INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Exploring multi-task learning and data augmentation in dementia detection with self-supervised pretrained models

Minchuan Chen, Chenfeng Miao, Jun Ma, Shaojun Wang, Jing Xiao

Ping An Technology, China
{CHENMINCHUAN109}@pingan.com.cn

Abstract

Detection of Alzheimer's Dementia (AD) is crucial for timely intervention to slow down disease progression. Using spontaneous speech to detect AD is a non-invasive, efficient and inexpensive approach. Recent innovations in self-supervised learning (SSL) have led to remarkable advances in speech processing. In this work, we investigate a set of SSL models using joint fine-tuning strategy and compare their performance with conventional classification model. Our work shows that fine-tuning the pretrained SSL models, in conjunction with multi-task learning and data augmentation, boosts the effectiveness of general-purpose speech representations in AD detection. The results surpass the baseline and are comparable to state-of-the-art performance on the popular ADReSS dataset. We also compare single- and multi-task training for AD classification, and analyze different augmentation methods to show how to achieve improved results.

Index Terms: Dementia detection, Self-supervised Learning, Speech representation, Multi-task learning

1. Introduction

As the proportion of people with Alzheimer's Disease increases year by year, research into the early assessment of dementia has become increasingly important. AD-related neurodegeneration can cause cognitive changes that lead to progressive declines in memory and language abilities [1]. Traditional methods of AD detection are primarily based on clinical tests of cognitive decline in daily activities. However, these diagnostics are limited by time and resource availability. Recently, numerous studies show that valuable cognitive clinical information can be obtained from spontaneous speech [2, 3]. Some works use speech analysis [4], natural language processing (NLP) [5, 6] and machine learning (ML) [7, 8] to distinguish healthy and cognitively impaired participants in datasets that include speech tasks, such as picture descriptions. These approaches show great potential as a tool for rapid and non-invasive assessment of an individual's cognitive state.

Feature extraction of speech and language is an important first step in AD detection, which helps to distinguish whether participants in the dataset are healthy or suffering from AD. Generated speech features can be divided into two main groups, text-based and acoustic-based features. Compared with text-based methods, acoustic features are considered to be less discriminative than linguistic ones [9]. As described in [10], most of the top-ranked features are linguistic features such as parts of speech (POS) and word categories, while acoustic features rank relatively low for AD detection. According to previous studies [5, 8], using manual transcripts of speech or a combination of transcripts and speech audio can generally lead to better performance compared to using audio alone. But in practical applications, text-based detection methods are limited by manual transcription or rely heavily on the accuracy of automatic speech recognition (ASR) systems. Exploring effective speech representations from audio for AD classifier remains a challenging problem, and more robust and sensitive detection systems need to be built. There are existing efforts on the acoustic characteristics of AD detection. In [11], several acoustic feature sets are assessed on the DementiaBank Pitt Corpus. The results show that combining the eGeMAPS feature set [12] with hard fusion acquires the best accuracy of 78.70%. [13] utilizes the bottleneck features extracted from audio using an ASR model to achieve an accuracy of 82.59% on the Pitt dataset. [8] uses low-level descriptors of IS10-Paralinguistics feature set [14] and Bag-of-Acoustic-Words (BoAW) for feature aggregation, and achieve an accuracy of 76.85 % on the ADReSS challenge dataset. In [15], the authors use bag-of-deep-neural-embeddings and ensemble learning approaches to detect dementia, and obtain 83.33% accuracy with Music-Linear-BoW features for audio modality on the same dataset.

Self-supervised learning (SSL) is a rapidly growing research field that utilizes information extracted from the unlabelled data to learn useful representations for downstream tasks. There are two stages in this framework. In the first stage, SSL is used to pretrain a representation model on a huge number of unlabeled speech data, also called an upstream model. In the second stage, downstream tasks use either the learned representation from the frozen model, or fine-tune the entire pretrained model in a supervised phase. Recently, the framework has shown great success in the speech processing community [16]. Numerous SSL models have been proposed [17, 18, 19, 20] for different speech tasks.

In this paper, we propose a solution to solve the data scarcity problem in AD detection by using the pretrained SSL models and fine-tuning them on the dementia speech dataset, combined with multi-task learning and data augmentation techniques. Then we use temporal average pooling and a linear layer on top of the feature extractor to carry out classification. The evaluation shows that using the proposed method, we can achieve better performance than the baseline and comparable to currently published state-of-the-art (SOTA) results. The contributions of this work are three-fold:

- We explore speech representations extracted from different pretrained SSL models for AD detection using spontaneous speech without manual or automatic transcripts.
- We combine multi-task learning and data augmentation with the SSL models to obtain comparable results with SOTA.
- We evaluate the impact of audio augmentation on SSL models fine-tuning in terms of data size and kinds of methods.

10.21437/Interspeech.2023-1623

The remainder of the work is organized as follows: Section 2 introduces the related work. Section 3 describes the SSL frameworks, multi-task training and data augmentation methods. Section 4 reports the experiments and results. Finally, Section 5 concludes the paper.

2. Related work

2.1. Self-supervised learning

Self-supervised learning has emerged as a new paradigm to learn general data representations from unlabeled examples, and then fine-tune models on labeled data. In natural language processing (NLP), self-supervision has been used to leverage huge amounts of unlabeled data to train a language model using masked prediction [21]. Studies have shown that using self-supervised pretraining can improve performance on a variety of speech tasks [16, 22]. Self-supervised pretraining becomes an effective method for neural networks to utilize unlabeled data, which makes it possible to develop powerful models where labeled data is scarce. Many previous works [4, 23, 24] have explored the potential of using SSL models in AD classification task. In this work, we use wav2vec 2.0 [17], HuBERT [18] and SSAST [25] frameworks as the backbone network to extract feature representations from speech data.

2.2. Multi-task learning

Multi-task learning (MTL) as a regularization method can help to solve the problem of overfitting in low-resource scenarios. By learning an auxiliary task that is different from but related to the main task, the model can be forced to improve generalization performance. MTL has been successfully used in multiple speech tasks, such as speech recognition [26], speaker verification [27] and speech emotion recognition [28]. Previous works [6, 29, 30, 31] suggest that pause (duration, frequency and distribution) and filler words (such as 'uh' or 'um') are speech signs of dementia. In this work, we firstly implement single-dataset MTL using gender recognition as the auxiliary task. We also introduce an extra dataset and choose stuttering events detection as the auxiliary task which have similar characteristics with AD classification, such as speech disfluency. We hypothesize that the models trained on the auxiliary task can benefit from the additional gender or disfluency information as well as regularization effect.

3. Methodology

In this work, we propose a method for AD detection using speech representations extracted from the SSL models, combined with multi-task training and data augmentation.

3.1. SSL frameworks

3.1.1. wav2vec 2.0

wav2vec 2.0 is a framework which masks latent representations of the raw waveform and solves a contrastive task over quantized speech representations. A multilayer convolutional neural network (CNN) is used to encode speech into a latent space. The latent space representations are then quantized and randomly masked before being passed to the multihead self-attention layers of the transformer encoder. The InfoNCE loss [32] is adopted to minimize the distance between contextualized speech representations and quantized target vectors.

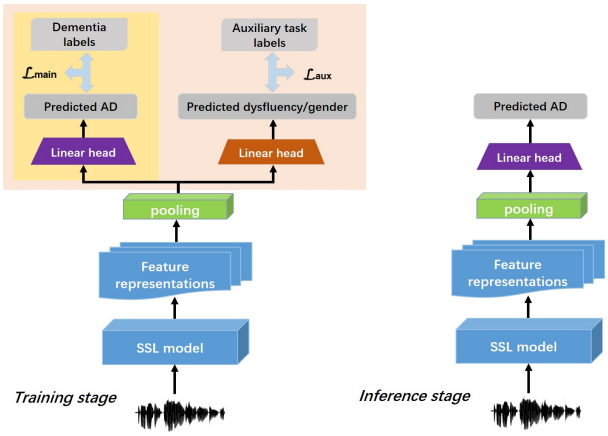

Figure 1: *The training and inference processes of the proposed method. The blue color represents the shared backbone. Yellow (upper left part in training stage) and pink (upper in training stage) are single-task and multi-task training, respectively.*

3.1.2. HuBERT

HuBERT shares the same architecture as wav2vec 2.0. In place of constructing a contrastive loss, HuBERT uses an offline clustering step to generate pseudo-labels for masked language model pretraining. The input speech is passed to the CNN encoder and randomly masked before passing to the transformer network, which then predicts the pseudo-labels of the masked regions. The target pseudo-labels are generated by performing two iterations of K-means clustering. The model evaluates cross entropy loss between the correct K-means cluster and the predicted one. The prediction loss is only applied to masked regions, forcing the model to learn high-level representations of unmasked inputs to infer the targets of masked ones correctly.

3.1.3. SSAST

SSAST is a self-supervised AST [33] model, which operates over patches of spectrogram and adopts joint discriminative and generative masked spectrogram patch modeling (MSPM). The 2D audio spectrogram is split into a sequence of patches and projected to a sequence of 1-D patch embeddings. Each patch embedding is added with a learnable positional embedding and then fed to the transformer encoder. During self-supervised pretraining, a portion of spectrogram patches are randomly masked. Two pretext tasks (finding the correct patch at each masked position and reconstructing the masked patch) are introduced to force the model to learn both the temporal and frequency structure of audio data.

3.2. Fine-tuning SSL models

The pretrained SSL models can be regarded as feature encoders, which yield contextualized speech representations for down-stream tasks. We extract the last hidden states of the pretrained models and use them as embedded representations of audio. A temporal average pooling and one linear head followed by softmax are added as a simple down-stream classifier. The temporal average pooling compresses variable time lengths into one, then the linear head implements an utterance-level classification that minimizes the cross-entropy loss (CEL). The training and inference processes are shown in Figure 1. In this work, we use joint fine-tuning of the upstream model and the classification layer.

Previous work [34] suggests that this is superior than fine-tuning the classification layer with frozen weights of the SSL model, since it can alleviate domain-shift between the pretrained and fine-tuned datasets.

3.3. Multi-task training strategies

To implement MTL, an additional output layer of the same dimensionality is added to the classification head as shown in Figure 1. This branch also takes the output of the same average pooling layer as the main branch as its input. We explore multi-task training in both single-dataset and multi-dataset scenarios. For single-dataset learning (SD-MTL), inspired by [35], we choose gender recognition as the auxiliary task which has shown to be an effective way in disfluency detection. The gender information of the ADReSS dataset can be obtained from the metadata provided by the sponsor. A weighted cross-entropy loss is computed for the auxiliary task, and the combined loss is calculated as below:

$$\mathcal{L}_{SD-MTL} = \lambda \mathcal{L}_{main} + (1-\lambda)\mathcal{L}_{aux} \quad (1)$$

where $\mathcal{L}_{main}$ is the cross-entropy loss of the AD classification task, $\mathcal{L}_{aux}$ is the cross-entropy loss of auxiliary task, and λ is a tunable task weight parameter. We experimentally set λ to 0.9.

For multi-dataset, multi-task learning (MD-MTL), we follow [36] which introduces a multi-task learning scheme that employs related tasks as well as related datasets in training process. In this work, we use the SEP-28k dataset [37] which is a corpus containing English stuttered speech. It consists of 28k annotated clips (23 hours) extracted from podcasts that contains five types of disfluencies: prolongations, blocks, sound repetitions, word repetitions and interjections. In experiment, we use a single network to train multiple tasks including stuttering events detection and AD classification, and extend it to handle tasks from different datasets. As shown in Figure 2, we mix batches with samples from the two datasets during training. For each batch, it contains data randomly selected from the two datasets for batch loss calculation, and each epoch consists of data from the combined training set. Since common loss averaging is not possible when batches are associated with different datasets and tasks, we subsample each batch (size S) based on its origin dataset and produce an effective batch per dataset (a or b). Then, we calculate each task's loss for the corresponding samples. The averaged losses are computed over the size of the effective batch and summed up as below:

$$\mathcal{L}_{MD-MTL} = (\sum_{a \in S} \mathcal{L}_{main}/a) + (\sum_{b \in S} \mathcal{L}_{aux}/b) \quad (2)$$

In the case that no sample within a mixed batch has label associated with one of the tasks, we zero out the corresponding dataset task's loss. The gradients for each task-specific layer are calculated based on the dataset task's loss, and accumulated before being backpropagated. Therefore, both tasks contribute to train the shared backbone network, regardless of the number of samples taken from each dataset. Because of the similarity of the datasets, the mixed batches may contain complementary information and prevent divergence in training.

3.4. Data augmentation

Data augmentation has been proved to be an effective way to deal with data sparsity and can improve performance on many speech tasks. Following [38], we select five types of audio augmentation approaches and combine them to generate extended training samples [1], including pitch shifting (Pitch), loudness shifting (Loudness), Vocal Tract Length Perturbation (VTLP), SpecAugment (SpecAug) and Voice Conversion (VC). Pitch shifting is a change in pitch without changing the speed of speech. We alter the pitch with a scale factor in the range of $(-15, 15)$. Loudness shifting increases or decreases the amplitude of the signal. We adjust the loudness of each sample by a scale factor in the range of (0.5, 5). VTLP performs warping along the frequency dimension on the Mel-spectrograms. We apply VTLP operations with a scaling factor between (0.3, 3). SpecAugment combines time warping, frequency and time masking on the log Mel-spectrograms. We also perform zero-shot voice conversion using a pretrained VC model. Lastly, we mix all augmented audios to create an extended training set. To evaluate the data size effect of the proposed method, we generate different scales (5, 10, 15 and 20x) of the original dataset.

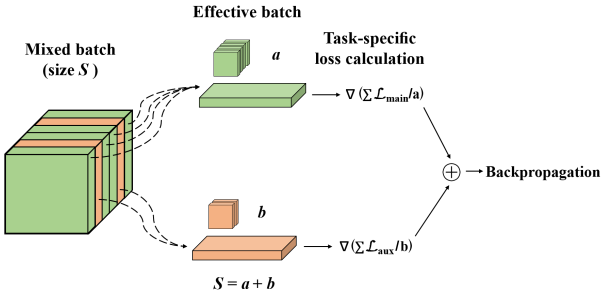

Figure 2: *Mixed-batch loss calculation. The loss from each task is averaged over its dataset's samples.*

4. Experiments

4.1. Experimental Setup

We work with the ADReSS dataset [2] which is a subset of Pitt dataset in Dementiabank. The dataset consists of speech samples from AD and non-AD English-speaking participants for the Cookie Theft picture description task, and is divided into train (108 participants, about 2 hours) and test (48 participants, about 1 hour) sets that balanced for age, gender and disease condition.

Since the ADReSS dataset is relatively small, we generate an augmented training set to fine-tune the pretrained SSL models with the single-task and multi-task schemes as described in 3.2 and 3.3. We split the training samples into chunks of 10s with a stride of 2s and apply data augmentation via the nlpaug package [2]. Both AD and non-AD in the ADReSS challenge's training set are used. For multi-task training, we split SEP-28k corpus to train and validation sets as 80%, 20%. We perform 5-fold cross validation to evaluate the models, and report the average accuracy and F1 score across 3 different seeds.

4.2. Comparison of single task training

We use Random Forest (RF) model trained with conventional acoustic features as the baseline. The eGeMAPS feature set [12] which contains 88 features is extracted from the augmented audios using the openSMILE toolkit.

Prior to extracting feature embeddings from the SSL models, we fine-tune different systems with the augmented data in single-task learning (STL) setting. For wav2vec 2.0 and HuBERT, we follow [17, 18] to freeze the CNN-based feature encoder, only fine-tune the parameters of the transformer blocks.

[1] https://github.com/hl-anna/DA4AD

[2] https://github.com/makcedward/nlpaug

Table 1: *Accuracy and F1 score results of single task training on the augmented dataset.*

Models	**Acc.**	**F1**
Baseline	0.678	0.691
wav2vec 2.0	0.737	0.728
HuBERT	**0.742**	**0.736**
SSAST	0.719	0.722
SOTA methods	**Acc.**	**F1**
Syed et al. [8]	0.768	-
Syed et al. [15]	0.833	-

Table 2: *Accuracy and F1 score results of multi-task training on the augmented dataset with respect to single-dataset (SD-MTL) and multi-dataset (MD-MTL).*

	SD-MTL		**MD-MTL**	
Models	**Acc.**	**F1**	**Acc.**	**F1**
wav2vec 2.0	0.752	0.738	0.766	**0.786**
HuBERT	**0.759**	**0.746**	**0.783**	0.779
SSAST	0.735	0.742	0.749	0.741

The wav2vec2-base and HuBERT-base upstream models are employed, both models are pretrained on 960 hours of unlabeled Librispeech dataset. We use the adam optimizer and fine-tune the models for 10 epochs with a learning rate of 10^{-5}, and batch size 32. We choose the frame-based SSAST model for experiment, which processes spectrogram in frame-by-frame manner and is pretrained on AudioSet and Librispeech. We fine-tune it up to 20 epochs, using a fixed learning rate of 10^{-4}.

The result in Table 1 shows that the SSL model-based systems outperform the RF baseline model, which proves that by leveraging powerful contextual speech representations we can obtain significant improvement from pre-training over conventional representations. It also shows that HuBERT is a better feature extractor compared to wav2vec 2.0 in AD classification, and the SSAST model achieve slightly lower result.

4.3. Comparison of multi-task training

For multi-task training, the SSL models are fine-tuned using the combined MTL loss from equation (1) or (2) instead of a single CEL term. The training parameters used for fine-tuning and the feature extraction schemes are identical to the ones described in section 4.2. we use batch size of 32 for both single-dataset and multi-dataset MTL experiments for comparison. To fine-tune on both the ADReSS set and SEP-28k corpus, we train the models for up to 20 epochs, with a patience of 3 epochs.

Results are shown in Table 2. The HuBERT based model gets the highest accuracy in both SD-MTL and MD-MTL, which is better or comparable to state-of-the-art results [8, 15]. The wav2vec 2.0 model achieves the best F1 score in MD-MTL. Comparing with STL, the results show that MTL significantly improves models' performance for AD classification task. For different multi-task training strategies, it shows that MD-MTL performs favorably over SD-MTL in all systems. We speculate that this improvement is due to the fact that stuttering detection is a task more related to AD classification and provides relevant aspects in speech representations, and the mixed batch strategy enable the network to acquire information across datasets.

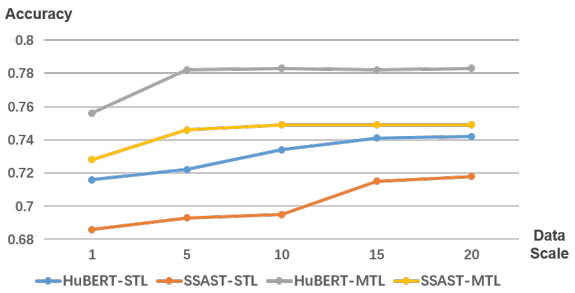

Figure 3: *The effect of data size for single-task learning (STL) and multi-task learning (MTL). The default 1x scale is the original training set of ADReSS corpus.*

Table 3: *Accuracy and F1 score results of training with different data augmentation methods.*

	HuBERT		**SSAST**	
Data configuration	**Acc.**	**F1**	**Acc.**	**F1**
ADReSS (no augment)	0.716	0.698	0.686	0.690
Pitch	0.725	0.711	0.703	0.706
Loudness	0.722	0.693	0.695	**0.708**
VTLP	0.719	0.718	0.687	0.692
SpecAug	**0.731**	0.715	**0.713**	0.697
VC	0.730	**0.722**	0.708	0.701

4.4. Effect of Data Augmentation

We investigate how data augmentation affects the performance of SSL models fine-tuning in terms of data size and augmentation method. We use two training settings (STL and MD-MTL) to test the HuBERT and SSAST models on the extended datasets of different scales. As shown in Figure 3, both models trained with the MTL scheme acquire the best performance at about 5x scale, more augmented data do not make more gain. For single-task training, the model need much more data (about 15x) to obtain the best result and get relative lower accuracy.

To evaluate different data augmentation approaches, we fine-tune the models with the STL scheme. Both models are also trained only on the original ADReSS dataset for comparision. The results in Table 3 show that all methods have positive effects on model's performance. SpecAugment is the most effective approach for SSL models fine-tuning. Voice conversion preserving speech prosody and duration also has significant benefits for model's training. VTLP, pitch shifting and loudness shifting make slightly improvements for AD classification task.

5. Conclusions

This work presents a comparative study of different embedding features abstracted from the pretrained SSL frameworks for AD detection. Experimental results on the ADReSS dataset reflect that combined with multi-task learning and audio data augmentation, the proposed method is able to detect AD with high accuracy that is comparable to the state-of-the-art results. In future work, we plan to improve the feasibility of pretrained embedding features for different kinds of dementia stages and extend our work to multi-class scenario.

6. References

[1] S. de la Fuente Garcia, C. W. Ritchie, and S. Luz, "Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: a systematic review," *Journal of Alzheimer's Disease*, 2020.

[2] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: the adress challenge," *Interspeech 2020*.

[3] S. Luz, F. Haider, S. de la Fuente, and et al., "Detecting cognitive decline using speech only: The adresso challenge," *Interspeech 2021*.

[4] A. Balagopalan and J. Novikova, "Comparing acoustic-based approaches for alzheimer's disease detection," *Interspeech 2021*.

[5] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection," *Interspeech 2020*.

[6] 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 *Interspeech 2020*.

[7] N. P. Shahla Farzana, "Exploring mmse score prediction using verbal and non-verbal cues," in *Interspeech 2020*.

[8] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for alzheimer's dementia through spontaneous speech." in *Interspeech 2020*.

[9] N. Cummins, Y. Pan, B. W. Ren, M. Magimai-Doss, H. Strik *et al.*, "A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition," in *Interspeech 2020*.

[10] J. Weiner, C. Frankenberg, J. Schröder, and T. Schultz, "Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews," in *ASRU 2019*.

[11] F. Haider, S. De La Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2019.

[12] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, and et al., "The geneva minimalistic acoustic parameter set gemaps for voice research and affective computing," *IEEE Trans. Affect. Comput. 2016*.

[13] Z. Liu, Z. Guo, Z. Ling, and Y. Li, "Detecting alzheimer's disease from speech using neural networks with bottleneck features and data augmentation," in *ICASSP 2021*.

[14] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge," 2010.

[15] 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. 88 377–88 390, 2021.

[16] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaloe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

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

[18] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[19] S. Chen, C. Wang, Z. Chen, Y. Wu, and et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[20] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," *Interspeech 2019*.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *NAACL 2019*.

[22] H. Aronowitz, I. Gat, E. Morais, W. Zhu, and R. Hoory, "Towards a common speech analysis engine," in *ICASSP 2022*, pp. 8162–8166.

[23] F. Braun, A. Erzigkeit, H. Lehfeld, T. Hillemacher, K. Riedhammer, and S. P. Bayerl, "Going beyond the cookie theft picture test: Detecting cognitive impairments using acoustic features," in *Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6–9, 2022, Proceedings*. Springer, 2022, pp. 437–448.

[24] 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 *Interspeech 2021*.

[25] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *AAAI 2022*.

[26] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for robust speech recognition," in *ICASSP 2020*.

[27] S. Sigtia, E. Marchi, S. Kajarekar, D. Naik, and J. Bridle, "Multi-task learning for speaker verification and voice trigger detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6844–6848.

[28] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning," in *Interspeech 2021*.

[29] J. Yuan, X. Cai, and K. Church, "Pause-encoded language models for recognition of alzheimer's disease and emotion," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7293–7297.

[30] P. Pastoriza-Domínguez, I. G. Torre, F. Diéguez-Vide, I. Gomez-Ruiz, S. Gelado, J. Bello-López, A. Ávila-Rivera, J. A. Matias-Guiu, V. Pytel, and A. Hernández-Fernández, "Speech pause distribution as an early marker for alzheimer's disease," *Speech Communication*, vol. 136, pp. 107–117, 2022.

[31] B. H. Davis and M. A. Maclagan, "Uh as a pragmatic marker in dementia discourse," *Journal of Pragmatics*, vol. 156, pp. 83–99, 2020.

[32] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[33] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *Interspeech 2021*.

[34] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[35] S. P. Bayerl, D. Wagner, E. Nöth, and K. Riedhammer, "Detecting dysfluencies in stuttering therapy using wav2vec 2.0," *Interspeech 2022*.

[36] G. Kapidis, R. Poppe, and R. C. Veltkamp, "Multi-dataset, multitask learning of egocentric vision tasks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 01, pp. 1–1, 2021.

[37] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. P. Bigham, "Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter," in *ICASSP 2021*.

[38] D. Woszczyk, A. Hedlikova, A. Akman, S. Demetriou, and B. Schuller, "Data augmentation for dementia detection in spoken language," *Interspeech 2022*.