Augmented Adversarial Self-Supervised Learning for Early-Stage Alzheimer’s Speech Detection

Longfei Yang1,∗, Wenqing Wei2,∗, Sheng Li3, Jiyi Li4, Takahiro Shinozaki1

1Tokyo Institute of Technology
2Japan Advanced Institute of Science and Technology
3National Institute of Information and Communications Technology
4University of Yamanashi

longfei.yang.cs@gmail.com, wqwei@jaist.ac.jp, shengli@nict.go.jp, jyli@yamanashi.ac.jp, shinot@ict.e.titech.ac.jp

Abstract

The early-stage detection of Alzheimer’s disease has been considered an important field of medical studies. While speech-based automatic detection methods have raised attention in the community, traditional machine learning methods suffer from data shortage because Alzheimer’s record data is very difficult to get from medical institutions. To address this problem, this study proposes an augmented adversarial self-supervised learning method for Alzheimer’s disease detection using limited speech data. In our approach, Alzheimer-like patterns are captured through an augmented adversarial self-supervised framework, which is trained in an adversarial manner using limited Alzheimer’s data with a large scale of easily-collected normal speech data and an augmented set of Alzheimer’s data. Experimental results show that our model can effectively handle the data sparsity problems and outperform the several baselines by a large margin. The performance for the “AD” class has been improved significantly, which is very important to actual AD detection applications.

Index Terms: Alzheimer’s disease detection, augmented adversarial self-supervised learning

1. Introduction

Alzheimer’s disease (AD) is a neurodegenerative disease that involves the decline of cognitive and functional abilities as the illness progresses [1]. It is the cause of 60-70% cases of dementia [2]. AD has been a global problem. As of 2020, approximately 50 million people were found being with AD worldwide [3]. Unfortunately, this disease still cannot be perfectly cured and the cause of AD even remains poorly understood [4]. Therefore, it is far-reaching and instrumental for relevant people if AD can be detected in the early stage. Conventionally, AD is usually diagnosed by requiring a comprehensive examination by medical experts according to the person’s medical history, history from relatives, behavioral observation, and so on [5]. Among behaviors in the early stage of AD, memory loss is generally regarded as the signature symptom [6]. As a consequence, language impairment may also appear in its early stage. Therefore, the nature of speech and language is considered a biomarker for AD and has been investigated by several studies [7, 8]. Several efforts have been made in previous works on establishing an early-stage AD detection system using speech data using state-of-the-art speech and language analysis technologies [9, 10, 11, 12] and other techniques [13, 14, 15]. Several studies indicate that the properties in speech can be regarded as a source of clinical information for AD detection, e.g., speech changes in fluency [16], prosody [17], and so on. Despite all this, it is still a challenge to achieve a high-performance AD detection system. Supervised learning is widely used in existing works on AD detection systems, which highly rely on plenty of training data. It is hard and time-consuming to collect a large number of speech data from AD patients because medical record data of specific patients are very difficult to get from each medical institution. Even though some techniques like data augmentation are proposed to deal with this problem, data sparsity makes it challenging to set up the model directly on a limited dataset. Recently, several substantial works have shown that the representations of different levels of information can be learned from raw data with self-supervised objectives, and these representations are effective for some downstream related tasks [18]. Oord et al. [19] proposed an unsupervised learning framework based on contrastive predictive coding to extract representations from high-dimensional data by predicting the future in latent space with auto-regressive models and a probabilistic contrastive loss that induces the latent space to capture information that is maximally useful to predict future samples. Schneider et al. [20] employed an unsupervised pre-training approach for automatic speech recognition and suggested the learned representation may contain some acoustic patterns. Yang et al. [21, 22] proposed a self-supervised model to learn the non-native patterns for the mispronunciation detection task. Kreuk et al. [23] proposed a self-supervised learning model for unsupervised phoneme boundary detection for speech segmentation. Inspired by these researches, we propose an augmented adversarial self-supervised learning method for the AD detection task in this paper. We expect that a large amount of easily-collected normal speech can be utilized to alleviate data sparsity problems in AD detection. In our work, an augmented self-supervised model is trained to capture the representation in which AD’s patterns can be found in acoustic information by aligning the distribution of the normal and AD in latent space using plenty of normal speech data and an augmented set on AD data. The experimental results on the dataset of AD2021 (in Subsection 3.1), i.e., Alzheimer’s Disease Recognition Evaluation, demonstrate that our proposed method is effective in handling the data sparsity problem for the AD detection task.

Specifically, we make the following contributions. (1) We propose to utilize a large amount of normal speech to handle the sparsity problem in the speech-based Alzheimer’s disease detection task. (2) We propose an adversarial self-supervised...
learning approach with augmented adversarial samples to capture relevant information from the limited AD dataset and a large-scale normal speech set. (3) The experimental results indicate that our proposed methods effectively improve the performance for AD detection.

2. Alzheimer’s Disease Detection based on Augmented Adversarial Self-Supervised Learning

Our goal is to obtain a model which can capture relevant patterns with limited AD speech data (“AD/MCI”) and a large number of normal speech (“HC”) data. Our proposed model contains two parts: the pre-trained part and the downstream AD part. Briefly, the large-scale normal speech and small-scale AD speech are firstly employed to train the pre-trained model. Then the pre-trained model extracts feature for the AD detection task. Figure 1 shows the architecture of our proposed approach.

2.1. Proposed Method

2.1.1. Self-Supervised Learning

The pre-training model in our proposed method is based on contrastive predictive coding, which is a slow feature analysis framework using unsupervised representation learning. It indicates that the information captured by predicting the observation span of many time steps may contain unique properties, e.g., phonemes and intonation in speech [19]. Prediction of different timescales will capture different levels of information [24]. The main reason behind it is that the main function of the brain is simply to minimize the errors between predicted input and the input received. This process is similar to the modern machine learning methods in which the model is trained to end with a global loss between the output and the ground truth.

In detail, let \( x = \{x_1, x_2, ..., x_i\}, x_i \in \mathbb{R} \), denotes a segment of raw signal of normal speech in \( L \) discrete time steps where \( x_i \) is the acoustic amplitude at time \( i \). \( x' = \{x'_1, x'_2, ..., x'_i\}, x'_i \in \mathbb{R} \) is a segment with \( L \) length of raw speech of AD. Firstly, an encoder \( g_{enc} \) encodes these two raw signal into embedding vector representations \( z = \{z_1, z_2, ..., z_i\}, z_i \in \mathbb{R}^{d_{z}} \) and \( z' = \{z'_1, z'_2, ..., z'_i\}, z'_i \in \mathbb{R}^{d_{z}} \), where \( d_{z} \) is the hidden representation dimension. In this work, the encoder contains several convolutional layers that have a ability to capture local information, i.e.,

\[
    z = g_{enc}(x_1, x_2, ..., x_i),
\]

\[
    z' = g_{enc}(x'_1, x'_2, ..., x'_i).
\]

Subsequently, a sequence model \( g_{seq} \) summarizes the past information in the vector embedding sequence, and generates corresponding context-aware representations for normal and AD speech, which can be denoted as \( c = \{c_1, c_2, ..., c_i\}, c_i \in \mathbb{R}^{d_{c}}, \) i.e.,

\[
    c = g_{seq}(x), \quad c' = g_{seq}(x').
\]

Then the model is trained to predict the future latent representation \( x_{i+k} \) using the context-aware representation at \( t \)-th timestep for normal speech. The objective function is designed by minimizing the InfoNCE [19]. InfoNCE is noise contrastive estimation-based loss function, which maximizes the mutual information lower bound between context aware embedding \( c_i \) and future latent representations \( z_{i+k} \) for \( k \in \{1, ..., K\} \). Given a set \( Z = \{z_1, z_2, ..., z_N\} \) which contains one positive sample from \( p(z_{i+k}|c_i) \) and \( N-1 \) negative samples from

![Figure 1: The architecture of our proposed adversarial self-supervised learning. The left is the pre-training part where the model is trained to learn the input representation from normal speech, and it is adversarial to the AD speech (“AD/MCI”). (Note that the hidden representation of AD speech does not take part in InfoNCE loss). The red arrow at the top means that the loss is maximized through a gradient reversal layer to confuse the classifier. The right is the downstream AD detection. The testee’s speech is fed into the pre-trained model to generate features and then detection model output indicating whether the owner of the input is “AD”, “MCI” or “HC.”](image-url)
“noise” distribution $p(z)$. For each step $t$, the definition of the InfoNCE loss is as follows:

$$L_{\text{NCE}}^t = -E \log f_k(c_t, z_{t+k}) \sum_{z \sim p(z)} f_k(c_t, z)$$

where $f_k(c_t, z_{t+k})$ is a scoring function that can be a bilinear model:

$$f_k(c_t, z_{t+k}) = \exp(c_t^T W_k z_{t+k})$$

where $W_k$ is the parameters in each model for each $k$. The total loss to be minimized is a sum of the InfoNCE loss for each step:

$$L_{\text{NCE}} = \sum t \sum_k L_{\text{NCE}}^t$$

which samples negative samples from the speech signal $z$ representations according to a uniform distribution. Note that the hidden representation of AD speech does not take part in the calculation of InfoNCE.

At the same time, another output is to determine the input speech is from the normal or AD speech with an explicit auxiliary task by sending the context vector of normal speech $c$ and AD speech $c^\epsilon$ to a classifier. The training process of this classifier is adversarial with respect to the shared hidden layer in the classifier by employing gradient reversal to maximize the classification loss $L_{\text{clas}}^t$ between the output of the classifier and the ground-truth. With it, the generated normal context vector may be drawn closer to that of AD. In other words, the model is trained to generate some specific patterns in AD’s speech, which may be hard to capture using limited training data.

$$L_{\text{clas}} = -\hat{y} \log y_w - (1 - \hat{y}) \log (1 - y_w)$$

where $\hat{y} \in \{0, 1\}$ denotes the ground-truth label, $y_w = p(y|\theta; h_{\text{clas}})$ is the softmax output, $W_{\text{clas}}$, $b \in \theta_{\text{clas}}$ are the output and weights of the final layer, and $h$ is the hidden representation sent to the final layer.

After obtaining the pre-trained model, we utilize it for the downstream AD detection task. The model of the AD detection task accepts the context-aware representations from the pre-trained model as input and makes classification of three classes “AD”, “MCI” and “HC” as output. The AD detection model (CNN and CNN-LSTM are used in our method) should be trained individually while the pre-trained model parameters are frozen. The detailed descriptions of our implementation can be found in Subsection 3.3.

2.1.2. Augmented Adversarial Samples

Since the number of speech of Alzheimer’s disease is much smaller than that of healthy people, the pattern of Alzheimer’s disease would be a little plain in the limited dataset. To overcome it, we utilize data augmentation making the Alzheimer’s patterns more diverse. We employ three augmentation schemes in this paper, which are as follows:

- Speed based augmentation. The speed function of $\text{Sox}$ is used to resample a signal to modify the speed of it. Ten random factors are chosen in the range of $[0.90, 1.10]$ to make ten copies of the original AD dataset.

- Tempo based augmentation. The tempo of the signal is modified while ensuring that the pitch and spectral envelope of the signal does not change. The WSOLA based implementation in the tempo command of the $\text{Sox}$ tool is employed to make it. Similar to the speed based augmentation, ten random factors are chosen in the range of $[0.90, 1.10]$ to generate a 10-fold dataset.

- Tremolo based augmentation. A tremolo effect is applied to the signal to amplitude low frequency. Ten random tremolo frequencies in $[2\text{KHz}, 6.5\text{KHz}]$ and the depth of 40 are applied for this augmentation.

3. Experiments

3.1. Data Description

For the Alzheimer’s disease (AD) dataset, it is provided by the 2021 NCMMSC Alzheimer’s Disease Recognition Challenge. In total, 7,16-hour speech from 39 male speakers and 54 female speakers are the training set (Training), and a 0.67-hour speech set from 15 male and 15 female speakers is the development set (Dev). The training and dev samples are segmented into 6s to keep consistent with the official short speech track test set with a 1.92-hour speech set (Testing). All the sentences have labels of three kinds: “AD”, “MCI” and “HC”.

For the normal speech dataset, since the AD dataset is a dataset of Chinese speakers, we choose AISHELL corpus as the source of normal speech, which is an open-source Mandarin Chinese speech corpus collected for speech-related research [25]. We randomly chose around 80 hours of data from it as our training set for pre-training.

3.2. Baselines

The convolutional neural network (CNN)-based system provided by the official organizer is employed as the baseline system. This model is directly trained on the AD dataset only. A 20-dimensional Mel frequency cepstral coefficients (MFCC) vector extracted with a 25ms window and 10ms frameshift is employed as the feature of each frame. The input layer consists of $259 \times 20$ nodes. Subsequent layers are comprised of 5 convolutional layers, and each convolutional layer is followed by one ReLU activation function and one max-pooling layer. Batch normalization is followed to standardize the inputs to the next layer for each mini-batch. The convolutional layers’ settings of number of filters, filters size, the strides and the paddings are sequentially arranged as $[32, 32, 32, 64, 128], [3, 3, 3, 3, 3], [1, 1, 1, 1]$ and $[2, 2, 2, 2, 2]$ respectively. The following layer is a convolutional pooling layer, a layer with 128 filters of size 1, stride 1, and padding 1 is employed. The output is fed into a ReLU layer and an adaptive 1-dimensional max-pooling layer. Finally, the last two layers are fully connected (FC) layers with 256 nodes before a 3-node output layer with a softmax function. A dropout factor of 0.3 is applied to the layer before the final output layer. The Adam optimizer is employed to train the model with a learning rate of 1e-3 and a mini-batch size of 128.

3.3. System Implementation

Pre-Trained Model Setup: The encoder stacks five 1-dimensional convolutional layers. They have the same down-sampling rate of 1/160 to get the same frame rate with the phoneme sequence labels from the Kaldi-based system (10ms). These layers have the same settings of filters size, strides, and paddings $[10, 8, 4, 4, 4], [5, 4, 2, 2, 2]$, and $[3, 2, 1, 1, 1]$ and followed by the Batch normalization and ReLU activation

1http://ncmmsc2021.org/competition.html
2Available at https://github.com/THU/satlab/AD2021
Table 1: Main experimental results. Aug-Adv-SSL denotes our proposed adversarial self-supervised model with contrastive predictive coding using augmented negative samples. -FT denotes the parameters of the pre-trained model that is fine-tuned at the downstream stage.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>72.3</td>
<td>73.7</td>
<td>71.8</td>
</tr>
<tr>
<td>Wav2vec2.0 [26]</td>
<td>77.9</td>
<td>77.2</td>
<td>77.2</td>
</tr>
<tr>
<td>Aug-Adv-SSL</td>
<td>83.61</td>
<td>83.0</td>
<td>83.15</td>
</tr>
<tr>
<td>Aug-Adv-SSL-FT</td>
<td>83.77</td>
<td>83.69</td>
<td>83.73</td>
</tr>
</tbody>
</table>

Table 2: Ablation study about the importance of each component in our proposed model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug-Adv-SSL</td>
<td>83.61</td>
<td>83.0</td>
<td>83.15</td>
</tr>
<tr>
<td>w/o Adv</td>
<td>81.9</td>
<td>81.5</td>
<td>81.5</td>
</tr>
<tr>
<td>w/o Aug</td>
<td>81.88</td>
<td>82.64</td>
<td>82.26</td>
</tr>
<tr>
<td>w/o Both</td>
<td>78.51</td>
<td>77.45</td>
<td>77.56</td>
</tr>
</tbody>
</table>

Figure 2: Confusion matrix of the results of baseline (CNN-based) and ours (CNN-based) models.

4. Results and Discussions

4.1. Main Results

The experimental results are reported in Table 1. The show that the overall performance is improved with our model. It is reflected in that the macro F1-score is improved from 71.8 of the baseline model to 83.15 of our proposed adversarial self-supervised model with contrastive predictive coding using augmented negative samples. We also notice that our proposed model outperforms the famous model proposed by [26] which exploits a pre-trained wav2vec model on the same tasks with the same dataset. In addition, fine-tuning work on the unfrozen parameters of the pre-trained model is effective in further improving the performance of the pre-trained model to handle the mismatch between different datasets.

In detail, the confusion matrices of the results are depicted in Figure 2. We notice that, with our proposed pre-trained methods, the accuracy of “AD” is significantly improved from 42.57% to 77.26%. It verifies our plan that the “AD” patterns are captured by our augmented adversarial self-supervised learning, and these patterns contribute to the system’s high detection accuracy. It can also be found that our proposed methods do not affect the performance for other classes, which is reflected in that the performance for “MCI” improved a little and that of “HC” does not change.

4.2. Ablation Study

To further investigate our proposed approach, we present the results that show the effectiveness of the components of the proposed approaches. Table 2 demonstrates the results when we progressively remove the components of our proposed model. As shown in Table 2 on one hand, when the adversarial components are removed, the model is an SSL model trained with a large of the normal speech and a set of augmented AD speech, in which the negative samples are drawn the same utterance to the input for normal and AD domain separately. The results decrease a lot for the F1-score from 83.15 to 81.5, for recall and precision as well. It indicates that the adversarial components are effective and important to the proposed model. When removing the augmentation components, the model is trained with the normal speech and a set of limited AD speech serving as negative samples without augmentation. The results decline as well because of the single patterns in the limited AD dataset. When removing both of them, the model is an SSL model with large-scale normal speech and a limited AD speech. Although the performance is far behind that of our proposed model, it still outperforms the baseline. It indicates that our expectation of introducing a large number of normal speech is useful for dealing with data sparsity problems for the AD detection task.

5. Conclusion

In this work, we propose to utilize a large amount of normal speech to handle the sparsity problem in the speech-based Alzheimer’s disease detection task. We propose an augmented adversarial self-supervised learning approach to capture relevant information from the limited AD dataset and a large-scale normal speech set. In the model, the limited AD dataset is applied with an augmentation to diversify the anomaly patterns. The experimental results indicate that our proposed methods effectively improve the performance for AD detection to a large margin compared to the official baseline and other models.
6. References


