IEEE Journal of Translational Engineering in Health and Medicine

Received 26 April 2017; revised 13 July 2017 and 25 August 2017; accepted 1 September 2017. Date of publication 15 September 2017; date of current version 29 September 2017.

Digital Object Identifier 10.1109/JTEHM.2017.2752152

Detecting Dementia Through Interactive Computer Avatars

HIROKI TANAKA[1], HIROYOSHI ADACHI[2], NORIMICHI UKITA[3], MANABU IKEDA[4], HIROAKI KAZUI[4], TAKASHI KUDO[2], AND SATOSHI NAKAMURA[1], (Fellow, IEEE)

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Nara 630-0101, Japan
[2]Health and Counseling Center, Osaka University, Osaka 560-0043, Japan
[3]Graduate School of Engineering, Toyota Technological Institute, Nagoya 468-8511, Japan
[4]Department of Psychiatry, Graduate School of Medicine, Osaka University, Osaka 565-0871, Japan

CORRESPONDING AUTHOR: H. TANAKA (hiroki-tan@is.naist.jp)

This work was supported in part by JSPS KAKENHI under Grant JP17H06101 and Grant JP16K16172 and in part by the Yanmar Innovation Lab. 2112.

ABSTRACT This paper proposes a new approach to automatically detect dementia. Even though some works have detected dementia from speech and language attributes, most have applied detection using picture descriptions, narratives, and cognitive tasks. In this paper, we propose a new computer avatar with spoken dialog functionalities that produces spoken queries based on the mini-mental state examination, the Wechsler memory scale-revised, and other related neuropsychological questions. We recorded the interactive data of spoken dialogues from 29 participants (14 dementia and 15 healthy controls) and extracted various audiovisual features. We tried to predict dementia using audiovisual features and two machine learning algorithms (support vector machines and logistic regression). Here, we show that the support vector machines outperformed logistic regression, and by using the extracted features they classified the participants into two groups with 0.93 detection performance, as measured by the areas under the receiver operating characteristic curve. We also newly identified some contributing features, e.g., gap before speaking, the variations of fundamental frequency, voice quality, and the ratio of smiling. We concluded that our system has the potential to detect dementia through spoken dialog systems and that the system can assist health care workers. In addition, these findings could help medical personnel detect signs of dementia.

INDEX TERMS Dementia, spoken dialogue, computer avatars, Alzheimer's disease, MMSE.

I. INTRODUCTION

Dementia is broadly defined as deterioration of memory, reasoning, language, and behavior, which decreases a person's ability to function independently [1]. Dementia is a neurodegenerative disorder that presents itself in different types (e.g., Alzheimer's disease, Primary Progressive Aphasia, Normal pressure hydrocephalus, and Dementia with Lewy Bodies). Its early diagnosis is obviously critical because this allows the patient and family to plan for the future and identify outside sources of assistance. As potentially useful and proven treatments become available, early diagnosis will become increasingly important [2]. However, dementia's early detection is challenging, especially in its most early stages [3].

In order to detect dementia, patients undergo a series of cognitive tests and assessments conducted by a professional team. Specifically, in the case of early-stage detection, complementary tests include the analysis of samples of cerebrospinal fluid from the brain, a functional magnetic resonance brain imaging (fMRI) test, and a blood test [4]. They are relatively expensive and require a significant amount of time and effort. Thus, there is increasing need for additional cost-effective tools that allow the identification of people with dementia in preclinical or early clinical stages [5].

Audiovisual cues extracted from conversation may be indicative of underlying cognitive processes (e.g., word retrieval, semantic difficulties, attention deficits). These features could thus be potentially useful for the detection of dementia [6]. It is hypothesized that dementia is characterized by longer gaps between conversation turns compared to healthy controls, as shown in the literature [7], [8].

Furthermore, recent research has examined a potential alternative clinical use of speech-language for processing speech samples to assess neurodegenerative impairments [7], [9]–[13]. These works used speech data from picture descriptions, narratives, writings, and cognitive tasks. For example, Roark et al. analyzed the recorded data of the Wechsler memory recall task [14] and identified language and acoustic markers for discriminating between healthy elderly subjects and those with mild cognitive impairment [12]. They demonstrated a statistically significant improvement in the area under the receiver operating characteristic (ROC) curve (best performance: 0.86) when using language (e.g., Yngve scoring, Frazier scoring, dependency distance) and acoustic (e.g., total pause time, total phonation time) features, in addition to neuropsychological test scores. Even though linguistic, acoustic, and image features associated with dementia have not been extensively examined under various conversation conditions, we hypothesize that they are potentially effective for detecting dementia.

On the other hand, one previous study applied computer avatars with spoken dialogues to assess neurodegenerative impairments [15]. Another study presented algorithms for inferring cognitive performance associated with keyboard entry and mouse movement by monitoring data from computer games played by the elderly [16]. A different computer avatar system was capable of speech-recognition and language-understanding functionalities, which are potentially useful for evaluating the daily cognitive status of seniors [17]. These types of systems offer the possibility of detecting neurodegenerative impairments automatically.

This paper proposes a novel approach to automatically detect dementia from conversations in human-agent interaction. We developed a computer avatar with spoken dialog functionalities that produces simple spoken queries based on such cognitive tests as the mini-mental state examination (MMSE) [18], the Wechsler memory scale-revised (WMS-R), and an objective structured clinical examination (OSCE) [19].

In this paper, we recorded the interaction data between a computer avatar and elderly participants with dementia as well as age-matched healthy controls. We created a model for detecting dementia using various audiovisual features and machine learning algorithms (support vector machines (SVM) and logistic regression). The following summarizes this paper's results:

1) Developed a spoken dialogue system that communicates with elderly people.
2) Achieved good performance in detecting dementia through conversational data.
3) Found important and promising features for identifying dementia in conversation.

We first summarize the relevant related work on detecting dementia from multiple features (Section 2), explain our proposed interactive system (Section 3), describe our method (Section 4) and give results (Section 5), offer a discussion of our findings (Section 6), and finally give our conclusion and future directions (Section 7).

This paper is an extension of a conference proceeding [20]. We increased the number of participants to provide more detailed analysis.

II. RELATED WORK ON DETECTING DEMENTIA FROM SPEECH, LANGUAGE, AND IMAGE FEATURES

Many works have detected dementia from speech and language features and demonstrated the potential of these approaches [11]–[13], [21], [22].

Orimaye et al. [21] proposed a diagnostic method to identify people with Alzheimer's disease using language features (e.g., syntactic features produced by the Stanford Parser) extracted from transcribed audio files from the DementiaBank dataset.* The dataset contains transcripts of verbal interviews with healthy controls, and people with Alzheimer's disease, including those with mild cognitive impairment. Interviews were conducted in the English language and were based on the description of the Cookie-Theft picture component, which is part of the Boston Diagnostic Aphasia Examination [23]. They used 242 sample files for both healthy controls and people with Alzheimer's disease. They compared five different machine learning algorithms and achieved 74% classification accuracy using an SVM classifier with 10-fold cross-validation. König et al. [22] experimentally used four cognitive vocal tasks (counting backward, sentence repetition, image description, and verbal fluency) with participants who were divided into three groups: healthy controls, people with mild cognitive impairment, and people with Alzheimer's disease. They extracted several vocal features from the audio recordings and classified them into healthy controls and mild cognitive impairment with 79% accuracy and healthy controls and Alzheimer's disease with 87% accuracy. Fraser et al. [11] achieved higher accuracy in a study on the potential of using language and speech features (large number of features (370) to capture a wide range of linguistic phenomena, e.g., part-of-speech, vocabulary richness, and mel-frequency cepstrum coefficients), to identify Alzheimer's disease with speech recordings from the DementiaBank dataset. They selected 240 speech recordings, applied two machine learning algorithms, and obtained the highest accuracy (92%) for distinguishing between healthy controls and Alzheimer's disease. Aramaki et al. [13] reported that people with mild cognitive impairment tend to speak less and use simpler words than healthy controls by analyzing several language features (type-token ratio, potential vocabulary size, dependency distance, and Yngve score) from written and spoken narratives regarding one of the happiest events in their lives.

On the other hand, speech prosody, mutual gaze, laughter and smile are all considered important human communication skills [24]. They convey much information during human-human interaction such as emotional state, social

*https://talkbank.org/DementiaBank/

IEEE Journal of Translational Engineering in Health and Medicine

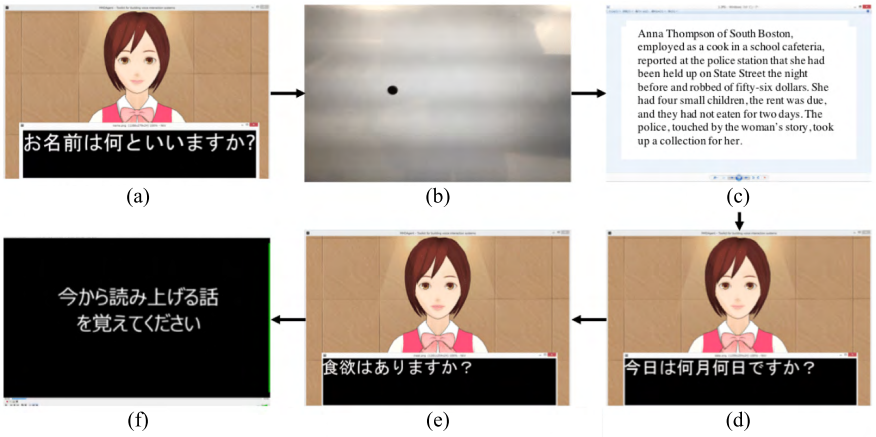

FIGURE 1. Dialogue procedures and subtitles. Japanese sentences and subtitles are translated into English (Sub). The system includes six continuous dialogue procedures. (a) Self-introduction Sub: what's your name? (b) Gaze. (c) Reading. (f) Retelling Sub: Please remember a story. (e) Random QA Sub: How is your appetite?

communication strategy, and personality. This kind of information also appears in human-agent interaction [25]. In particular, a few papers have examined the facial expressions of dementia. For example, exhibiting facial expressions was studied in four people with severe Alzheimer's disease (determined by the facial action coding system (FACS)) under pleasant and unpleasant stimulus conditions [26]. That study's results showed that some types of dementia revealed no variations in facial expressions. For people with Alzheimer's disease, mutual gaze was preserved, reflecting relative maintenance of this aspect of social behavior despite increasing cognitive impairment. In contrast, abnormalities in mutual gaze were apparent for frontotemporal dementia and people with semantic dementia [27]. For the skills of comprehending communicative emotions, deteriorating emotion-recognition ability, rather than deterioration of general cognition, influences the indifferent and awkward interpersonal behaviors of people with Alzheimer's disease [28] and frontotemporal dementia [29].

We can see that social conversation is associated with some cognitive factors and related to dementia types. In this study, we propose an approach that automatically detects dementia through interactive computer avatars.

III. INTERACTIVE SYSTEM

In this section, we describe the proposed system (Fig. 1). It uses, as a computer avatar, an MMDAgent[†] on a Windows application. Our proposed system is a Japanese spoken dialogue system that integrates speech recognition, dialogue management, text-to-speech, and behavior generation. The system was adopted for elderly people by displaying subtitles, producing a lower pitch, and using a slower speaking rate. The development process was carried out in consultation with a professional psychiatrist. For example, the articulation rate, which divides the number of words (measured in MeCab) by voiced seconds, is 3.38 for an avatar's utterance ''kyo wa nan gatsu nan nichi desu ka.'' The pitch was 0.7% lower than the default female character's voice. Subtitles were made using approximately 80-pt Gothic type.

We selected an animated, human-like female character instead of an animal-like character by referring to [20] and [30]. Moreover, it provides a context that is closer to real clinical settings, that is, interacting with a caregiver. When the system recognizes an utterance, after a few seconds the character nods its head. The nodding-behavior motions were created using MikuMikuDance.[‡] By default, the system naturally moves the character's body, which stares to the front without blinking. The avatar is displayed from the front without background distractions. Users can operate and communicate with it using speech throughout the interaction. All of the dialogue system's utterances were created using templates.

We prepared six continuous dialogue procedures, which are summarized as follows:

(a) Self-introduction: The system introduces herself and asks two questions: 1) What's your name? and 2) How old are you?. This first step is important to get the users comfortable with interaction. The purpose of this small talk is primarily to build rapport and trust among the interlocutors [31], [32]. In addition, this is aimed at helping users understand how to use the system.

(b) Gaze: On a computer screen, the system displays a small moving dot at which the user is directed to gaze. This procedure continues for approximately 2 minutes. We previously evaluated this procedure using reaction delay analysis [33].

(c) Reading: The system displays a passage from the Wechsler logical memory task in the WMS-R [12], [14] that users read aloud. This resembles a language-independent task. We used a regular Windows viewer application to display it.

[†]http://www.mmdagent.jp/

[‡]http://www.geocities.jp/higuchuu4/

(d) Fixed QA: The system asks three fixed queries based on MMSE:
   1) What's the date today? (Q1)
   2) Tell me something interesting about yourself. (Q2)
   3) How did you come here today? (Q3)

(e) Random QA: We prepared 13 questions and randomly produced five queries, based on a previous work [19]. The question sets were randomly selected when the system started to avoid habituation to the same questions. The following is the complete list of questions:
   1) What season of the year are we in?
   2) What year is it?
   3) Who is Japan's Prime Minister?
   4) Are you left- or right-handed?
   5) Do you sleep well?
   6) How is your appetite?
   7) Please tell me something that has recently been stressful in your life.
   8) Please tell me about Hibari Misora (a famous singer).
   9) Please tell me about Yujiro Ishihara (a famous actor).
   10) Please tell me about Shigeo Nagashima (a famous baseball player).
   11) What's your hobby?
   12) What's your favorite song?
   13) Please tell me about your family.

(f) Retelling: The system reads aloud from a different part of the Wechsler logical memory in the WMS-R [14], and users paraphrase the passage. Here, we recorded the audio sound spoken by a clinical psychologist in advance. The system plays the sound.

During interactions with the computer avatar, the system recorded user video and audio with a built-in camera and microphone. The system waits for 10 seconds after the user's final utterance for closed/easy-to-answer questions and waits for 15 seconds for manually selected difficult questions before going to the next question. Waiting 10 to 15 seconds to ask a question produced a very long silence compared to what has been reported in the literature for human-human conversations [34]. We observed some users spoke with very small amplitude, which thus could not be recognized by the system. Therefore, we extended the gap pauses.

In order to confirm the load of each task, two Nara Institute of Science and Technology graduate students evaluated the difficulty of each task by sorting. Then we measured agreement of the two students and ordered tasks, based on difficulty, in ascending order from procedure (a) to procedure (f). Completing all of the dialogues required around 10-15 minutes.

IV. METHODS

In this section, we present an experimental evaluation using the system.

A. PARTICIPANTS

We recruited 33 Japanese participants. The Research Ethics Committees of Osaka University Hospital and Nara Institute of Science and Technology reviewed and approved this experiment. Written informed consent was obtained from all participants before the experiment.

In all, 16 participants (12 males and 4 females) were recorded in a clinical room at Osaka University Hospital as the dementia group, and 17 participants (13 males and 4 females) were recorded in a seminar room at the Nara Institute of Science and Technology as healthy controls. We confirmed no dyslexia, dysarthria, hearing loss, visual impairment, or history of other neurological conditions in either group by asking them directly about their health background.

As shown in Table 1 (age, diagnosis, education in years, sex, MMSE score, and time to complete the system), we finally selected 29 participants. Four were removed from the experiment because two in the dementia group had not been diagnosed with dementia yet and two in the healthy controls obtained MMSE scores of 22 and 23, which are below the cut-off score.[§] The dementia group participants were diagnosed as having mild cognitive impairment or dementia (including nine Alzheimer's disease, one Normal pressure hydrocephalus, one Alzheimer's disease + Normal pressure hydrocephalus, and one Dementia with Lewy Bodies) by certified psychiatrists at the Osaka University Hospital. This was done by applying the diagnostic and statistical manual of mental disorders, 4th edition (DSM-IV-TR) [35].

The MMSE scores were significantly different between the two groups. This was confirmed using a two-tailed Welch's t-test ($p < 0.05$) after confirming the variance of the two groups was significant (F test) ($p < 0.05$) and normal using a one-sample Kolmogorov-Smirnov test ($p > 0.05$). The ages were matched ($p > 0.05$). The genders were matched ($p > 0.05$) by a binomial test. Education in years was significantly different between the two groups by using a two-sample t-test ($p < 0.05$) after confirming normality and equal variances. We found fewer years of education in the healthy control. A significance test for education in years and MMSE score was performed based on the individual's abnormality test score against norms ($p < 0.05$, two tailed) [36], as indicated in bold type in Table 1.

B. DATA COLLECTION

Prior to data collection, we tried to test the system by people with dementia and modify it several times to ensure that participants understood/hear the instructions well.

[§] We performed experimental evaluation including these four persons and found a lowered detection performance.

IEEE Journal of Translational Engineering in Health and Medicine

TABLE 1. Participant demographics.

ID	Diagnosis	Age	Education	Sex	MMSE score	Usage time [min:sec]
1c	Control	73	12	male	26	12:16
2c	Control	77	12	male	25	09:45
3c	Control	83	12	male	24	11:11
4c	Control	67	12	female	26	09:21
5c	Control	77	12	female	28	08:35
6c	Control	72	14	female	29	12:42
7c	Control	72	12	male	28	11:03
8c	Control	73	9	male	26	10:40
9c	Control	72	9	male	26	10:13
10c	Control	68	10	male	30	07:04
11c	Control	75	6	female	30	09:03
12c	Control	72	6	male	27	10:45
13c	Control	71	6	male	27	11:12
14c	Control	80	10	male	29	09:46
15c	Control	79	16	male	30	09:43
M	nc	74.1	10.5	nc	27.5	nc
SD	nc	4.4	2.9	nc	2.0	nc
Range	nc	72-83	6-16	nc	24-30	nc
1d	AD	89	16	male	24	10:57
2d	NPH	82	16	male	22	08:08
3d	AD+NPH	80	9	male	**10**	10:52
4d	AD	65	**21**	male	23	10:54
5d	AD	78	12	male	**18**	10:10
6d	AD	68	16	male	30	09:50
7d	DLB	79	**18**	male	23	22:19
8d	AD	71	16	male	**21**	09:30
9d	AD	76	12	female	**16**	09:04
10d	MCI	72	16	female	27	10:57
11d	AD	87	8	male	**18**	09:27
12d	MCI	78	8	male	25	11:20
13d	AD	65	14	female	**20**	09:23
14d	AD	78	16	female	22	12:36
M	nc	76.3	14.1	nc	21.4	nc
SD	nc	7.4	3.9	nc	4.9	nc
Range	nc	65-89	8-21	nc	10-30	nc

AD: Alzheimer's disease, NPH: Normal Pressure Hydrocephalus, DLB: Dementia with Lewy Bodies, MCI: Mild Cognitive Impairment, and nc: not calculated.

Data was collected in the same way across groups. We used a Windows laptop to record the interactions and confirmed that all 29 participants completed every dialogue procedure. The total amount of time needed to complete is shown in Table 1.

Participants were given instructions about data collection by an experimenter. To the extent possible, we maintained consistent microphone amplitude (loud value, 0.8% of maximum volume) and distance between the users and the laptop. Gain of the built-in microphone was consistently set to 70 dB. Participants of both groups were recorded with normal level lighting and quiet conditions. Sampling rate of video was 30 fps and audio was 16,000 Hz.

C. DATA ANNOTATION

After recording, we separated audio files from the recorded video files. Then, one person manually transcribed the recorded data using the Audacity application. All utterances were transcribed based on the USC Rachel corpus manual [37]. In the transcription manual, if the speaker pauses for more than 1 second, the speech is transcribed as separate utterances.

We analyzed the answer videos of the three fixed queries. The time needed for data annotation was around 10 minutes per person.

D. FEATURE EXTRACTION

The following features were studied, and then compared between groups using the Wilcoxon rank sum test and effect size (Cohen's D).

We extracted the speech, language, and image features from the answer videos of the three fixed queries and averaged each feature of the three answers. We selected user-independent and significant audiovisual features based on previous works that found differences between neurodegenerative disorders or dementia and healthy controls [7], [13], [15]. In addition, reducing the number of features avoids over-fitting in the case of supervised machine learning [38] because our data were limited by the number of participants.

1) SPEECH FEATURES

For speech feature extraction, we used the Snack sound toolkit[¶] and considered the fundamental frequency, power, and voice quality. Although mean or SD values of the fundamental frequency were associated with intention and emotional aspects, we did not extract them because they are strongly related to individuality such as sex. Instead we extracted the following statistics of the coefficient of

[¶]http://www.speech.kth.se/snack/

variation (F0cov) from per utterance:

$$F0_{cov} = \frac{sd(f0)}{mean(f0)}. \quad (1)$$

We extracted the mean values of the amplitude (Power) and the articulation rate (AR), which divides the number of words by voiced time. Voice quality (breathiness) was also computed using the difference between the amplitudes of the first harmonic (h1) and the third formant (a3) (h1a3), defined as the spectral tilt [39]:

$$h1a3 = h1 - a3. \quad (2)$$

Furthermore, we extracted silences between the avatar's question and the user's response (Gap). Here, we denote the gap before new turns as the time between the end of the avatar's question (t_q) and the start of the user's answer (t_a) as follows:

$$Gap = t_a - t_q. \quad (3)$$

The maximum waiting time was 10 or 15 seconds in the system, and thus the feature ranges between 0 and 15.

Pauses were manually measured within responses as the number of pauses between sentences (# of pauses) and the proportion of silence length during responses (Prop. of pauses).

2) LANGUAGE FEATURES

We used Mecab[‖] for part-of-speech tagging in the Japanese utterances. We extracted the number of tokens (Tokens) as well as the ratio of hesitations (Hesitations) (e.g., "umm" or "ehh") from the Mecab output. The type-token ratio (TTR), which represents the ratio of the total vocabulary to the overall words, is a simple measure of vocabulary size. We extracted TTR [40] and Simpson's D value, which does not depend on sample size [41], as follows:

$$TTR = \frac{\sum_{all\ m} V(m,N)}{\sum_{all\ m} mV(m,N)}. \quad (4)$$

$$D = \sum_{all\ m} V(m,N)\frac{m}{N}\frac{m-1}{N-1}. \quad (5)$$

Here, N is the number of tokens, and $V(m,N)$ represents the number of words used m times. We also computed the percentage of nouns (Nouns) and verbs (Verbs) according to [42] as well as correction rates (Corrections).

3) IMAGE FEATURES

To analyze the recorded video, we first extracted the number of facial features using 2D facial landmark detection [43] based on Openface.[**] From 68 facial landmarks, following Naim et al. [44], we calculated the distance of the heights of the outer eye-brow, the inner eyebrow, the outer lip, and the inner lip as well as the eye opening and lip corner distance. Using these features, we modeled smiling faces using an expression database [45] that contains 213 images of seven facial expressions by ten Japanese female models. In the database, we used 31 samples of happy faces and 30 samples of neutral faces and trained a model of these two types of facial expressions using SVM with linear kernels. Precision, recall, and the F-measure of the leave-one-out cross-validation over the database were .90, .87, and .88, respectively. For the video, we predicted whether the label belongs to the smiling or neutral class in each frame. The proportion of smiling frames among all of the frames in an utterance was called the smiling ratio (Smile) based on [20].

E. CLASSIFIERS

We used two machine learning algorithms for detecting dementia from the healthy controls. An SVM with linear kernel and logistic regression were used as classifiers, which are supervised machine learning algorithms. SVMs are useful in many pattern recognition tasks such as depression detection [46]. Because they are binary classifiers, they are well suited to the classification task of the dementia group vs. healthy controls. In this experiment, we used selected audiovisual features as input of the classifiers, which were trained to predict the dementia group and healthy control labels with default parameters. Feature sets were automatically selected based on the p-values of the t-test in the training dataset (top six features). The feature values were normalized to fall between 0 and 1 in each training set by identifying the maximum and minimum values for each feature in the training set; the same normalization function was applied to test participants. We evaluated the classification performance with leave-one-participant-out cross-validation and plotted the ROC curve with the areas under it (AUC) [12]. In the ROC curve, we plotted the true and false positive rates. An ROC curve along a diagonal shows a random classification model.

V. RESULTS

A. FEATURES AND CLASSIFIERS

First, we sorted each feature by the p-values of Wilcoxon rank sum test and effect size as shown in Table 2.

Gaps were significantly different between the two groups (p=0.02), indicating people with dementia tended to have more delay responses to the system than the healthy controls. Such features as h1a3, F0cov, and smile were also significantly different ($p < 0.05$).

Table 3 shows the relationship between gap values and question types, which obtained the highest p-value and effect size. While in Table 2 we averaged each question, the question type affected the p-values. Q1 and Q2 were effective for distinguishing dementia; Q3 inadequately identified dementia (Q1: What's the date today?, Q2: Tell me something interesting about yourself, and Q3: How did you come here today?).

[‖] http://taku910.github.io/mecab/

[**] https://github.com/TadasBaltrusaitis/OpenFace

IEEE Journal of Translational Engineering in Health and Medicine

TABLE 2. Feature ranking sorted by p-values. Fourth column shows effect size computed by Cohen's D. Fifth column (Trend) shows trend direction (increasing or decreasing) with dementia group. Bold type indicates statistically significant differences ($p < 0.05$).

Rank	Feature	P-value	Cohen's D	Trend
1	Gap	**0.02**	1.18	↑
2	F0cov	**0.02**	0.98	↑
3	Smile	**0.02**	0.75	↑
4	h1a3	**0.03**	0.99	↑
5	Power	0.11	0.69	↓
6	Verbs	0.22	0.52	↓
7	Corr.	0.25	0.56	↑
8	# of pause	0.31	0.27	↓
9	Prop. of pause	0.34	0.21	↓
10	Hesitations	0.43	0.13	↓
11	Tokens	0.49	0.25	↓
12	Nouns	0.61	0.16	↑
13	AR	0.67	0.15	↓
14	TTR	0.69	0.15	↑
15	D value	0.84	0.01	↓

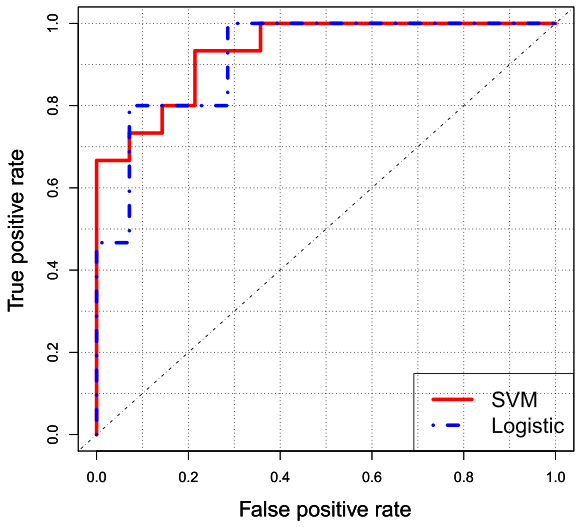

FIGURE 2. ROC curve. For true and false positive rates, red and blue lines show SVM and logistic regression, respectively.

Figure 2 shows the ROC curve. The machine learning algorithms classified the two groups with 0.93 for SVM and 0.91 for logistic regression, as measured by AUC. The unweighted accuracy was 83% (SVM) and 79% (logistic regression).

Here, we trained the SVM with only people with Alzheimer's disease (n=9), and the overall performance was AUC 0.89. It also showed that Gap and F0cov were significantly different between people with Alzheimer's disease and healthy controls ($p < 0.05$). In contrast, smile and h1a3 were not significantly different. As an additional analysis, we also trained the SVM without two MCI participants, and the overall performance slightly decreased (AUC 0.87).

B. ERROR ANALYSIS

Both classifiers failed to correctly detect two persons in the dementia group because those persons (ID: 1d and 13d) answered the questions too quickly. Two participants (ID: 1c and 15c) among the healthy controls was not classified accurately because they delayed their answers and/or their h1a3 scores were relatively high.

TABLE 3. Gap feature of dementia group and healthy controls by question types with mean and SD values. Bold type indicates significant differences ($p < 0.05$).

Question	Control	Dementia	P-value
Q1	0.71 (0.3)	3.58 (4.2)	**0.02**
Q2	3.35 (2.4)	5.87 (3.7)	**0.04**
Q3	1.45 (0.8)	1.36 (0.7)	0.94

VI. DISCUSSION

Our study is the first attempt to detect dementia from audio-visual features through interaction with computer avatars. We extracted generally accepted features that are easy to understand based on several previous works from not only dementia research but also conversation studies.

Our results indicate that higher or at least equivalent detection performance was obtained compared with previous works [7], [10]–[12]. This result establishes a high degree of accuracy [47]. This system may facilitate not only detecting dementia but also making intermediate or follow-up assessments. The SVM performance was superior to that of logistic regression, which is consistent with previous findings [20], [46]. For SVM, we did not control each parameter. To obtain better results, we will attempt a grid search method to find the best parameter sets for a development dataset.

Members of the dementia group delayed in answering the avatar's questions, a finding that is consistent with previous works where dementia significantly delayed responses to stimuli [8]. Gap or hesitations are indicative of cognitive load or speech planning [48]. If the questions require more cognitive load for the dementia than for controls, and therefore longer planning, then a longer gap would be expected. We also showed that question type is related to the detection performance of dementia (see Table 3). Q1 and Q2 were effective for distinguishing dementia; Q3 inadequately identified dementia. Q1 and Q2 may require more time for the dementia to plan what they have to say compared to Q3. Analyzing the differences of fixed and random QA, including open- and closed-type questions [49], would be interesting.

We also found that the voices of the dementia group were not as clear (breathiness, reflected by the high h1a3 values) as the controls. Although elderly voices typically have increased breathiness [50], this characteristic's relation to dementia is still unclear. Further study is needed to investigate the relationship between dementia and voice quality. The pitch variation results were not consistent with previous works that had significantly less pitch modulation based on the Alzheimer's type [51], [52]. Variation of pitch is associated with emotion [53]. Moreover, we found that the participants in the dementia group tended to smile more than the healthy controls. This might be related to smiling out of frustration when having difficulty in answering questions [54]. We found

that healthy controls simply spoke because the system seemed like a cognitive test, which was not so difficult for them.

For generalizing the system to other languages, we should take into consideration that pitch is part of prosody and that it varies across languages. Furthermore, some facial expressions may be culture-dependent [55], for which it may be necessary to train the system with other populations. The language-dependent vs. language-agnostic aspect of audio-visual features in this context needs to be investigated.

In addition, multimodal integration should be considered, such as eye gaze tracking (large delay of eye tracking response in dementia [33]), mutual gaze, actual user responses, and richer linguistic features measured from syntactic parse trees [56].

In this study, several modules were not adapted to elderly people. For example, to achieve more precise prediction of facial expressions, we should train a smile model that uses data specifically recorded from elderly people because, for example, their eye openings are different from the people in our Japanese female facial expression database [57]. Consequently, we need a new dataset of elderly people to construct a more robust smile detection model.

This work is limited in terms of the participants population. Although this was increased from our previous work [20], it remains too small to validate system effectiveness. We need to increase the number of participants to obtain stronger results by considering education in years, interaction type, age, dementia type, and severity [9]. The system may detect impairments in many other neurological populations, but the avatar was not tested in people with other etiologies (e.g., stroke, brain tumors, depression). For example, gaps and delayed responses are not unique to people with dementia but also occur in people after stroke or brain tumors [58], [59]. In addition, we relied on manual transcription of the audio data. It is a clear limitation of using speech-based methods as we have the advantage of being quick and cost effective. We should integrate an automatic speech recognition to the audio data in the future. Furthermore, in this study, the interactive model relies on a fixed threshold of 10-15-second gaps only at turn exchanges; however, this is not truly representative of real-life interactions [60]. A model that adapts the avatar gaps/overlaps to the user style would be more appropriate.

In this study, the two groups differed in education (individuals with dementia have attained higher education in years). While previous works reported there may be aggravating effects of lower education on risk for dementia and language function [61], [62], our study seems to differ from those works by showing results for higher education (years) in the dementia group. Consequently, the differences between groups may not be due to education.

We implemented various ideas to simplify our system and facilitate its use. For instance, since elderly people have difficulty listening to fast speech and high-pitched voices, we reduced the articulation rate of the avatar voice and lowered its mean pitch value based on consultation with a professional psychiatrist. Subtitles helped seniors (especially the dementia group) understand the avatar's utterances by integrating visual and auditory sensors. However, no existing work has investigated the effects of human-computer interaction and human-human interaction in people with dementia. Thus, we need to scrutinize such effects with these interaction types.

A. SINGLE-CASE ANALYSES

As shown in Table 1, we compared each individual with dementia with the healthy controls. Some individuals had significantly lower MMSE scores than the healthy controls. For example, ID: 3d had a significantly low MMSE score of 10, and thus it might have been easier to detect him than the other individuals with dementia. However, while ID: 13d scored an MMSE of 20, we misclassified her by both classifiers due to her rapid answers. Other participants who scored low MMSEs, such as ID: 5d, ID: 8d, ID: 9d, and ID: 11d, were correctly classified, and individuals with higher education, such as ID: 3d and ID: 7d, were also correctly classified.

VII. CONCLUSION

We developed a new computer avatar with spoken dialog functionalities. We recorded 29 participants and extracted features of their behaviors. The analysis results showed that several features effectively distinguished between dementia and healthy controls. We also confirmed that SVM can classify two groups with 0.93 detection performance as measured by AUC with 83% unweighted accuracy. Our system has the potential to detect dementia through spoken dialogs with a computer avatar. In addition, we newly identified some contributing features, e.g., gaps and variations of fundamental frequency. These findings could help medical personnel detect signs of dementia in human-human interaction.

We plan to develop a recommendation system that includes a dementia detection module that is easy to use, is cost-effective, provide repeatable measurements, and does not require a significant amount of time and effort.

ACKNOWLEDGMENT

The authors would like to thank Yuriko Taniguchi, Osaka University, who helped us with the MMSE data collection, and Yuko Kawaguchi who recorded the speech data of our paraphrase component. The authors would also thank the graduate students of the Department of Psychiatry, Osaka University Health Care Center, for their valuable comments that helped us to improve our system.

REFERENCES

[1] G. M. McKhann *et al.*, "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's Dementia*, vol. 7, no. 3, pp. 263–269, 2011.

[2] K. S. Santacruz and D. Swagerty, "Early diagnosis of dementia," *Amer. Family Phys.*, vol. 63, no. 4, pp. 703–713, 2001.

IEEE Journal of Translational Engineering in Health and Medicine

[3] V. Taler and N. A. Phillips, "Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review," *J. Clin. Experim. Neuropsychol.*, vol. 30, no. 5, pp. 501–556, 2008.

[4] C. Laske *et al.*, "Innovative diagnostic tools for early detection of Alzheimer's disease," *Alzheimer's Dementia*, vol. 11, no. 5, pp. 561–578, 2015.

[5] D. S. Geldmacher, "Cost-effective recognition and diagnosis of dementia," in *Seminars Neurology*, vol. 22. New York, NY, USA: Thieme Medical Publishers, 2002, pp. 063–070.

[6] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[7] S. Kato, H. Endo, A. Homma, T. Sakuma, and K. Watanabe, "Early detection of cognitive impairment in the elderly based on Bayesian mining using speech prosody and cerebral blood flow activation," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2013, pp. 5813–5816.

[8] F. J. Pirozzolo, K. J. Christensen, K. M. Ogle, E. C. Hansch, and W. G. Thompson, "Simple and choice reaction time in dementia: Clinical implications," *Neurobiol. Aging*, vol. 2, no. 2, pp. 113–117, 1981.

[9] W. Jarrold *et al.*, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proc. ACL Workshop Comput. Linguistics Clin. Psychol.*, 2014, pp. 27–36.

[10] M. Lehr and E. T. Prud, "hommeaux, I. Shafran, and B. Roark, "Fully automated neuropsychological assessment for detecting mild cognitive impairment," in *Proc. INTERSPEECH*, 2012, pp. 1039–1042.

[11] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *J. Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2015.

[12] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2081–2090, Sep. 2011.

[13] E. Aramaki, S. Shikata, M. Miyabe, and A. Kinoshita, "Vocabulary size in speech may be an early indicator of cognitive impairment," *PloS ONE*, vol. 11, no. 5, p. e0155195, 2016.

[14] D. Wechsler, *WAIS-III: Administration and Scoring Manual: Wechsler Adult Intelligence Scale*. San Antonio, TX, USA: Psychological Corp., 1997.

[15] H. Tanaka *et al.*, "Automated social skills trainer," in *Proc. 20th Int. Conf. Intell. User Interfaces*, 2015, pp. 17–27.

[16] H. Jimison, M. Pavel, J. McKanna, and J. Pavel, "Unobtrusive monitoring of computer interactions to detect cognitive status in elders," *IEEE Trans. Inf. Technol. Biomed.*, vol. 8, no. 3, pp. 248–252, Sep. 2004.

[17] Y. Sakai, Y. Nonaka, K. Yasuda, and Y. I. Nakano, "Listener agent for elderly people with dementia," in *Proc. 7th Annu. ACM/IEEE Int. Conf. Hum.-Robot Interact.*, Mar. 2012, pp. 199–200.

[18] 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," *J. Psychiatric Res.*, vol. 12, no. 3, pp. 189–198, 1975.

[19] R. M. Harden and F. Gleeson, "Assessment of clinical competence using an objective structured clinical examination (OSCE)," *Med. Edu.*, vol. 13, no. 1, pp. 39–54, 1979.

[20] H. Tanaka, H. Adachi, N. Ukita, T. Kudo, and S. Nakamura, "Automatic detection of very early stage of dementia through multimodal interaction with computer avatars," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, 2016, pp. 261–265.

[21] S. O. Orimaye, J. S.-M. Wong, and K. J. Golden, "Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances," in *Proc. 1st Workshop Comput. Linguistics Clin. Psychol. (CLPsych)*, 2014, pp. 78–87.

[22] A. König *et al.*, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's Dementia, Diagnosis, Assessment Disease Monitor.*, vol. 1, no. 1, pp. 112–124, 2015.

[23] E. Kaplan, H. Goodglass, and S. Weintraub, *Boston Naming Test*. Austin, TX, USA: Pro-ed, 2001.

[24] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, 2009.

[25] F. Yang, M. A. Sehili, C. Barras, and L. Devillers, "Smile and laughter detection for elderly people-robot interaction," in *Proc. Int. Conf. Social Robot.*, 2015, pp. 694–703.

[26] K. Asplund, A. Norberg, R. Adolfsson, and H. M. Waxman, "Facial expressions in severely demented patients—A stimulus–response study of four patients with dementia of the Alzheimer type," *Int. J. Geriatric Psychiatry*, vol. 6, no. 8, pp. 599–606, 1991.

[27] V. E. Sturm *et al.*, "Mutual gaze in Alzheimer's disease, frontotemporal and semantic dementia couples," *Social Cognit. Affective Neurosci.*, vol. 6, no. 3, pp. 359–367, 2010.

[28] A. Shimokawa *et al.*, "Influence of deteriorating ability of emotional comprehension on interpersonal behavior in alzheimer-type dementia," *Brain cognition*, vol. 47, no. 3, pp. 423–433, 2001.

[29] D. Fernandez-Duque and S. E. Black, "Impaired recognition of negative facial emotions in patients with frontotemporal dementia," *Neuropsychologia*, vol. 43, no. 11, pp. 1673–1687, 2005.

[30] A. L. Baylor and Y. Kim, "Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role," in *Intelligent Tutoring Systems* (Lecture Notes in Computer Science) vol. 3220. Berlin, Germany: Springer, 2004, pp. 592–603.

[31] T. Bickmore and J. Cassell, "Small talk and conversational storytelling in embodied conversational interface agents," in *Proc. AAAI Fall Symp. Narrative Intell.*, 1999, pp. 87–92.

[32] H. Tanaka, K. Yoshino, K. Sugiyama, S. Nakamura, and M. Kondo, "Multimodal interaction data between clinical psychologists and students for attentive listening modeling," in *Proc. Conf. Oriental Chapter Int. Committee Coordination Standardization Speech Databases Assessment Techn. (O-COCOSDA)*, 2016, pp. 95–98.

[33] T. Endo *et al.*, "Initial response time measurement in eye movement for dementia screening test," in *Proc. 15th IAPR Int. Conf. Mach. Vis. Appl. (MVA)*, May 2017, pp. 262–265.

[34] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *J. Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[35] American Psychiatric Association, "Diagnostic and statistical manual, 4th edn, text revision (DSM-IVTR)," *Amer. Psychiatric Assoc., Washington*, 2000.

[36] J. R. Crawford and P. H. Garthwaite, "Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences," *Neuropsychologia*, vol. 40, no. 8, pp. 1196–1208, 2002.

[37] E. Mower, M. P. Black, E. Flores, M. Williams, and S. Narayanan, "Rachel: Design of an emotionally targeted interactive agent for children with autism," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2011, pp. 1–6.

[38] C. M. Bishop, "Pattern recognition," in *Machine Learning*, vol. 128. New York, NY, USA: Springer-Verlag, 2006, pp. 1–58.

[39] M. Hanson, "Glottal characteristics of female speakers," Ph.D. dissertation, Dept. Division Appl. Sci., Harvard Univ., Cambridge, MA, USA, 1995.

[40] R. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, 2000.

[41] H. H. Wright, S. Silverman, and M. Newhoff, "Measures of lexical diversity in aphasia," *Aphasiology*, vol. 17, no. 5, pp. 443–452, 2003.

[42] R. Herbert, J. Hickin, D. Howard, F. Osborne, and W. Best, "Do picture-naming tests provide a valid assessment of lexical retrieval in conversation in aphasia?" *Aphasiology*, vol. 22, no. 2, pp. 184–203, 2008.

[43] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1034–1041.

[44] I. Naim, M. I. Tanveer, D. Gildea, and M. (Ehsan) Hoque. (Apr. 2015). "Automated analysis and prediction of job interview performance." [Online]. Available: https://arxiv.org/abs/1504.03425

[45] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 1998, pp. 200–205.

[46] J. F. Cohn *et al.*, "Detecting depression from facial actions and vocal prosody," in *Proc. 3rd Int. Conf. Affective Comput. Intell. Interact. Workshops (ACII)*, Sep. 2009, pp. 1–7.

[47] C. E. Metz, "Basic principles of ROC analysis," *Seminars Nucl. Med.*, vol. 8, no. 4, pp. 283–298, Oct. 1978.

[48] R. Lunsford, P. A. Heeman, L. Black, and J. V. Santen, "Autism and the use of fillers: Differences between 'um' and 'uh,'" in *Proc. DiSS-LPSS Joint Workshop*, 2010, pp. 107–110.

[49] T. Chaspari, D. Bone, J. Gibson, C.-C. Lee, and S. Narayanan, "Using physiology and language cues for modeling verbal response latencies of children with ASD," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2013, pp. 3702–3706.

[50] L. Zhou, K. C. Fraser, and F. Rudzicz, "Speech recognition in Alzheimer's disease and in its assessment," in *Proc. INTERSPEECH*, 2016, pp. 1948–1952.
[51] J. J. Meilán, F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana, "Speech in Alzheimer's disease: Can temporal and acoustic parameters discriminate dementia?" *Dementia Geriatric Cognit. Disorders*, vol. 37, nos. 5–6, pp. 327–334, 2014.
[52] K. Horley, A. Reid, and D. Burnham, "Emotional prosody perception and production in dementia of the Alzheimer's type," *J. Speech, Lang., Hearing Res.*, vol. 53, no. 5, pp. 1132–1146, 2010.
[53] A. Paeschke and W. F. Sendlmeier, "Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements," in *Proc. ISCA Tuts. Res. Workshop (ITRW) Speech Emotion*, 2000, pp. 75–80.
[54] M. Hoque and R. W. Picard, "Acted vs. Natural frustration and delight: Many people smile in natural frustration," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops (FG)*, Mar. 2011, pp. 354–359.
[55] G. Trovato *et al.*, "A novel culture-dependent gesture selection system for a humanoid robot performing greeting interaction," in *Proc. Int. Conf. Social Robot.*, 2014, pp. 340–349.
[56] B. Roark, M. Mitchell, and K. Hollingshead, "Syntactic complexity measures for detecting mild cognitive impairment," in *Proc. Workshop Biolo., Transl., Clin. Lang. Process. (BioNLP)*, 2007, pp. 1–8.
[57] U. Hess, R. B. Adams, A. Simard, M. T. Stevenson, and R. E. Kleck, "Smiling and sad wrinkles: Age-related changes in the face and the perception of emotions and intentions," *J. Experim. Social Psychol.*, vol. 48, no. 6, pp. 1377–1380, 2012.
[58] E. M. Saffran, R. S. Berndt, and M. F. Schwartz, "The quantitative analysis of agrammatic production: Procedure and data," *Brain Lang.*, vol. 37, no. 3, pp. 440–479, 1989.
[59] J. Vermeulen, R. Bastiaanse, and B. Van Wageningen, "Spontaneous speech in aphasia: A correlational study," *Brain Lang.*, vol. 36, no. 2, pp. 252–274, 1989.
[60] C. Chao and A. L. Thomaz, "Turn taking for human–robot interaction," in *Proc. AAAI Fall Symp. Dialog Robots*, 2010, pp. 132–134.
[61] E. S. Sharp and M. Gatz, "The relationship between education and dementia an updated systematic review," *Alzheimer Disease Assoc. Disorders*, vol. 25, no. 4, p. 289, 2011.
[62] D. Leibovici, K. Ritchie, B. Ledésert, and J. Touchon, "Does education level determine the course of cognitive decline?," *Age Ageing*, vol. 25, no. 5, pp. 392–397, 1996.

HIROKI TANAKA received the master's and Ph.D. degrees from the Nara Institute of Science and Technology, Japan, in 2012 and 2015, respectively. He is an Assistant Professor with the Graduate School of Information Science, Nara Institute of Science and Technology. His research interest is assisting people with disabilities through human–computer interaction.

HIROYOSHI ADACHI received the Ph.D. degree from Osaka University, Osaka, Japan, in 2004. He is an Associate Professor with the Health and Counseling Center, Osaka University. His main research interest is the early detection of dementia through sleep disturbance.

NORIMICHI UKITA received the Ph.D. degree in informatics from Kyoto University, Japan, in 2001. He was a Research Scientist with PRESTO, Japan Science and Technology Agency, from 2002 to 2006, and a Visiting Research Scientist with the Robotics Institute, Carnegie Mellon University, from 2007 to 2009. He was an Assistant Professor with the Nara Institute of Science and Technology, then an Associate Professor in 2007. He has been a Professor with the Toyota Technological Institute, Japan, since 2016. His main research interests are multi-object tracking and human pose estimation. He has received the Best Paper Award from the IEICE in 1999.

MANABU IKEDA received the degree from the Department of Anthropology, Faculty of Science, The University of Tokyo, in 1984, and the degree from the Osaka University Medical School in 1988. He has been engaged with neuropsychology, neuropathology, and old age psychiatry research in Osaka University, Psychiatric Research Institute of Tokyo, Hyogo Institute for Aging Brain and Cognitive Disorders, University of Cambridge, Ehime University School of Medicine, and Kumamoto University. He was the Chairman and a Professor with the Department of Neuropsychiatry, Kumamoto University, in 2007, and has been the Chairman and a Professor with the Department of Psychiatry, Graduate School of Medicine, Osaka University, since 2016. With his background of neuropsychiatry, neuropsychology, and old age psychiatry, his research covers a wide area in clinical psychiatry.

HIROAKI KAZUI received the degree from the Tottori University School of Medicine in 1989, and the degree from the Graduate School of Medicine, Osaka University, in 1995. He trained in neuropsychiatry with Osaka University Hospital and emergency medicine with Hyogo Medical University Hospital. He was with the Hyogo Institute for Aging Brain and Cognitive Disorders, Himeji, Japan, he became a Research Associate in 2002 and an Associate Professor in 2006 with the Department of Psychiatry, Graduate School of Medicine, Osaka University. His research subjects were the elucidation of neural bases of cognitive impairment and neuropsychiatric symptoms in patients with mental disorders, including dementia. He has been a Principal Investigator of national research projects for the development of treatment of neuropsychiatric symptoms of dementia for over ten years in Japan. He is also interested in idiopathic normal pressure hydrocephalus because of its treatable nature and is involved in researching the syndrome from the point of view of neuropsychiatry.

IEEE Journal of Translational Engineering in Health and Medicine

TAKASHI KUDO was born in Osaka, Japan, in 1958. He received the M.D. degree from Osaka Medical College in 1986, and the Ph.D. degree from the Graduate School of Medicine, Osaka University, in 1990. From 1990 to 1992, he holds a post-doctoral position with the New York State Institute Basic Research in Developmental Disabilities. He was an Assistant Professor in 1994–1999 and an Associate Professor in 1999–2013, respectively, with the Department Psychiatry, Graduate School of Medicine, Osaka University. Since 2013, he has been a Professor with the Department of Mental Health Promotion, Graduate School of Medicine, Osaka University.

SATOSHI NAKAMURA (F'16) received the B.S. degree from the Kyoto Institute of Technology in 1981, and the Ph.D. degree from Kyoto University in 1992. He was an Associate Professor with the Graduate School of Information Science, Nara Institute of Science and Technology, from 1994 to 2000. He was the Director of ATR Spoken Language Communication Research Laboratories from 2000 to 2008 and the Vice President of ATR in 2007–2008. He was the Director General of Keihanna Research Laboratories and the Executive Director of the Knowledge Creating Communication Research Center with the National Institute of Information and Communications Technology, Japan, in 2009–2010. He is a Professor with the Graduate School of Information Science, Nara Institute of Science and Technology, Japan, the Honorar Professor of with the Karlsruhe Institute of Technology, Germany, and ATR Fellow. He is currently the Director of the Data Driven Science Center and Augmented Human Communication Laboratory, and a Full Professor with the Graduate School of Information Science, Nara Institute of Science and Technology. He is interested in the modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received the LREC Antonio Zampoli Award 2012 and many domestic awards. He has been Elected Board Member of the International Speech Communication Association, ISCA, since 2011, the *IEEE Signal Processing Magazine* Editorial Board Member during 2012–2014, and the IEEE SPS Speech and Language Technical Committee Member since 2013.