeddings for the audio and the text modalities. Then we introduce the measurement of task-related pertinence using these embeddings. And finally, we illustrate the feature extractor and classifier for the AD detection task.

2.1. Acoustic & Linguistic Embeddings

To obtain rich characteristics for the AD detection, we adopt different pretrained models for speech and text to extract acoustic and linguistic features respectively.

For the acoustic features, we compared several self-supervised (SSL) pretrained models (e.g., Wav2Vec 2.0 [23], HuBERT [24], WavLM [25]), and a weakly-supervised Whisper model [26]. The Wav2vec 2.0 is a model that jointly learns contextualized speech representations and an inventory of discretized speech units. The HuBERT introduces a prior lexicon based on offline clustering to provide pseudo labels for speech units. 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. 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 [25]. The gated relative position bias is utilized to capture the sequence ordering of input speech better. With these improvements, WavLM is effective for not only the ASR task, but also several downstream tasks [27].

However, the speech of individuals with AD may differ from that of a general speaker, which may affect the performance of the self-supervised pretrained models on AD detection. A recent work, named Whisper [26], has achieved state-of-the-art performance on many ASR tasks, which is trained on a web-scale 680,000 hours in a weakly supervised multilingual multitask fashion. Therefore, compared with previous models, Whisper is more advantageous and robust for application to different downstream tasks. Considering these merits of Whisper in processing speech tasks, we adopt it as the default acoustic feature extractor. For linguistic features, we choose BERT [28], a Transformer-based [29] pretrained model, as the basic backbone network, which adopts masked language modeling and next sentence prediction as the pre-training objective.

Whisper calculates the logarithm Mel-Spectrum and then connects with two Convolution layers, followed by the Transformer layers. We feed audio into Whisper and use its produced features as acoustic representation. So for each second of audio, we can obtain features from different Transformer layers, which have a dimension of $50 \times H_a \times L_a$ (dimensions of time, feature and layer respectively). And for the text modality, we feed the transcribed text from the audios of patients into BERT model to obtain the features as $1 \times H_t \times L_t$ of each token.

2.2. Task-related Information

For the Cookie Theft Picture Description Task, we aim to integrate information from the picture using a series of pre-defined keywords, as shown in Fig. 1. The keywords include the named entities (nouns) and actions (verbs) happening in the picture. We calculate the correlation between the linguistic embeddings of the spoken utterances (describing the picture) and the pre-defined keywords for the picture. Formally, let the extracted embeddings of spoken utterances and task-related keywords be z_u and z_k, then the task-related correlation is $Corr. = z_u \times z_k$ by element-wise production.

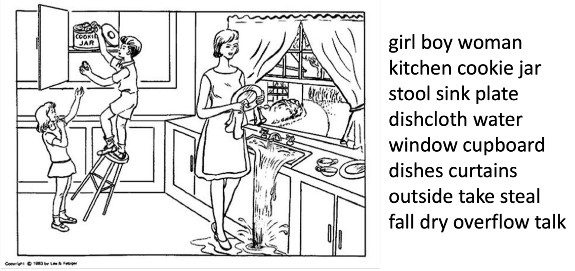

Fig. 1. The Cookie Theft Picture and pre-defined keywords.

2.3. Feature Extractor

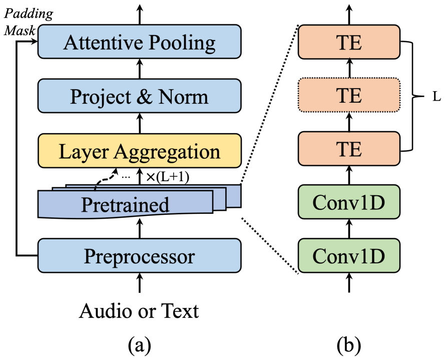

Fig. 2. Feature extraction with pretrained models. "TE" denotes the Transformer Encoder, "L" denotes the number of TE layers. Compared to BERT, Whisper has two Conv1D before the stack of TEs.

As mentioned earlier, the raw embeddings extracted from the Whisper or BERT for the two modalities are large and redundant. To further condense the information which should be helpful for the AD detection task, we propose several subsequent modules as illustrated in Fig. 2. First, we perform layer aggregation to calibrate layer-wise feature responses. Here, we adopt two strategies: the first is "weighted sum (WS)", which sums up all elements with learnable weights; the second is "maximum single (MS)", which selects one single layer with the best performance. Second, we project the calibrated features into a lower dimensionality to reduce feature redundancy while retaining intra-class variability. The projector is an MLP with Layer-Normalization, which can also map audio and text modalities into the same dimensional space for fusion. Third, we use an attentive temporal pooling layer [30] to compress the sequence with variant time lengths into a fixed-length vector, and capture richer statistics of the temporal features for the AD detection task. Finally, we average the features from different segments for each speaker.

2.4. Classifier

The extracted features are either fed into a classifier directly or combined with others with concatenation operation. The classifier used in this work is a fully-connected layer that produces the probabilities of binary classification between an individual having AD or being a Healthy Control (HC).

3. EXPERIMENTS

In this section, we describe the corpus, experimental setup and results.

3.1. Corpus

The corpus used in this work comes from the Alzheimer's Dementia Recognition Through Spontaneous Speech (ADReSS) Challenge 2020 corpus [22]. This challenge selects a sub-task of Pitt Corpus in the DementiaBank database [31], which requires all the participants to describe the Cookie Theft picture as shown in Fig. 1. The ADReSS corpus consists of 156 different English speakers' audio samples with corresponding transcripts. Among them, 78 of the speakers are healthy control (35 male, 43 female) while the rest are with AD (35 male, 43 female). The corpus is divided into a standard train (108 speakers, about 2 hours) and test (48 speakers, about 1 hour) sets with balanced distributions of age, gender and disease conditions.

3.2. Experimental Setup

3.2.1. Data Preprocessing

At the beginning of the experiments, we preprocess the speech with enhancement [32] and normalization methods for internal consistency of the data. In addition, we use data augmentation to enrich the data and improve the robustness by slightly changing the acoustic characteristics with minor distortions. Specifically, we used three strategies, pitch-shifting, speed-perturbation, and dithering for each input waveform during the training stage [33–35]. The shifted ranges of pitch and speed are [-100, 100] semitones and [-0.05, +0.05] rates, respectively.

Notably, the data on AD investigated in this work are in long-form, e.g., several minutes of spontaneous speech associated with transcripts from the picture description task. However, the inputs of the high-level pretrained models are usually within 30 seconds for audio or 512 tokens for text. To deal with this issue, we slice the experimental audio into 30-second segments with a hop ratio of 0.25, and obtain the aligned transcripts that are less than 512 tokens. We then aggregate the extracted segment-level features for each speaker.

3.2.2. Model & Training Details

The variants of the Whisper and BERT models adopted in this work are respectively the small and base-uncased versions pretrained on English corpora , both of which output 768-dimensional embeddings. The projector is composed of two stacked 8-dimensional linear layers with layer normalization, and the classifier is a fully-connected linear classifier. We insert a dropout layer with a rate of 0.25 between the projector and temporal pooling layer.

The training loss of this work is set to be the binary cross-entropy loss. We use AdamW [36] as our optimizer with a learning rate of $1e-4$ and a weight decay of $1e-5$. The models are trained with a batch size of 16 for 50 epochs.

3.2.3. Evaluation Protocols

We evaluate the model performance on the unseen ADReSS test data, with the metrics of classification accuracy and macro F1 scores, the mean of class-wise F1-scores, that averaged on five random runs. To better compare our proposed methods with previous literature, we choose multiple baselines, including Luz *et al.* which used ComParE and Linguistic feature sets [22], Koo [12] and Syed *et al.* [14] which used acoustic pretrained models, and Yuan [16], Matej [17] and Yi *et al.* [19] which used linguistic pretrained models.

3.3. Results

3.3.1. Results using Extracted Embeddings

We compare the performance of acoustic and linguistic features with different aggregation strategies on the layer and time dimensions, including weighted-sum (WS) or single selected (Top, MS) for the layers, and mean (Mean) or attentive (Attention) pooling for the time-axis, as listed in the Table. 1.

Feature	Layer AGG	Time AGG	Accuracy(%)	F1-score(%)
ComParE [22]	-	Mean	62	62
Linguistics [22]	-	Mean	75	71
VGGish [12]	Top	Mean	72.92	72.62
OpenL3 [14]*	Top	Mean	81.25	81.20
ERNIE [16]	Top	Mean	85.4	85.3
Glove [17]	Top	Mean	89.6	-
BERT+Roberta [19]*	Top	Mean	**91.7**	**91.7**
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
BERT	WS	Mean	87.50	87.47
WavLM	WS	Attention	82.33	82.33
Whisper	WS	Attention	81.47	81.46
BERT	WS	Attention	88.79	88.76
WavLM	MS	Attention	85.78	85.78
Whisper	MS	Attention	**88.79**	**88.79**
BERT	MS	Attention	**90.09**	**90.07**

Table 1. Performance based on acoustic features or representations. "AGG" denotes "aggregation", "WS" denotes "weighted sum", and "MS" denotes "maximum single". "$\star$" denotes ensemble systems.

It can be observed that, (i) the best single selected method outperforms the weighted-sum method for layer-wise representations; (ii) the attentive pooling method outperforms the mean pooling for the time-axis information; (iii) Whisper outperforms other acoustic models with the MS and attentive pooling methods; and (iv) linguistic representations generally outperform the acoustic representations.

For the first observation, we find that the weights in the WS method are not well-matched with the performance distributions using a single layer. For example, as shown in Fig. 3, the topmost layers (larger layer No.) of Whisper and the middle layers of BERT show higher performance for the AD detection task, but the learned weights may also be distracted by the bottom layers in the WS method, which harms the performance. We study the performance on each layer in depth. For the acoustic models, the observation of that the topmost layers outperforms others coincides with previous research that higher layers capture more word and semantic information [37], 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 [12, 14, 38]. For the linguistic models, the observation of that middle layers outperforms others implies that the syntactic information is more important than the semantic information [20] for the AD detection, which is also intuitive, since the syntactic information can model the cognitive disorder better.

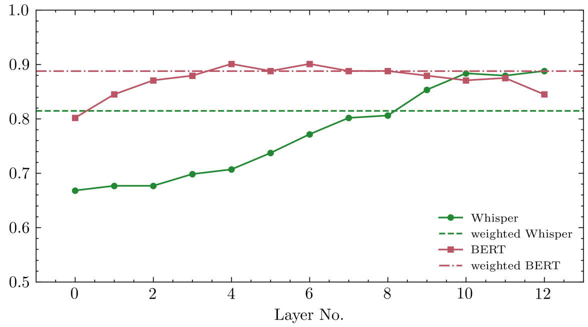

Fig. 3. Effectiveness of various layers of pretrained Whisper and BERT. The solid and dashed lines denote systems using only one single layer or a weighted sum of all layers respectively.

The second observation supports that the attentive pooling method captures richer statistics of temporal features than the mean pooling method. The first and second observations also show that the layer and time dimensions of pretrained models have different importance in the AD detection task. Take the Whisper model as an example, the Attention-based time aggregation improves the accuracy scores by about 2.7% relatively, while the MS layer aggregation is more effective and further improves it by about 9% relatively.

The third observation reflects the robustness and effectiveness of the Whisper model for AD detection. And the forth observation coincides with previous research [10, 18]. It is also interesting to find that the performance of acoustic models is now comparable with that of linguistic ones, which worse a lot than the latter in the past. The underlying mechanism of pretrained encoders is difficult to interpret, but we could intuitively explain the finding in terms of representation and pathology. On the one hand, the high-level pretrained acoustic encoders could extract the semantic information, especially from the top layers, that is similar to the linguistic features and helpful for the AD detection task. On the other hand, AD affects the participants' phonology and articulation [5, 6], such as dysfluencies (aphasic to some degree) and hesitations, while acoustic encoders could also extract these features that may not be easily extracted from the text. These encoded semantic and acoustic features could make acoustic models comparable to linguistic methods in the AD detection task. The promising performance of acoustic models not only promotes the fully-automation of the AD detection task, but also could be helpful for multilingual generalization since some acoustic characteristics are more "universal" across languages than linguistic ones.

3.3.2. Results using Task-related Information

We also compare the task-related correlation features with different keyword lists, including "None" (empty), "Nouns"-only (named entities), "Verbs"-only (actions) and combination of "Nouns" and "Verbs". It can be found that using keywords of "Nouns" performs better than using that of "Verbs" for AD detection task, which implies the ability of named entities retrieval are affected by Alzheimer's Disease and coincides with the fact that Named Task is important in the clinical cognitive tasks. We also decouple the effect of task-related correlation by using an empty keyword list, and the performance drops to 52.34%, which can be view as a random guess. It can be also observed that using correlation with task-related keywords can achieve over 80% accuracy scores, which is better than the linguistic measures in [22].

Type	Accuracy(%)	F1-score(%)
None	52.34	34.36
Nouns	85.16	85.11
Verbs	83.59	83.58
Nouns + Verbs	**85.94**	**85.88**

Table 2. Results using correlation features between text and keywords of "None" (empty), "Nouns", "Verbs" and both.

3.3.3. Feature Combination Results

Finally, we compare the performance of a combination of the acoustic (Whisper), linguistic (BERT) and task-related correlation (Corr.) features, as listed in Table 3. Generally, the combination of different modalities outperforms the other systems with a superior performance of 91.41% accuracy, which implies that complementary information from various modalities helps AD detection.

Feature	Accuracy(%)	F1-score(%)
OpenL3 + Roberta [14]⋆	89.85	89.54
Temporal + Glove [17]	91.67	-
Whisper	88.79	88.79
BERT	90.09	90.07
Corr.	85.94	85.88
Whisper + BERT	91.19	91.19
Whisper + Corr.	89.84	89.83
BERT + Corr.	90.62	90.60
Whisper + BERT + Corr.	**91.41**	**91.38**

Table 3. Results on a combination of features. "Corr." denotes correlation embeddings between utterance and keywords from picture. "⋆" denotes ensemble systems.

4. CONCLUSION

In this work, we explored pretrained representations from different modalities for Alzheimer's Disease detection. Experiments on the ADReSS corpus have shown superior performance by using acoustic and linguistic embeddings, as well as task-related keywords. Results indicate that the top layers of pretrained acoustic models and the middle layers of pretrained linguistic models provide features that are more important for the AD detection task. It also shows that the correlation and richness measured by task-related keywords and described utterances could also help the AD detection task.

Future work will include multimodal pretrained embeddings to model visual-textual relationships, more efficient fusion strategies to boost performance, as well as using acoustic embeddings for automatic and multilingual AD detection tasks.

5. ACKNOWLEDGEMENTS

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

6. REFERENCES

[1] S. G. Mueller, M. W. Weiner, et al., "Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni)," *Alzheimer's & Dementia*, 2005.

[2] J. Rasmussen and H. Langerman, "Alzheimer's disease–why we need early diagnosis," *Degenerative neurological and neuromuscular disease*, 2019.

[3] J. Appell, A. Kertesz, and M. Fisman, "A study of language functioning in alzheimer patients," *Brain and language*, 1982.

[4] J. L. Cummings, A. Darkins, et al., "Alzheimer's disease and parkinson's disease: comparison of speech and language alterations," *Neurology*, 1988.

[5] K. Croot, J. R. Hodges, et al., "Phonological and articulatory impairment in alzheimer's disease: a case series," *Brain and language*, 2000.

[6] F. Gayraud, H.-R. Lee, and M. Barkat-Defradas, "Syntactic and lexical context of pauses and hesitations in the discourse of alzheimer patients and healthy elderly subjects," *Clinical linguistics & phonetics*, 2011.

[7] G. Gainotti, D. Quaranta, et al., "Neuropsychological predictors of conversion from mild cognitive impairment to alzheimer's disease," *Journal of Alzheimer's disease*, 2014.

[8] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, 2016.

[9] J. Weiner, C. Frankenberg, et al., "Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews," in *ASRU*. IEEE, 2019.

[10] M. L. B. Pulido, J. B. A. Hernández, et al., "Alzheimer's disease and automatic speech analysis: a review," *Expert systems with applications*, 2020.

[11] C. Frankenberg, J. Weiner, et al., "Verbal fluency in normal aging and cognitive decline: Results of a longitudinal study," *Computer Speech & Language*, 2021.

[12] J. Koo, J. H. Lee, et al., "Exploiting multi-modal features from pre-trained networks for alzheimer's dementia recognition," in *INTERSPEECH*, 2020.

[13] A. Balagopalan and J. Novikova, "Comparing acoustic-based approaches for alzheimer's disease detection," *arXiv preprint arXiv:2106.01555*, 2021.

[14] Z. S. Syed, M. S. S. Syed, et al., "Automated recognition of alzheimer's dementia using bag-of-deep-features and model ensembling," *IEEE Access*, 2021.

[15] A. Balagopalan, B. Eyre, et al., "To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection," *arXiv preprint arXiv:2008.01551*, 2020.

[16] J. Yuan, Y. Bian, et al., "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease.," in *INTERSPEECH*, 2020.

[17] M. Martinc, F. Haider, et al., "Temporal integration of text transcripts and acoustic features for alzheimer's diagnosis based on spontaneous speech," *Frontiers in Aging Neuroscience*, 2021.

[18] J. Li, J. Yu, et al., "A comparative study of acoustic and linguistic features classification for alzheimer's disease detection," in *ICASSP*. IEEE, 2021.

[19] Y. Wang, T. Wang, et al., "Exploring linguistic feature and model combination for speech recognition based automatic ad detection," *INTERSPEECH*, 2022.

[20] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?," in *ACL*, 2019.

[21] L. Hernández-Domínguez, S. Ratté, et al., "Computer-based evaluation of alzheimer's disease and mild cognitive impairment patients during a picture description task," *Alzheimer's & Dementia: DADM*, 2018.

[22] S. Luz, F. Haider, et al., "Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge," *INTERSPEECH*, 2020.

[23] A. Baevski, Y. Zhou, et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, 2020.

[24] W.-N. Hsu, B. Bolte, et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM TASLP*, 2021.

[25] S. Chen, C. Wang, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Top. in Signal Process.*, 2022.

[26] A. Radford, J. W. Kim, et al., "Robust speech recognition via large-scale weak supervision," Tech. Rep., OpenAI, 2022.

[27] S. Horiguchi, Y. Fujita, et al., "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," *INTERSPEECH*, 2020.

[28] J. Devlin, M.-W. Chang, et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[29] A. Vaswani, N. Shazeer, et al., "Attention is all you need," *NeurIPS*, 2017.

[30] C. d. Santos, M. Tan, et al., "Attentive pooling networks," *arXiv preprint arXiv:1602.03609*, 2016.

[31] J. T. Becker, F. Boiler, et al., "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of neurology*, 1994.

[32] X. Hao, X. Su, et al., "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP*. IEEE, 2021.

[33] P. A. Cariani and B. Delgutte, "Neural correlates of the pitch of complex tones. ii. pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch," *Journal of neurophysiology*, 1996.

[34] J. A. Colosi and M. G. Brown, "Efficient numerical simulation of stochastic internal-wave-induced sound-speed perturbation fields," *The Journal of the Acoustical Society of America*, 1998.

[35] L. Schuchman, "Dither signals and their effect on quantization noise," *IEEE TCT*, 1964.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[37] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *ASRU*. IEEE, 2021.

[38] A. Balagopalan and J. Novikova, "Comparing acoustic-based approaches for alzheimer's disease detection," *arXiv preprint arXiv:2106.01555*, 2021.