Detecting Alzheimer’s Disease Using Gated Convolutional Neural Network from Audio Data

Tifani Warnita, Nakamasa Inoue, Koichi Shinoda
Tokyo Institute of Technology, Tokyo, Japan
{tifani,inoue}@ks.cs.titech.ac.jp, shinoda@c.titech.ac.jp

Abstract

We propose an automatic detection method of Alzheimer’s diseases using a gated convolutional neural network (GCNN) from speech data. This GCNN can be trained with a relatively small amount of data and can capture the temporal information in audio paralinguistic features. Since it does not utilize any linguistic features, it can be easily applied to any languages. We evaluated our method using Pitt Corpus. The proposed method achieved the accuracy of 73.6%, which is better than the conventional sequential minimal optimization (SMO) by 7.6 points.

Index Terms: Alzheimer’s disease, computational paralinguistics, convolutional neural network, gating mechanism

2. Related Works

Linguistic information had been widely utilized in order to automatically detect AD patients [12, 13, 9]. For instance, Wankerl et al. (2017) [9] employed a pure linguistic approach based on n-gram on Pitt Corpus [14], in which subjects undergo a picture description task. From the study, they found that patients often not only uttered incomplete phrases but also interrupted others. These degraded the intelligibility of their speech [15].

Some studies utilized both linguistic features and acoustic features to detect people having AD [15, 3, 16]. Fraser et al. (2016) [3] utilized multilinear logistic regression and selected 35 top-ranked features out of 370 features by using Pitt Corpus [14]. Khodabakhsh et al. (2015) [15] used the recordings of conversational speech of 32 AD patients and 51 control subjects for this combination. Weiner et al. (2016) [17] employed only acoustic information from German conversational recordings. However, the study relied on the transcription for calculating some features such as silence-to-word ratio and word rate.

In contrast with these studies, we focused on non-linguistic approach where only speech audio features of the subjects are utilized. It can be easily applied to other languages and expected to be more robust against environmental noise and channel differences. As features, we used the paralinguistic features that frequently used in emotion recognition task since previous studies showed that AD people tend to show emotional prosodic impairment [18]. Numerous approaches for detecting emotions emerged from people have been carried out. Schuller et al. (2009) [19] provided the baseline for emotion recognition in the INTERSPEECH 2009 Emotion Challenge, which uses sequential minimal optimization (SMO). Huang et al. (2014) [10] employed deep learning based approach in which convolutional layers were employed. Keren and Schuller (2016) [11] used not only convolutional but also recurrent layers.

3. Database

In this study, we used the picture description task session data of Pitt Corpus [14], which is a part of DementiaBank, a multimedia database for studying people having dementia. In the picture description task, patients were asked to describe what happens in a picture, Cookie Theft Picture of the Boston Diagnostic Aphasia Examination [20]. Pitt Corpus consists of speech data and their transcription from 244 control (healthy) subjects as well as 309 patients having dementia such as mild cognitive impairment (MCI), vascular dementia, and AD. In this study, we used only the data from AD patients and from patients suspected of having AD (probable AD). It should be noted that, even though we used the same database as in [3, 9], the number of the subjects in our study was slightly different compared to theirs according to the following three reasons. First, the size of the database has increased over time. Second, we ex-
cluded speech data with overlapped sounds from the other in-
terview sessions. Third, we only used data with both audio data and transcription information. Consequently, the number of sessions is 488 (253 AD, 233 control) from 267 subjects (169 AD, 98 control). Similar to [3], due to a limited amount of data, we employed a 10-fold cross-validation scheme for our evaluation. We designed the ten subsets so that no subject appeared in more than one subset in order to avoid any speaker dependencies.

In this study, we performed four preprocessing stages. First, we decrease the sampling rate into 16kHz. After that, we normalizе each audio signal using the average value of decibels relative to full scale (dBFS) in the data. Then, we segmented the audio data of each subject into utterances according to the transcription information. We obtained the total of 6267 utter-
ances (3276 AD, 2991 control). Finally, we added 10ms at the beginning and at the end of each utterance segment with 15ms fade-in and fade-out.

4. Features

In this study we used openSMILE [21] to extract acoustic fea-
tures from Pitt Corpus. OpenSMILE consists of several differ-
ent configurations of acoustic feature extraction. It has been
mainly used for emotion recognition but we believe it is also
effective for our purpose. While AD patients can easily get de-
pressed, anxious, or even upset [1], they find it difficult to ex-
press their emotions in prosodies such as tempo alteration and
powerful intonation [18]. Based on this finding, we used the
following feature sets.

1. INTERSPEECH 2009 Emotion Challenge Features (IS09) [19]

In this feature set, there are 16 types of low-level descrip-
tors (LLDs) extracted from the frame level. The LLDs include several prosodic features such as pitch and ener-
gy which are expected to represent the emotional pat-
tern of the speech. The delta coefficients are also cal-
culated hence producing the total of 32 LLDs. In or-
der to get the utterance-level features from the LLDs, 12
functionals (e.g. the values of minimum and maximum, mean, and range) are applied to each LLDs. As a result, 384 features are extracted from one utterance.

2. INTERSPEECH 2010 Paralinguistic Challenge Features (IS10) [22]

The additional LLDs to IS09 are PCM loudness, eight
log Mel frequency band (0-7), eight line spectral pairs (LSP) frequency (0-7), F0 envelope, voicing probability, jitter local, jitter DPP, and shimmer local. These features are expected to give broader coverage of paralinguistic
events in speech. In addition, more MFCC features are extracted (0-14 compared to 1-12). Finally, we get 76 LLDs for one frame and the total of 1582 features for one utterance.

3. INTERSPEECH 2011 Speaker State Challenge Features (IS11) [23]

Compared to the previous feature sets, IS11 provides the
derived loudness measure and the employment of Rela-
tive Spectral Analysis (RASTA)-style filtered auditory
spectra resulting in 118 LLDs for the frame-level fea-
tures. The total number of features in one utterance is
4368.

4. INTERSPEECH 2012 Speaker Trait Challenge Features (IS12) [24]

In this feature set, some of the LLDs added compare to IS11 include harmonic-to-noise ratio (HNR), spectral
harmonicity, and psychoacoustic spectral sharpness re-
sulting in 120 LLDs for the frame-level features. After being applied by functionals, the total number of features in one utterance is 6125 features.

5. Method

5.1. Convolutional Neural Network

In this study, assuming that temporal features are well repre-
sented in the frame-level features, we employed a Convolutional Neural Network (CNN) [25] as a classifier. A CNN was inspired by the structure of animal visual cortex for perceiving lights [26] and has yielded supreme results in an abundance of tasks in the past few years [27]. Furthermore, it needs a relatively small amount of training data compared to the other networks since it has a much smaller number of connection weights [28].

The convolution operation over the input aims to extract the
temporal information by sliding through the use of a kernel
(filter). In our study, we feed utterance segment features \( X \in \mathbb{R}^{F \times T} \) into the CNN, where \( F \) and \( T \) represent the dimension of LLDs and the number of time frames, respectively. The size of the sliding window is \( F \times N \) where \( N \) is the window length of the kernel in the time axis. The convolution operation between the kernel and the input will produce a scalar output. Output from the convolution layer at time \( i \), \( i = n + 1, \ldots, T \) is then defined as,

\[
y_i = g \left( \sum_{j=1}^{F} \sum_{n=1}^{N} w_{f,n} x_{f,j-n} + b \right),
\]

where \( g \cdot \) denotes the bias and activation function re-
spectively. \( x_{c,d} \) denotes the (\( c,d \)) element of the input \( X \), and \( w_{c,d} \) is the (\( c,d \)) element of the weight matrix \( W \). Both \( W \) and \( b \) are the learnable parameters of the network that we train. When the kernel is sliding through the input matrix over the time dimen-
sion, we multiply the overlapping element of the two matrices as in Eq. 1. We employed a rectified linear unit (ReLU) [29] as the activation function \( g \). This network is also referred as Time-Delay Neural Network (TDNN) [30].

Since the input dimension for CNN should be fixed, we set the segment length as that of the longest utterance segment in the dataset. After that, we applied zero padding for the rest of the
utterance segment. We use the number of kernels \( K = 64 \)
and the window length of \( N = 3 \). Accordingly, the calculation of every patch of the input will produce the complete output \( Y \in \mathbb{R}^{M \times K} \) where \( M \) is the number of time frames, \( T - N + 1 \).

We added batch normalization before each activation func-
tion [31]. We also used random weight initialization for each convolution layer. After each convolution layer, we put a max-
pooling layer [32].

The output of the last convolution layer is flattened into one
feature vector. For example, the flattened output \( Y \in \mathbb{R}^{M \times N} \) will produce a vector \( Z \in \mathbb{R}^{O} \) where \( O = M \times N \). This vec-
tor became the input to a fully-connected layer with activation
function ReLU consisting of 256 hidden neurons. We also em-
ployed batch normalization and initialized the weight matrix as
random uniform. Furthermore, we also applied dropout with the
value of 0.5 before the output layer. The output layer consists of
one hidden neuron with a sigmoid function. We used binary
cross-entropy as the loss function and Adam [33] as the opti-
mizer. We trained the network with the pair of LLDs of each
In addition to the standard deep convolutional neural network, we introduce a gating mechanism after each convolution. The resulting network is called Gated Convolutional Neural Network (GCNN). A gate represents an information controller which manages the information that flows into the succeeding layer. The gate has the ability to prevent the gradient from being vanished during backpropagation [34]. Recently, it has been often used in convolutional neural networks with various objectives such as for conditional image modeling [35], language modeling [34], and speech synthesis [36]. A previous study [34] showed that gated linear unit (GLU) outperformed gated tanh unit (GTU) which is used in [35]. Therefore, we employed GLU in our study.

Similar as in Section 5.1, we give the input of extracted LLDs features of each utterance segment \( X \in \mathbb{R}^{T \times F} \). We used a sigmoid function as the activation function \( g \) which is multiplied by a linear gate. Eq. 1 is modified as:

\[
y_i = \mathbf{v} \cdot g\left( \sum_{f=1}^{F} \sum_{n=1}^{N} v_{f,n} x_{f,i-n} + b \right),
\]

where \( v_{c,d} \) represents the element of \( V \) at position \( (c, d) \). The \( V \in \mathbb{R}^{F \times F} \) and \( e \in \mathbb{R} \) are the kernel weight matrix and bias for the linear gate. For GCNN, we used \( N = 2 \) and applied the same padding.

After that, we halved the output length in a max-pooling layer [32]. Figure 1 shows the visualization of our GCNN with the depth = 1 where the depth represents the number of gated convolution layers. In the figure, one gated convolution layer lies between the dotted horizontal line which followed by the max-pooling layer. Deeper networks consist of more gated convolution layers. We also defined the same layers after the max-pooling layer as in Section 5.1.

5.3. Framework Overview

While we need to classify each subject based on his/her whole data, we performed the classification for each utterance instead. This is based on our assumption that the information about a patient having AD or not can still be obtained from a smaller but appropriate length of a segment. After the utterance-based classification is performed, we make the final verdict for each subject based on the proportion of each class; we classified a subject into AD if he/she has AD percentage above the half. The utterance-level subject classification is expected to give more detailed information about the symptoms while most of the previous studies conducted the subject-level classification [15, 3, 12, 13, 9, 16].

6. Experimental Results

We used the average accuracy from the 10-fold cross-validation scheme on Pitt Corpus [14] (see Section 3 for details) for evaluating our method. Our first experiment was the employment of the standard deep CNN without gates on the four feature sets, IS09, IS10, IS11, and IS12. Figure 2 shows the classification result of AD or non-AD with different numbers of hidden layers (1, 2, 3, 4, 6, 8, 10). The performance result given is from the subject-level classification after majority voting from the utterance-level classification. Since there is no established
Table 1: Comparison of the employment of the standard CNN and GCNN by using IS10 feature set represented in the average accuracy (%) from 10-fold cross-validation.

<table>
<thead>
<tr>
<th>Depth</th>
<th>Standard CNN</th>
<th>Linear GCNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Utterance</td>
<td>Subject</td>
</tr>
<tr>
<td>1</td>
<td>64.2</td>
<td>66.0</td>
</tr>
<tr>
<td>2</td>
<td>64.2</td>
<td>66.6</td>
</tr>
<tr>
<td>3</td>
<td>64.2</td>
<td>69.2</td>
</tr>
<tr>
<td>4</td>
<td>64.9</td>
<td>68.7</td>
</tr>
<tr>
<td>6</td>
<td>65.5</td>
<td>68.6</td>
</tr>
<tr>
<td>8</td>
<td>66.1</td>
<td>69.0</td>
</tr>
<tr>
<td>10</td>
<td>65.2</td>
<td>70.4</td>
</tr>
</tbody>
</table>

Table 2: Confusion matrix depicting the prediction results from all folds using IS10 feature set and GCNN with the depth of 8.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Utterance-Level</th>
<th>Subject-Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AD</td>
<td>Control</td>
</tr>
<tr>
<td>AD</td>
<td>2340</td>
<td>1213</td>
</tr>
<tr>
<td>Control</td>
<td>936</td>
<td>1778</td>
</tr>
</tbody>
</table>

baseline for this database and the number of instances used are different from one experiment to another experiment, we give the result of the four feature sets with baseline methods, which is the subject-level classification using sequential minimal optimization (SMO) \cite{19, 22}. In Figure 2, they are marked by a star symbol (*).

From Figure 2, the best result is obtained when we use 10-layer CNN with IS10 feature set which is better than SMO by 2.4 points. Furthermore, we can see that the use of CNN with both IS11 and IS12 could not yield better result compared to using SMO while it can improve the performance of using IS09 and IS10. We can also see that the IS10 feature set outperformed the rest of the feature sets when we use CNN.

When we compare between IS09 and IS10, we can see that IS10 covers more features in the paralinguistic aspect of speech as it was used as the age-gender and level-of-interest classification. The performance of the feature sets IS09 was worse than that of IS10 especially when the subjects did not have any specific emotion (neutral). Some emotions appeared at the beginning before the subject begins to describe the image, in the middle when the subject begins to confuse with the things he/she wants to describe (laughs), and at the end after completing the task.

Noticeable differences between IS09 and IS10 include the use of voicing probability in which might more represent the sound-silence pattern in the subjects. Further, jitter and shimmer in IS10 might give more representation of the hesitation rate in the subjects. Those LLDs also appear in both IS11 and IS12. However, the higher dimension of the two feature sets might be too big for the input of the CNN.

Next, we investigated the effectiveness of the gating mechanism. The experiment was carried out by using IS10 feature set. The comparison of the standard CNN and the gating mechanism is shown in Table 2. From the table, we can see that the employment of linear gate improved the average accuracy from the 10-fold cross-validation scheme into 73.6%. We can see that the gating mechanism yields better results.

Furthermore, we also present the confusion matrix of the best model showing the aggregated values from the ten folds in Table 1, both from the utterance-level and the subject-level. The confusion matrix shows the balance performance in predicting both AD and Control subject. We can also see that predicting the individual utterance of the subject can be considered as a difficult task. However, combining several utterances for one subject can improve the subject-level prediction result.

Lastly, we investigated the importance of utterance length information in the classification performance. We tried a set of different segment length $L$, which are 500ms, 1000ms, 2000ms, 4000ms, and also 4295ms which is the segment length of the longest utterance in the dataset. In this case, we segmented each subject data into segments with a predetermined length $L$ without taking into consideration the oracle utterance length. Accordingly, the zero padding is added only for the last utterance segment of the subject if its length is less than $L$. The experiment was carried out by using the best scheme from the previous experiment, which is GCNN. We only tried to use the gated CNN with the number of layers as 6, 8, and 10.

As depicted in Figure 3, we can see that shorter segment length yields worse results. However, we can still get close results using the utterance length of 4000ms (69.1%, 70.8%, and 69.8% for 6, 8, and 10 layers respectively) compared to the cases when we use the oracle utterance length (72.2%, 73.6%, and 69.8% for 6, 8, and 10 layers respectively). For 10 layers, we got the similar result between using the utterance length of 4000ms and the oracle utterance length with 69.8%. This result suggests that we can use the approach even if we do not have the transcription.

7. Conclusion

We present our study in the non-linguistic approach for detecting AD by utilizing only the speech audio data. The employment of GCNN yielded the best result of 73.6%. Since it does not utilize linguistic information, we can easily apply it to low-resource languages.

There are still a lot of remaining tasks. In the near future, we will evaluate our current approach on different languages especially on low-resource languages. The study on how to early detect AD patients is also a major concern. Other possible directions include estimating the severity of the disease and also evaluating its temporal change.

8. Acknowledgements

This work was supported by JSPS KAKENHI 16H02845 and by JST CREST Grant Number JPMMCR1687, Japan.
9. References


