T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題 | |
| Title | Speech Paralinguistic Approach for Detecting Dementia Using Gated Convolutional Neural Network |
| 著者 | R Makiuchi Mariana, Warnita Tifani, 井上 中順, 篠田 浩一 |
| Authors | Mariana RODRIGUES MAKIUCHI, Tifani WARNITA, Nakamasa INOUE, Koichi SHINODA, Michitaka YOSHIMURA, Momoko KITAZAWA, Kei FUNAKI, Yoko EGUCHI, Taishiro KISHIMOTO |
| 出典 | IEICE TRANSACTIONS on Information and Systems, Vol. E104-D, No. 11, pp. 1930-1940 |
| Citation | IEICE TRANSACTIONS on Information and Systems, Vol. E104-D, No. 11, pp. 1930-1940 |
| 発行日 / Pub. date | 2021, 11 |
| 権利情報 / Copyright | 本著作物の著作権は電子情報通信学会に帰属します。<br>Copyright(c) 2021 IEICE |

Powered by T2R2 (Science Tokyo Research Repository)

IEICE TRANSACTIONS

# on Information and Systems

**VOL. E104-D NO. 11**
**NOVEMBER 2021**

**The usage of this PDF file must comply with the IEICE Provisions on Copyright.**
**The author(s) can distribute this PDF file for research and educational (nonprofit) purposes only.**
**Distribution by anyone other than the author(s) is prohibited.**

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY
The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

IEICE TRANS. INF. & SYST., VOL.E104–D, NO.11 NOVEMBER 2021

PAPER

# Speech Paralinguistic Approach for Detecting Dementia Using Gated Convolutional Neural Network

**Mariana RODRIGUES MAKIUCHI**[†], **Tifani WARNITA**[†], **Nakamasa INOUE**[†], **Koichi SHINODA**[†], **Michitaka YOSHIMURA**[††], **Momoko KITAZAWA**[††], **Kei FUNAKI**[††], **Yoko EGUCHI**[††], ***and*** **Taishiro KISHIMOTO**[††a)], ***Nonmembers***

**SUMMARY** We propose a non-invasive and cost-effective method to automatically detect dementia by utilizing solely speech audio data. We extract paralinguistic features for a short speech segment and use Gated Convolutional Neural Networks (GCNN) to classify it into dementia or healthy. We evaluate our method on the Pitt Corpus and on our own dataset, the PROMPT Database. Our method yields the accuracy of 73.1% on the Pitt Corpus using an average of 114 seconds of speech data. In the PROMPT Database, our method yields the accuracy of 74.7% using 4 seconds of speech data and it improves to 80.8% when we use all the patient's speech data. Furthermore, we evaluate our method on a three-class classification problem in which we included the Mild Cognitive Impairment (MCI) class and achieved the accuracy of 60.6% with 40 seconds of speech data.

***key words:*** *convolutional neural network, dementia detection, gating mechanism*

## 1. Introduction

Dementia is an umbrella term for a group of medical signs and symptoms associated with the cognitive-related deficiency due to damage in neurons [1]. Types of dementia include Alzheimer's disease (AD), vascular dementia, dementia with lewy body (DLB) and frontotemporal lobar degeneration (FTLD). Dementia have various characteristics representing cognitive dysfunction such as poor narrative memory when recalling experiences [2] as well as difficulties in making plans, solving problems, and completing daily tasks [1].

The increasing number of people living with dementia has gained a lot of attention. AD, which takes the biggest proportion of dementia, has become the 6$^{th}$ leading cause of death in the United States of America [3]. Moreover, according to the World Health Organization [4], in 2015, dementia affected 47 million people worldwide and it is estimated that, by 2050, this number will be nearly triplicated.

Unfortunately, there is no clear protocol on how to detect dementia in an accurate and effective manner [3]. The most common approach is to perform various clinical assessments of the patients such as examining their medical history, conducting cognitive tests (e.g., memory tasks, executive function tasks, picture description tasks, naming tasks), assessing their mood and mental status, as well as performing brain imaging; i.e., computerized tomography (CT), magnetic resonance imaging (MRI), single-photon emission computed tomography (SPECT), positron emission tomography (PET), and blood/cerebrospinal fluid testing.

The careful diagnosing process can be invasive, time-consuming and costly. However, early cognitive deficiency treatments can help patients preserve their cognitive functions [5] as some causes of dementia are remediable in early stages [6]. Thus, faster and more cost-effective dementia detection approaches have been strongly demanded.

Most approaches for the automatic dementia detection relied on linguistic information [7]–[11] since cognitive dysfunctions in patients typically appear as linguistic impairments. While these methods are effective, their major drawback is the requirement of transcriptions of patient's speech. Manual transcription is costly, and Automatic Speech Recognition (ASR) is often erroneous.

In order to address this issue, we propose a dementia detection method that relies on speech audio data only. Moreover, since geriatric clinical assessment presents several challenges [12] and patients may feel fatigued by it, this work focuses on using less patient speech data as possible to make the diagnosis in real time and physically less demanding for the patients.

We employ Gated Convolutional Neural Networks (GCNN) in order to capture the temporal pattern in the extracted features. We chose the GCNN architecture due to its superior performance in several tasks [13], [14], including tasks with limited amount of data [15]. As an extension of our previous work [16], we evaluate our method on two datasets, the DementiaBank Pitt Corpus (English) and the PROMPT Database (Japanese) collected by our own.

We further explore the detection of Mild Cognitive Impairment (MCI) patients. MCI is the stage of cognitive impairment between the expected cognitive decline of normal ageing and early dementia [17]. MCI is characterized by a cognitive decline that is greater than the age-related expectation, but that cannot be defined as dementia yet [18].

The experimental results demonstrate the effectiveness of our method in terms of cost and time required for each prediction.

Manuscript received October 1, 2020.
Manuscript revised July 16, 2021.
Manuscript publicized August 3, 2021.

[†]The authors are with Department of Computer Science, School of Computing, Tokyo Institute of Technology, Tokyo, 152–8550 Japan.
[††]The authors are with Department of Neuropsychiatry, Keio University School of Medicine, Tokyo, 160–8582 Japan.
a) E-mail: taishiro-k@mti.biglobe.ne.jp

DOI: 10.1587/transinf.2020EDP7196

Copyright © 2021 The Institute of Electronics, Information and Communication Engineers

## 2. Related Works

### 2.1 Dementia Assessment

Various evaluation methods have been defined to diagnose dementia in people. In the medical field, commonly used approaches are the Clinical Dementia Rating (CDR) [19], the Clock Drawing Test (CDT) [20], the Neuropsychiatric Inventory (NPI) [21] and the Mini-Mental State Examination (MMSE) [22]. CDR uses an interview protocol to assess dementia severity as mild or severe while, in CDT, the patients are asked to draw a clock with a specific time to get a score, which can aid the diagnosis of neurological disorders. The NPI test assess the disruptions of several behavioural functioning.

The MMSE is an extensively-used screening test that quantifies patients' cognitive function as the total score of a series of questions and problems [23]. The test itself is designed to aid the dementia diagnosis. The MMSE is used in this work and, based on [24]–[27], we define the score ranges of 0–23, 24–26 and 27–30 to respectively represent dementia, MCI and healthy categories. While most of the current medical research works present similar cut-off points to represent these classes, the definition of those score ranges has not been standardized [28]–[30]. In addition, there is not a consensus about the definition and diagnosis of MCI, since its symptoms are vast and subtle [27].

### 2.2 Features

Several types of features can be used to identify dementia. In this section, image-based features and linguistic-based features will be presented, as well as acoustic features obtained from the patient's speech, which are the features used in this work.

#### 2.2.1 Image

Structural brain images from Magnetic Resonance Imaging (MRI) can be used to identify AD patients [31]. Brain imaging plays an important role in neurodegenerative disorder detection because it provides useful information regarding anatomical changes in patients' brain [32]. The combination of MRI and fluorodeoxyglucose-positron emission tomography (FDG-PET) images was used to identify MCI patients who would further progress to dementia [33]. However, despite their effectiveness, medical image acquisition is costly and not easily accessible.

To address this issue, some works have proposed non-invasive image-based approaches to detect dementia. Work [34] proposed to detect dementia with an SVM from gait information recorded with a Kinect sensor. In their recordings, the participants walked on place and simultaneously performed a cognitive task. Other works considered computer vision techniques to detect cognitive decline from facial expressions [35]. Although not as accurate as the MRI-based approaches, these non-invasive methods show promising results that could be further improved with a combination of data modalities (e.g., gait, face frames, speech and text).

#### 2.2.2 Linguistic Data

Language deficiency becomes a prominent and perceivable symptom of dementia patients. Several syntactic, lexical and $n$-gram features were used for detecting AD on the DementiaBank Pitt Corpus [7], [9]. [10] used $n$-gram and MMSE score correlation analysis. More recently, [36] defined a hybrid RNN-CNN architecture with an attention mechanism to detect AD from Pitt Corpus' transcriptions' textual embeddings. Work [37] proposed an LSTM-based neural network language model whose prediction is calculated from their model's perplexity.

Several other works have studied the combination of linguistic and acoustic features. [38] combined features inspired in the conversation analysis of clinical interviews, lexical information extracted with an ASR and acoustic features. They further input these features to a support vector machine (SVM) to classify the patients of their own dataset into dementia or functional memory disorder. [39] and [40] extracted phonetic-based features with an ASR in order to detect MCI patients from their speech. [8] fused transcription-based linguistic features with acoustic features such as Mel-frequency cepstral coefficients (MFCC). [41] utilized the combination of speech duration, pause-related features, pitch-related features and other prosodic features, as well as linguistic features acquired from a customized ASR adapted for dementia patients.

Even though those approaches have shown good results, most of them are limited by the availability of transcription data and/or ASR, which often has poor performance due to the degraded speech intelligibility of the patients.

### 2.3 Speech

Several works proposed the dementia identification from ASR-independent speech features. Features such as silence ratio were found to be more meaningful than other linguistic features when applied to an SVM classifier [42]. Moreover, the usage of acoustic and context-free linguistic features to classify patients showed promising results on the Carolina Conversations Collections dataset [43].

Besides having problems with language deficits, people with AD, specially in the early stages of the disease, might become apathetic and have a tendency to get depressed [1]. People with AD usually suffer from prosodic impairment due to which they will find difficulties in expressing their emotions [44]. Those signs suggest the presence of paralinguistic cues in the speech of people who suffer from this cognitive dysfunction.

OpenSMILE [45] is a commonly used tool for feature extraction in speech tasks [46], [47]. It describes a series of

default feature sets, such as the INTERSPEECH 2010 Paralinguistic Challenge Feature Set (IS10) [47], which is used in this work. The IS10 defines 76 Low-Level Descriptors (LLD) features for each time frame. In this work, we define a time frame as 25ms, sampled at a rate of 10ms. Those LLD are a combination of several speech descriptors that were independently used in previous works, such as pitch, voicing probability [42] and MFCC [8]. When compared to other feature sets defined in OpenSMILE, the IS10 yielded the best result in the AD detection task [16].

### 2.4 Classifiers

Support Vector Machine (SVM) classifiers [48] were widely used as the baseline method of several paralinguistic tasks, such as emotion recognition [46] and age-gender classification [47]. While training the SVM, the Sequential Minimal Optimization (SMO) [49] algorithm is used. The SMO solves the Quadratic Programming (QP) optimization in the SVM by dividing the QP into the smallest possible QP subproblems, allowing the SMO to handle large amounts of data.

Recently, deep learning-based approaches have become extremely popular due to their success in a wide range of tasks. [50] applied Convolutional Neural Networks (CNN) [51] to speech emotion recognition. For the same task, a Recurrent Neural Network (RNN) [52] was added on top of CNN layers [53] to capture the speech's dynamic features.

RNNs can accommodate the temporal pattern change, but they require a long training time [54] and a large amount of training data. On the other hand, CNNs need little training data compared to other existing networks [55] due to their reduced number of connection weights. Moreover, even without any explicit sequential mechanism, CNNs are still able to model the temporal context in the data by means of their convolution operations [54].

There have been various studies that incorporated gating mechanisms to convolution layers achieving state-of-the-art performance on tasks such as conditional image modelling [56], language modelling [13], speech synthesis [14] and generative image inpainting [57]. When applied to RNNs, such as Long Short-Term Memories (LSTM) [58], gating mechanisms were shown to be effective in handling the long-term dependencies problem. Gating mechanisms in CNNs can be used to manage the information flow as well as to mitigate the vanishing gradient problem [13]. Moreover, it was shown that the combination of gating mechanisms and CNNs can achieve superior performance in tasks with limited amount of data, such as speech recognition for low-resource languages [15] and speech keyword spotting [59]. Therefore, inspired by these advantages and the effectiveness of the combination of CNNs and gating mechanisms applied to different tasks, we hypothesize that the automatic dementia detection can also benefit from this architecture.

## 3. Gated Convolutional Neural Network

A Gated Convolutional Neural Network (GCNN) consists of convolution layers and gating mechanisms. In our case, each convolution layer is expected to extract the salient information from the combined LLD features for every short period of time. Thus, the temporal pattern change will be encapsulated within the combination of several extracted patches of features.

The convolution operation "slides" a kernel $k$ over the input features to extract their prominent cues. In our work, since we want to model the correlation between all the LLD features, captured at each time frame, we use the one-dimensional (1D) CNN, hence each kernel slides only in the time axis, as represented in Fig. 1. This network is also referred to as Time-Delay Neural Network (TDNN) [60].

The gating mechanism applies this convolution operation to the input in two different paths, as shown in Fig. 2. The gate in this network controls the information flow between succeeding layers, hence preventing the vanishing gradient problem.

Following the Gated Linear Unit (GLU) architecture proposed by [13], we feed a speech feature matrix $X \in \mathbb{R}^{F \times T}$ into our network, in which $F$ and $T$ are the dimension of the

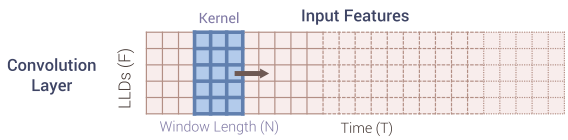


**Fig. 1** A convolution layer over LLD features extracted with the openSMILE toolkit.

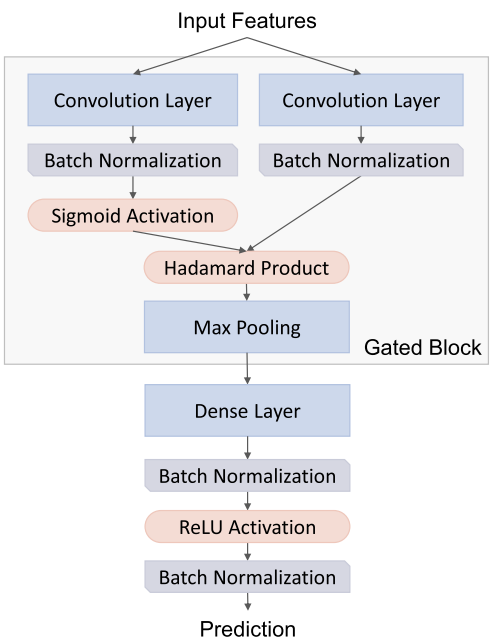


**Fig. 2** A GCNN with one gated block. A deeper network can be made by stacking gated blocks.

LLD features and the number of time frames, respectively. We further convolve these input features with a kernel of dimension $F \times N$, in which $N$ is the length of the kernel in the time axis. At each convolution operation between this kernel and the input, a scalar output is produced. The result of the gated convolution operation before the max pooling for the kernel $k$ at position $i$ is given by

$$y_{k,i} = \left( \sum_{f=1}^{F} \sum_{n=1}^{N} v_{(k)_{f,n}} x_{f,i-n} + e_{(k)} \right) \cdot g\left( \sum_{f=1}^{F} \sum_{n=1}^{N} w_{(k)_{f,n}} x_{f,i-n} + b_{(k)} \right), \quad (1)$$

in which $v_{(k)_{c,d}}$ is the element of kernel $k$'s matrix $V$ at position $(c, d)$ and $w_{(k)_{c,d}}$ is the element of kernel $k$'s matrix $W$ at position $(c, d)$. $V \in \mathbb{R}^{F \times N}$ and $e \in \mathbb{R}$ are the linear gate kernel weight matrix and bias respectively (i.e., they represent the convolution operation in the right stream of the gated block in Fig. 2), $W \in \mathbb{R}^{F \times N}$ and $b \in \mathbb{R}$ are the respective weight matrix and the bias of the convolutional operation in the left stream of Fig. 2 and $g$ is the sigmoid function.

In Eq. (1), both summations enclosed by parenthesis represent a convolution operation that results in one scalar, and the term to which the sigmoid function $g$ is applied is the gate operation that controls the linear gate output.

The resulting $Y$ matrix formed by the elements $y_{k,i}$ will have the dimensions $K \times M$, in which $K$ is the number of kernels and $M$ is the output segment length. After the convolution operation, the matrix $Y$ has its length halved in the time-axis by the max-pooling layer [61] to get its most significant information and reduce its dimensionality.

Figure 2 shows our GCNN with a single gated block, which is formed by two convolution layers and one max-pooling layer. A deeper GCNN would consist of multiple stacked gated blocks.

The output of the network's last gated block, $Y' \in \mathbb{R}^{K' \times M'}$, with $K'$ as the number of kernels of the last gated block and $M'$ as the final output segment length, is then flattened into one feature vector $Z \in \mathbb{R}^{O}$, in which $O = K'M'$. This vector is input to a fully-connected (dense) layer with the ReLU activation function. We also apply batch normalization [62] at the end of each dense and convolution layer.

## 4. Experiments

### 4.1 Evaluation Metrics

We use classification accuracy as the main evaluation metric in our experiments. This reflects previous works on the Pitt Corpus dataset [8], [10] and on the other related datasets [41], [42]. We compute the accuracy averaged over the 10-fold cross-validation results. At each fold, we partition the dataset in 10 subsets, from which we select one for testing and the remaining for training. We design these subsets so that no subject's data appears in both training and testing.

### 4.2 Configuration

We split the patient's speech of each interview session into several segments of length $L$, zero-padding the speech segments that are shorter than $L$. In this work, a speech segment is defined as a short slice of the patient's speech. In order to obtain the speech segments, we first concatenate all the patient's speech utterances in an interview session. Then, we segment this combined patient's speech into consecutive and non-overlapping speech segments, which are applied as input to our model. On the Pitt Corpus, the patient's speech is extracted by using the speaker turns information. On the PROMPT Database we separate the patient and doctor speech by applying the Cross-Channel Spectral Subtraction method [63] followed by a Voice Activity Detection (VAD) approach. We classify each speech segment using our Gated Convolutional Neural Network architecture, and, after aggregating the scores from multiple segments, we conduct a majority voting to determine the session-level dementia classification.

In our binary GCNN, we consistently use the window length $N = 2$ and the kernel size $K = 64$ in every convolution operation of each gated block. We have tested our model with 6, 8 and 10 stacked gated blocks. The dense layer after the last gated block has 256 hidden neurons. We apply 0.5 dropout before the output layer for regularization. The output layer consists of one neuron with a sigmoid function. We trained the network using each segment's IS10 LLD features and their corresponding binary label (i.e., healthy or AD) on a 10-fold cross-validation scheme.

We used binary cross-entropy as the loss function and the Adam [64] optimizer with learning rate equal to $10^{-3}$ and exponential decay rate respectively defined as 0.9 and 0.999 for the first and second moment estimates. A batch size equal to 32 was consistently used over all the experiments and the input $X \in \mathbb{R}^{F \times T}$ is composed of 76 LLD features $F$ per time frame and 397 time frames $T$.

### 4.3 Datasets

In this work, we use two datasets containing the speech of people with and without dementia: the Pitt Corpus [65], in which the subjects speak English, and the PROMPT Database, in which the subjects speak Japanese.

#### 4.3.1 Pitt Corpus

The Pitt Corpus, a part of the DementiaBank, contains speech data and the corresponding transcription information of healthy people (Control group) and of people with Alzheimer's disease (AD group) speaking in English. The audio files in this dataset contain speech from clinicians and patients. We apply three constraints to select data from this dataset.

First, we only consider the data drawn from the picture description task. This task is considered an approximation

of real-life spontaneous conversations [66], in which the subjects are asked to describe the Cookie Theft Picture of the Boston Diagnostic Aphasia Examination [67].

Second, from the AD group, we select the sessions that correspond to patients with a diagnosis of either AD or probable AD. There are no specific restrictions to select sessions from the control group. It should be noted that, even though we select sessions using the same method as in [8], [10], and [36], the number of sessions is slightly different from those works since the dataset was modified over time.

Third, we only select sessions with both the audio and the transcription information. The transcripts provide the speaker turns information. We use them to remove the clinician's speech in the data preparation step.

As a result, the data we use comprises 488 sessions (255 dementia, 233 healthy), with an average duration of 114 seconds, recorded from 267 participants (169 dementia, 98 healthy).

We perform three preprocessing stages on the data. First, we normalize each audio signal using the average value of decibels relative to full scale (dBFS) in the data. Then, we use the speech turns information available in the dataset transcriptions to directly extract the participant's speech segments. Each segment corresponds to a participant's speech turn during the interview, thus obtaining a total of 6,267 segments (3,276 dementia, 2,991 healthy). Finally, we extend the duration of these segments by 10ms at the beginning and 10ms at the end as an attempt to mitigate speech discontinuities due to imprecisions in the turns information. In these preprocessing stages, we only use the transcripts to extract the speaker turns information, disregarding the transcript's content.

The audio files in the Pitt Corpus are single channelled (mono), sampled at a frequency of 44.1kHz and stored as PCM encoded wave files.

#### 4.3.2 PROMPT Database

The PROMPT database is part of a larger project of Keio University School of Medicine: the Project for Objective Measures Using Computational Psychiatry (PROMPT)[†] [68].

All the patients have given their written consent before participating in the study and, in cases in which patients were judged to be decisionally impaired, the patients' guardians provided consent. Participants were able to leave the study at any time.

In this work, we use the PROMPT Database collected from May 2, 2016 to March 31, 2019 at seven hospitals and three outpatient clinics in five different Japanese prefectures.

Speech data were recorded when the participant had free-discussion and performed several clinical tasks with trained research psychiatrists and/or psychologists. The session interviews were recorded from two synchronized microphones: one positioned close to the participant and the other placed near the clinician. The session recordings were acquired under various unconstrained acoustic conditions (i.e., with different microphones and in rooms with different reverberation characteristics).

In this work, as discussed in Sect. 2.1, we categorize cognitive impairment based on the MMSE score, and dementia, MCI and healthy classes are defined as a MMSE score in the respective ranges of 0-23, 24-26 and 27-30. The inter-rater reliability for the MMSE score was examined and evaluated in terms of interclass correlation coefficient (ICC). The ICC assesses the consistency in the score annotations made by different raters. For the MMSE annotation in the PROMPT Database, the ICC is 0.996 (95% CI=0.990-0.999, $p < 0.01$).

The PROMPT data used in this work comprises 496 session recordings (153 dementia, 111 MCI and 232 healthy) with an average duration of 1,487 seconds, from 163 participants (49 dementia, 42 MCI and 72 healthy).

Since the PROMPT Database collects the audio recordings from two synchronized microphones, we adopted the Cross-Channel Spectral Subtraction (CCSS) method [63] followed by a Voice Activity Detection (VAD) approach to extract the patient-only speech. The CCSS is a source separation method able to, in a meeting setting in which one microphone is prepared to each speaker, estimate the speech of a given speaker by suppressing other speakers' speech. We utilize the CCSS to suppress the doctor's speech. The resulting speech signal is then applied to a VAD, which can detect and segment the patient's speech.

After extracting the patient's speech, we apply the three preprocessing stages described in Sect. 4.3.1 with the exception that we do not use speaker turns information to segment the patient speech.

Following the results obtained in Sect. 4.4, we divide the patient's speech into segments of 4 seconds in a fixed-length fashion, hence resulting in 184,337 speech segments (39,593 dementia, 27,234 MCI and 117,510 healthy). In addition, we adopt a GCNN with 8 gated blocks since it resulted in the best accuracy during the experiments with the Pitt Corpus.

All the resulting audio files are single channelled (mono), sampled at the frequency of 16kHz stored as PCM encoded wave files.

### 4.4 Pitt Corpus

We present the average accuracy result over the 488 selected sessions of the Pitt Corpus in Table 1. This accuracy is computed over the session-level classification, which is obtained by applying majority voting over the speech segments' predictions. We employ SMO on the IS10 features for comparison. Our method yields the accuracy of 73.1%, which outperforms the SMO result of 67.5%. In Table 1 we also

[†]On the 9th of March 2016, the PROMPT project and its medical data collection were approved by the ethics committee and the Institutional Review Board of Keio University School of Medicine and by all of the other participating facilities. PROMPT protocols have been registered with the University Hospital Medical Information Network (UMIN) (UMIN ID: UMIN000023764).

**Table 1** Comparison of dementia detection methods on the Pitt Corpus interview sessions.

| Method | Accuracy (%) |
|---|---|
| SMO baseline (Speech) | 67.5 |
| Wankerl (Linguistic) [10] | 77.1 |
| Fritsch (Linguistic) [37] | 85.6 |
| Chen (Linguistic) [36] | 97.4 |
| Fraser (Speech + Linguistic) [8] | 81.9 |
| Ours | 73.1 |

**Table 2** The confusion matrix depicting the classification results from all folds using the IS10 feature set as input and a GCNN with eight gated blocks on the Pitt Corpus.

| | | Predicted | | |
|---|---|---|---|---|
| | | Dementia | Healthy | Total |
| Actual | Dementia | 2,340 | 936 | 3,276 |
| | Healthy | 1,213 | 1,778 | 2,991 |
| | Total | 3,553 | 2,714 | Accuracy 65.7% |

(a) Segment-Level Classification

| | | Predicted | | |
|---|---|---|---|---|
| | | Dementia | Healthy | Total |
| Actual | Dementia | 189 | 66 | 255 |
| | Healthy | 65 | 168 | 233 |
| | Total | 254 | 234 | Accuracy 73.1% |

(b) Session-Level Classification

report the accuracy of methods that rely on linguistic features and on the combination of linguistic and speech features, which respectively result in the best accuracies of 97.4% and 81.9%. Although our method has a worse performance than the linguistic features-based works presented in Table 1, it does not require ASR or transcription information, hence being more cost-effective and more appropriate to fast diagnosis.

Table 2 shows the confusion matrix of our best model on the Pitt Corpus for the aggregated values from the ten folds both from the speech segment-level and the session-level predictions. The model is composed of eight gated blocks and it uses the speech segments obtained from the turns information in the Pitt Corpus. The confusion matrices show that predicting from a single speech segment in a session is a difficult task since the amount of information in one segment might be too limited. Thus, combining several segments for one session improves the prediction result.

Although our method does not depend on linguistic features, we still require the speaker turns information to segment the patient speech. In order to investigate our model's performance when the speech is partitioned into fixed-length segments and to further determine the shortest speech segment length that allows accurate data classification, we concatenate each patient's speech turns segments and we divide this concatenated speech into fixed-length segments of duration $L$. Each segment is input to the model and we apply majority voting over the segments' predictions to obtain the session-level classification. The experiment is carried out using GCNNs with a number of gated blocks

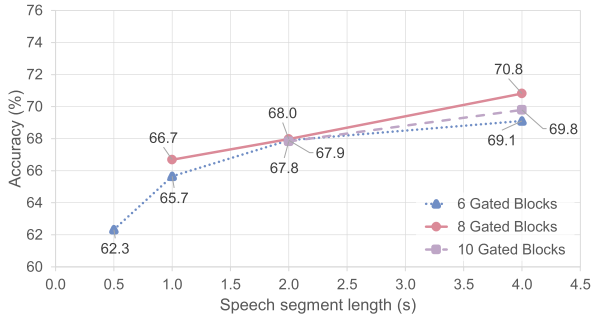


**Fig. 3** Accuracy of GCNNs with different speech segment length on the Pitt Corpus

equal to 6, 8 and 10 and durations $L$ of 0.5 s, 1 s, 2 s and 4 s. The results are summarized in Fig. 3.

Figure 3 shows that using only 4 seconds speech segments yields results almost as good as applying the segmentation based on the speaker turns information, which suggests that there exist discriminative dementia cues in a short duration of speech data. In addition, the results in Fig. 3 indicate that segmenting the subject's voice in the middle of their speech does not significantly degrade the performance.

### 4.5 PROMPT Database

We evaluate our approach on the binary classification (i.e., dementia *vs* healthy) obtaining the average session-level accuracy of 80.8%. The session-level prediction is computed as the majority voting result over all 4-seconds speech segments' predictions within a session. It should be noted that the sessions in the PROMPT Database have a longer duration compared to the Pitt Corpus, hence the accuracy is higher on the PROMPT Database.

We further evaluate our model's performance over speech intervals of different durations. Taking the 4-second fixed segment length as our duration unit, we experiment with different duration configuration for each session data, which are 4 seconds, 8 seconds, 20 seconds, 40 seconds, 1 minute, 5 minutes and all of the session speech data (i.e., the session-level prediction). In all cases, our model performs the classification over each 4-seconds speech segment. For duration configurations longer than 4 seconds, we apply majority voting to determine the prediction for each speech interval.

We report our results in accuracy in Fig. 4. The figure shows that performance degrades if we apply shorter speech durations. However, we obtain the average accuracy of 77.1% by using only 20 seconds of data for each session and the average accuracy of 74.7% when we use only 4 seconds of data. Although there is still room for improvement and our model alone cannot be used in practical dementia detection applications yet, this result represents an important step towards using automatic dementia detection tools in real-world diagnosis, in which, every so often, there is little available speech data. We have additionally reported the segment-level classification confusion matrix for the 4-

second segmentation of one of our folds in Table 3.

We further perform a more comprehensive experiment on the binary classification with various configurations. Apart from distinguishing dementia and healthy sessions, we perform the classification on the dementia versus non-dementia case and the healthy versus non-healthy case by respectively adding the participants with a MMSE score in the MCI range to the healthy and to the dementia groups. Table 4 depicts these configurations in the column "Conditions" and it reports the results in terms of accuracy for these configurations and different speech durations.

In terms of session classification accuracy, adding the MCI class yields worse performance on both Condition D *vs* M + H (78.6%) and Condition D + M *vs* H (75.9%) compared to Condition D *vs* H (80.8%) as it can be seen in Table 4. Moreover, it is possible to observe that, for the experiments that include the MCI class data, the accuracy does not increase monotonously with the amount of speech data. This result suggests that we should not combine the MCI participants either with healthy or dementia participants. This might be explained from the fact that MCI participants cannot be considered healthy due to their cognitive ability decline, but MCI cannot be framed as dementia either, since this decline is less severe compared to dementia. It is also interesting to see that combining MCI with dementia patients yields worse performance than combining MCI with healthy subjects. Further investigation on the closer relation between MCI patients and healthy subjects might be needed based on this result.

Finally, we also conduct the three-class classification to distinguish between sessions with dementia, MCI and healthy subjects. The results are reported in Table 5. We obtain the average accuracy of 65.0% using 4 seconds of session data. While MCI patients might present subtle different visible characteristic from healthy or dementia patients, they are very different in actual. Detecting MCI patients is important for the early prediction of dementia but it is difficult due to the nearly ambiguous nature of the data and the lack of a medical standard to classify MCI patients, as discussed in Sect. 2.1.

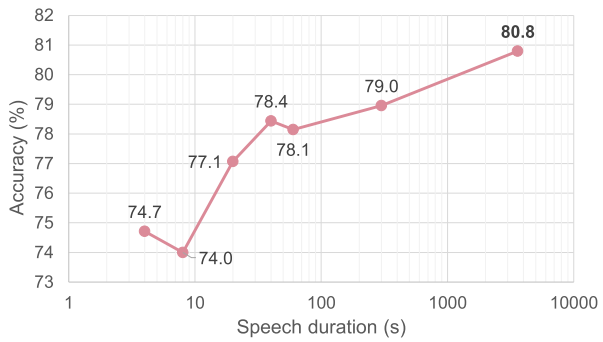


**Fig. 4** Accuracy of GCNN with different duration of data used for each session on the PROMPT Database. The horizontal axis is presented in logarithmic scale.

**Table 3** Segment-level confusion matrix of one of the folds on the PROMPT database for the classification based on 4 seconds long speech segments.

| | | Predicted | | |
|---|---|---|---|---|
| | | Dementia | Healthy | Total |
| Actual | Dementia | 3,882 | 380 | 4,262 |
| | Healthy | 3,227 | 7,418 | 10,645 |
| | Total | 7,109 | 7,798 | Accuracy 75.80% |

## 5. Discussion

Our experiments suggest that our model's accuracy has a tendency to improve as we obtain more training data. Moreover, when compared to conventional diagnosis methods based on biomarkers, our method has the advantage of being simpler and less costly. However, as shown in Table 4 and as mentioned in Sect. 4, our model cannot be used alone to diagnose dementia in practical applications because its accuracy is still not high enough [69], [70].

We believe that the best solution to this problem would be to combine our speech-based approach with methods that use other data modalities, in a multimodal manner. In fact, works in similar fields, such as depression detection, showed the advantages of multimodal approaches [71]. By combining facial expression and body movement informa-

**Table 4** Experimental results on the PROMPT Database for different experiment conditions for the binary classification. The dementia, MCI and healthy classes are respectively represented by the characters 'D', 'M' and 'H'.

| **Conditions** | **Accuracy (%)** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **4 sec** | **8 sec** | **20 sec** | **40 sec** | **1 min** | **5 min** | **all data** |
| D *vs* M + H | 73.8 | 74.3 | 77.9 | **78.6** | 78.3 | 77.0 | 77.6 |
| D + M *vs* H | 71.3 | 70.9 | 73.6 | 74.0 | 74.1 | **75.9** | 74.3 |
| D *vs* H | 74.7 | 74.0 | 77.1 | 78.4 | 78.1 | 79.0 | **80.8** |

**Table 5** Experimental results on the PROMPT Database for the three-class classification of D *vs* M *vs* H. The results are reported in accuracy (%).

| **4 sec** | **8 sec** | **20 sec** | **40 sec** | **1 min** | **5 min** | **all data** |
|---|---|---|---|---|---|---|
| 65.0 | **61.1** | 58.4 | 60.6 | 59.2 | 57.7 | 58.3 |

tion from the interview recordings, for example, with our speech-based approach, we believe that the accuracy could be improved, while keeping the method non-invasive.

Another way of improving our model's accuracy is to collect more speech data. In both ways (i.e., multimodal and using more data), the cost in terms of time, patient extenuation and health care professionals' effort, is much less than the invasive methods. For example, it is possible to collect data by monitoring patients' voices with mobile phones or wearable devices, and this data collection has a cost that is much smaller than that of the clinical assessment.

As the number of dementia patients rapidly increase, a two-step diagnosis, first with an automatic low-cost diagnosis method and second with health care professionals and medical treatment, is a promising future direction.

## 6. Conclusion

We present a method for dementia detection solely based on speech data. Using a GCNN architecture on top of the IS10 paralinguistic feature set yields the best accuracy of 73.1% in an English dataset, the Pitt Corpus, and 80.8% in a Japanese dataset, the PROMPT Database. We achieve the accuracy of 77.1% by using only 20 seconds of data on the PROMPT Database and 74.7% when we consider only 4 seconds of data. These results show our model's capability of making predictions with a reduced amount of data, which is important for real-world dementia diagnosis. We further perform the three-class classification of dementia, MCI and healthy subjects on the PROMPT Database, which yields the accuracy of 60.6%.

## Acknowledgements

This work was supported by JST CREST [grant numbers JPMJCR1687, JPMJCR19F5]; JSPS KAKEN [grant number 16H02845], and the Japan Agency for Medical Research and Development (AMED) [grant number JP18he1102004].

### References

[1] Alzheimer's Association, "2018 Alzheimer's disease facts and figures," Alzheimer's & Dementia, vol.14, no.3, pp.367–429, 2018.
[2] E.T. Prud'hommeaux and B. Roark, "Extraction of narrative recall patterns for neuropsychological assessment," Proc. Interspeech 2011, pp.3021–3024, 2011.
[3] Alzheimer's Association, "2017 Alzheimer's disease facts and figures," Alzheimer's & Dementia, vol.13, no.4, pp.325–373, 2017.
[4] World Health Organization, Global action plan on the public health response to dementia: 2017-2025, World Health Organization, 2017.
[5] H. Martono, Buku Ajar Boedhi-Darmojo: Geriatri (Ilmu Kesehatan Usia Lanjut) [Boedhi-Darmojo Textbook: Geriatrics (Elderly People Health Science)], Balai Penerbit FKUI, 2000.
[6] T. Manjari, V. Deepti, et al., "Reversible dementias," Indian Journal of Psychiatry, vol.51, no.5, pp.52–55, 2009.
[7] V.C. Zimmerer, M. Wibrow, and R.A. Varley, "Formulaic language in people with probable Alzheimer's disease: a frequency-based approach," Journal of Alzheimer's Disease, vol.53, no.3, pp.1145–1160, 2016.
[8] K.C. Fraser, J.A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," Journal of Alzheimer's Disease, vol.49, no.2, pp.407–422, 2016.
[9] S.O. Orimaye, J.S. Wong, K.J. Golden, C.P. Wong, and I.N. Soyiri, "Predicting probable Alzheimer's disease using linguistic deficits and biomarkers," BMC bioinformatics, vol.18, no.1, p.34, 2017.
[10] S. Wankerl, E. Nöth, and S. Evert, "An *N*-gram based approach to the automatic diagnosis of Alzheimer's disease from spoken language," Proc. INTERSPEECH 2017, pp.3162–3166, 2017.
[11] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Detecting signs of dementia using word vector representations," Interspeech, pp.1893–1897, 2018.
[12] I. Singh, "Assessment and management of older people in the general hospital setting," Challenges in Elder Care, vol.37, 2016.
[13] Y.N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," International conference on machine learning, pp.933–941, 2017.
[14] A. Van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.
[15] J. Kang, W. Zhang, and J. Liu, "Gated convolutional networks based hybrid acoustic models for low resource speech recognition," 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp.157–164, 2017.
[16] T. Warnita, N. Inoue, and K. Shinoda, "Detecting Alzheimer's disease using gated convolutional neural network from audio data," Proc. Interspeech 2018, pp.1706–1710, 2018.
[17] R.C. Petersen, "Mild cognitive impairment as a diagnostic entity," Journal of internal medicine, vol.256, no.3, pp.183–194, 2004.
[18] R.C. Petersen, G.E. Smith, S.C. Waring, R.J. Ivnik, E.G. Tangalos, and E. Kokmen, "Mild Cognitive Impairment: Clinical characterization and outcome," Archives of neurology, vol.56, no.3, pp.303–308, 1999.
[19] J.C. Morris, "The clinical dementia rating (cdr): Current version and scoring rules," Neurology, vol.43, no.11, pp.2412–2414, 1993.
[20] T. Sunderland, J.L. Hill, A.M. Mellow, B.A. Lawlor, J. Gundersheimer, P.A. Newhouse, and J.H. Grafman, "Clock drawing in alzheimer's disease: a novel measure of dementia severity," Journal of the American Geriatrics Society, vol.37, no.8, pp.725–729, 1989.
[21] J.L. Cummings, M. Mega, K. Gray, S. Rosenberg-Thompson, D.A. Carusi, and J. Gornbein, "The neuropsychiatric inventory: comprehensive assessment of psychopathology in dementia," Neurology, vol.44, no.12, p.2308, 1994.
[22] M.F. Folstein, S.E. Folstein, and P.R. McHugh, "'Mini-mental state": a practical method for grading the cognitive state of patients for the clinician," Journal of psychiatric research, vol.12, no.3, pp.189–198, 1975.
[23] V.C. Pangman, J. Sloan, and L. Guse, "An examination of psychometric properties of the mini-mental state examination and the standardized mini-mental state examination: implications for clinical practice," Applied Nursing Research, vol.13, no.4, pp.209–213, 2000.
[24] J. Saxton, L. Morrow, A. Eschman, G. Archer, J. Luther, and A. Zuccolotto, "Computer assessment of mild cognitive impairment," Postgraduate medicine, vol.121, no.2, p.177, 2009.
[25] D.I. Kaufer, C.S. Williams, A.J. Braaten, K. Gill, S. Zimmerman, and P.D. Sloane, "Cognitive screening for dementia and mild cognitive impairment in assisted living: comparison of 3 tests," Journal of the American Medical Directors Association, vol.9, no.8, pp.586–593, 2008.
[26] T.N. Tombaugh and N.J. McIntyre, "The mini-mental state examination: a comprehensive review," Journal of the American Geriatrics Society, vol.40, no.9, pp.922–935, 1992.
[27] I. Arevalo-Rodriguez, N. Smailagic, M.R. i Figuls, A. Ciapponi, E. Sanchez-Perez, A. Giannakou, O.L. Pedraza, X.B. Cosp, and

S. Cullum, "Mini-mental state examination (MMSE) for the detection of alzheimer's disease and other dementias in people with mild cognitive impairment (MCI)," Cochrane Database of Systematic Reviews, no.3, CD010783, 2015.

[28] I. Onwuekwe, "Assessment of mild cognitive impairment with mini mental state examination among adults in southeast nigeria," Annals of medical and health sciences research, vol.2, no.2, pp.99–102, 2012.

[29] S.E. O'Bryant, J.D. Humphreys, G.E. Smith, R.J. Ivnik, N.R. Graff-Radford, R.C. Petersen, and J.A. Lucas, "Detecting dementia with the mini-mental state examination in highly educated individuals," Archives of neurology, vol.65, no.7, pp.963–967, 2008.

[30] R. Kochhann, J.S. Varela, C.S. de Macedo Lisboa, and M.L.F. Chaves, "The mini mental state examination: review of cutoff points adjusted for schooling in a large southern brazilian sample," Dementia & Neuropsychologia, vol.4, no.1, p.35, 2010.

[31] S. Farhan, M.A. Fahiem, and H. Tauseef, "An ensemble-of-classifiers based approach for early diagnosis of alzheimer's disease: classification using structural features of brain images," Computational and mathematical methods in medicine, vol.2014, 2014.

[32] L. Narayanan and A.D. Murray, "What can imaging tell us about cognitive impairment and dementia?," World journal of radiology, vol.8, no.3, pp.240–254, 2016.

[33] D. Lu, K. Popuri, G.W. Ding, R. Balachandar, and M.F. Beg, "Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural mr and fdg-pet images," Scientific reports, vol.8, no.1, p.5697, 2018.

[34] K. Aoki, T.T. Ngo, I. Mitsugami, F. Okura, M. Niwa, Y. Makihara, Y. Yagi, and H. Kazui, "Early detection of lower mmse scores in elderly based on dual-task gait," IEEE Access, vol.7, pp.40085–40094, 2019.

[35] Z. Fei, E. Yang, D.D.U. Li, S. Butler, W. Ijomah, and H. Zhou, "A survey on computer vision techniques for detecting facial features towards the early diagnosis of mild cognitive impairment in the elderly," Systems Science & Control Engineering, vol.7, no.1, pp.252–263, 2019.

[36] J. Chen, J. Zhu, and J. Ye, "An attention-based hybrid network for automatic detection of alzheimer's disease from narrative speech," Proc. Interspeech 2019, pp.4085–4089, 2019.

[37] J. Fritsch, S. Wankerl, and E. Nöth, "Automatic diagnosis of Alzheimer's disease using neural network language models," Proc. ICASSP 2019, pp.5841–5845, 2019.

[38] B. Mirheidari, D. Blackburn, T. Walker, M. Reuber, and H. Christensen, "Dementia detection using automatic analysis of conversations," Computer Speech & Language, vol.53, pp.65–79, 2019.

[39] G. Gosztolya, L. Tóth, T. Grósz, V. Vincze, I. Hoffmann, G. Szatlóczki, M. Pákáski, and J. Kálmán, "Detecting mild cognitive impairment from spontaneous speech by correlation-based phonetic feature selection," Proc. Interspeech 2016, pp.107–111, 2016.

[40] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, G. Szatlóczki, E. Biró, F. Zsura, M. Pákáski, and J. Kálmán, "Automatic detection of mild cognitive impairment from spontaneous speech using asr," Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[41] R. Sadeghian, J.D. Schaffer, and S.A. Zahorian, "Speech processing approach for diagnosing dementia in an early stage," Proc. INTERSPEECH 2017, pp.2705–2709, 2017.

[42] A. Khodabakhsh, F. Yesil, E. Guner, and C. Demiroglu, "Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech," EURASIP Journal on Audio, Speech, and Music Processing, vol.2015, no.1, p.9, 2015.

[43] S. Luz, S. de la Fuente, and P. Albert, "A method for analysis of patient speech in dialogue for dementia detection," Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp.35–42, 2018.

[44] G. Tosto, M. Gasparini, G. Lenzi, and G. Bruno, "Prosodic impairment in Alzheimer's disease: assessment and clinical relevance," The Journal of neuropsychiatry and clinical neurosciences, vol.23, no.2, pp.E21–E23, 2011.

[45] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in OpenSMILE, the munich open-source multimedia feature extractor," Proc. MM'13, pp.835–838, 2013.

[46] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," Proc. INTERSPEECH 2009, pp.312–315, 2009.

[47] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," Proc. INTERSPEECH 2010, pp.2794–2797, 2010.

[48] B.E. Boser, I.M. Guyon, and V.N. Vapnik, "A training algorithm for optimal margin classifiers," Proc. fifth annual workshop on Computational learning theory, pp.144–152, ACM, 1992.

[49] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Tech. Rep. MSR-TR-98-14, Microsoft, April 1998.

[50] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," ACM Multimedia, 2014.

[51] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel, "Backpropagation applied to handwritten zip code recognition," Neural computation, vol.1, no.4, pp.541–551, 1989.

[52] J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," Proc. national academy of sciences, vol.79, no.8, pp.2554–2558, 1982.

[53] G. Keren and B. Schuller, "Convolutional RNN: an enhanced model for extracting features from sequential data," 2016 International Joint Conference on Neural Networks (IJCNN), pp.3412–3419, 2016.

[54] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[55] Y. Bengio, "Learning deep architectures for AI," Foundations and trends® in Machine Learning, vol.2, no.1, pp.1–127, 2009.

[56] A. Van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with pixelCNN decoders," Proceedings of the 30th International Conference on Neural Information Processing Systems, pp.4797–4805, 2016.

[57] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T.S. Huang, "Freeform image inpainting with gated convolution," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp.4471–4480, 2019.

[58] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol.9, no.8, pp.1735–1780, 1997.

[59] A. Coucke, M. Chlieh, T. Gisselbrecht, D. Leroy, M. Poumeyrol, and T. Lavril, "Efficient keyword spotting using dilated convolutions and gating," ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6351–6355, IEEE, 2019.

[60] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang, "Phoneme recognition using time-delay neural networks," in Readings in speech recognition, pp.393–404, Elsevier, 1990.

[61] Y. Zhou and R. Chellappa, "Computation of optical flow using a neural network," Proc. IEEE 1988 International Conference on Neural Networks, vol.27, pp.71–78, 1988.

[62] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," Proc. 32nd International Conference on Machine Learning, pp.448–456, 2015.

[63] Y. Nasu, K. Shinoda, and S. Furui, "Cross-channel spectral subtraction for meeting speech recognition," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4812–4815, IEEE, 2011.

[64] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[65] J.T. Becker, F. Boiler, O.L. Lopez, J. Saxton, and K.L. McGonigle, "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," Archives of Neurology, vol.51, no.6, pp.585–594, 1994.
[66] E. Giles, K. Patterson, and J.R. Hodges, "Performance on the Boston Cookie Theft picture description task in patients with early dementia of the Alzheimer's type: missing information," Aphasiology, vol.10, no.4, pp.395–408, 1996.
[67] E. Kaplan, H. Goodglass, S. Weintraub, and S. Brand, Boston Naming Test, Lea & Febiger, 1983.
[68] T. Kishimoto, A. Takamiya, K.C. Liang, K. Funaki, T. Fujita, M. Kitazawa, M. Yoshimura, Y. Tazawa, T. Horigome, Y. Eguchi, T. Kikuchi, M. Tomita, S. Bun, J. Murakami, B. Sumali, T. Warnita, A. Kishi, M. Yotsui, H. Toyoshiba, Y. Mitsukura, K. Shinoda, Y. Sakakibara, and M. Mimura, "The project for objective measures using computational psychiatry technology (prompt): Rationale, design, and methodology," medRxiv doi:10.1101/19013011, 2019.
[69] Z.S. Nasreddine, N.A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J.L. Cummings, and H. Chertkow, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," Journal of the American Geriatrics Society, vol.53, no.4, pp.695–699, 2005.
[70] Y. Imon, "The imon cognitive impairment screening test (icis): A new brief screening test for mild cognitive impairment or dementia," Nihon Ronen Igakkai zasshi, Japanese journal of geriatrics, vol.51, no.4, pp.356–363, 2014.
[71] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," Proc. 9th International on Audio/Visual Emotion Challenge and Workshop, pp.55–63, 2019.

**Mariana Rodrigues Makiuchi** received her BS degree from the University of Brasilia in 2018 in Mechatronics Engineering. She is currently an Artificial Intelligence Master's student in the Computer Science Department of Tokyo Institute of Technology. Her research interests include representation learning, deep learning, and emotion recognition.

**Tifani Warnita** received her BS degree in Informatics from Bandung Institute of Technology, Indonesia, in 2017. She was awarded the Gold Award in Research and Development Category of ASEAN ICT Awards (AICTA) from ASEAN for her project Hoax Analyzer. She also won first place in Microsoft Imagine Cup - South East Asia regional competition with the project. She received her M.Eng. degree in Artificial Intelligence from Tokyo Institute of Technology, Japan, in 2019. Her research interests include paralinguistic in speech, affective computing, and deep learning.

**Nakamasa Inoue** is an Assistant Professor in the department of computer science at Tokyo Institute of Technology, Japan. He received his B.E., M.E., and Ph.D. degrees in computer science from Tokyo Institute of Technology in 2009, 2011, and 2014, respectively. His main research interests lie in multimedia processing including video retrieval, image recognition, and speech recognition.

**Koichi Shinoda** received the BS and MS degrees from the University of Tokyo, Tokyo, Japan in 1987 and 1989, respectively, both in physics, and the DEng degree in computer science from the Tokyo Institute of Technology, Japan, in 2001. In 1989, he joined NEC Corporation, Japan, where he was involved in research on automatic speech recognition. From 1997 to 1998, he was a visiting scholar with Bell Labs, Lucent Technologies, Murray Hill, NJ. From October 2001 to March 2003, he was an associate professor with the University of Tokyo, Japan. He is currently a professor with the Tokyo Institute of Technology. His research interests include speech recognition, video information retrieval, statistical pattern recognition, and human interfaces. He received the Awaya Prize from the Acoustic Society of Japan in 1997 and the Excellent Paper Award from the IEICE in 1998. He was a publicity chair in INTERSPEECH2010, a video program co-chair in ACM Multimedia 2012, APSIPA distinguished lecturer from 2013 to 2015, and the chair of SIG-SLP (Spoken Language Processing) in Information Processing Society of Japan from 2014 to 2016. He is currently an associate editor of Computer Speech and Language. He is a fellow of IEICE, a senior member of IEEE and IPSJ.

**Michitaka Yoshimura** is a clinical psychologist with 13 years of experience in the field of mental health. He is currently an Associate Professor of Psychology, doing research and teaching. His specialization is behaviour change of mentally handicapped people using cognitive behaviour therapy and digital devices. He is currently involved in several clinical studies, including assessment and counselling for people with mental disorders.

**Momoko Kitazawa** is a clinical psychologist and researcher working in the fields of psychology, neuropsychiatry and ophthalmology. Her research interests focus on the relationship between facial expressions or blinking and mental disorders. She was appointed as project researcher associate in 2019 and has been involved in several studies on objective severity assessment of mental disorders.

**Kei Funaki** is a Project Assistant Professor in Neuropsychiatry at the Keio University Hospital. He received his PhD in 2019 from the Keio University School of Medicine. He is a member of the Japanese Society of Psychiatry and Neurology and of the Japanese Psychogeriatric Society. His main research interests include dementia, geriatric psychiatry and psychopharmacology.

**Yoko Eguchi** received her Ph.D. from Keio University School of Medicine. Her specialty is clinical neuropsychology, and her primary research interest is the decision-making ability of cognitively impaired patients and elderly patients regarding their health care.

**Taishiro Kishimoto** is the Professor of Future Preventive Medicine and Wellness Laboratory at Keio University School of Medicine, Tokyo, Japan. Dr. Kishimoto graduated from Keio University School of Medicine in 2000. After he was awarded his PhD degree from Keio University in 2009, he joined the Zucker Hillside Hospital in New York where he worked for 3.5 years as a post-doctoral research fellow. He returned to Japan in 2013 to rejoin the Neuropsychiatry Department at Keio University School of Medicine, and began his current position in June 2021.