Automatic Detection of Very Early Stage of Dementia through Multimodal Interaction with Computer Avatars

Hiroki Tanaka
Department of Information Science, NAIST
Ikoma-shi, Nara, Japan
hiroki-tan@is.naist.jp

Hiroyoshi Adachi
Department of Psychiatry, Osaka University Health Care Center
Toyonaka, Osaka, Japan
hadachi@psy.med.osaka-u.ac.jp

Norimichi Ukita
Department of Information Science, NAIST
Ikoma-shi, Nara, Japan
ukita@is.naist.jp

Takashi Kudo
Department of Psychiatry, Osaka University Health Care Center
Toyonaka, Osaka, Japan
kudo@psy.med.osaka-u.ac.jp

Satoshi Nakamura
Department of Information Science, NAIST
Ikoma-shi, Nara, Japan
s-nakamura@is.naist.jp

ABSTRACT

This paper proposes a new approach to detecting very early stage of dementia automatically. We develop a computer avatar with spoken dialog functionalities that produces natural spoken queries referring to Mini Mental State Examination, Wechsler Memory Scale-Revised and other related questions. Multimodal interactive data of spoken dialogues from 18 participants (9 dementias and 9 healthy controls) are recorded, and audiovisual features are extracted. We confirm that the support vector machines can classify into two groups with 0.94 detection performance as measured by areas under ROC curve. It is found that our system has possibilities to detect very early stage of dementia through spoken dialog with our computer avatars.

CCS Concepts

•**Applied computing** → *Health care information systems;*

Keywords

Dementia, spoken dialogue, avatars

1. INTRODUCTION

Dementia is broadly defined as deterioration in memory, thinking, and behavior that decreases a person's ability to function independently [14]. Early diagnosis of dementia is important for several reasons. The most important reason is that early diagnosis allows the patient and family to plan for the future and identify outside sources of assistance. Moreover, as potentially useful and proven treatments become available, early diagnosis of dementia will become increasingly important [19]. The early detection of dementia is challenging, especially in its very early stages [21]. Currently, there is no powerful tool that gives a reliable detection of dementia: rather, the patient has to go through a series of cognitive tests conducted by a professional neurologist for assessments. This process involves a certain amount of anxiety and stress. Especially in the case of the very early stage detection, complementary tests include the analysis of samples of cerebrospinal fluid taken from the brain, a magnetic resonance brain imaging test, and a blood test [11]. Such methods are invasive, bring discomfort to the participants, are relatively costly and require a significant amount of effort and time. Thus, there is an increasing need for additional noninvasive and/or cost-effective tools, allowing identification of participants in the preclinical or early clinical stages of dementia.

As noninvasive and cost-effective approaches, previous works have attempted to detect dementia from their speech and language attributes [8, 12, 5, 18, 1]. For example, Aramaki et al., [1] reported that mild cognitive impairments tend to speak less and use easier words than healthy control. However, there is no existing powerful tool to detect dementia from audiovisual features, and most of them used non-interactive data such as picture description, narrative, and cognitive tasks. In contrast, there was a study that applies interactive computer avatars with spoken dialogue to detect user's social behaviors [23].

This paper proposes a new approach to detect very early stage of dementia automatically. We develop a computer avatar with spoken dialog functionalities that produces simple natural spoken queries referring to cognitive tests such as Mini Mental State Examination (MMSE) and Wechsler Memory Scale-Revised (WMS-R) as well as other related medical questions [7].

2. RELATED WORK

Several works have attempted to detect or diagnose dementia from speech and language features, and have demonstrated the potential of the approaches to identifying dementia.

Orimaye et al [16] proposed a diagnostic method to identify people with Alzheimer's disease using a large number of language features extracted from transcribed audio files from the DementiaBank dataset[1]. They used 242 sample files for both healthy controls and people with Alzheimer's disease. They compared five different machine learning algorithms, achieving a 74% classification accuracy using a support vector machine (SVM) classifier with 10% cross-validation. Konig et al [10] performed an experiment of using four cognitive vocal tasks (a counting backward task, a sentence repeating task, an image description task, and a verbal fluency task) with participants divided into three groups: healthy controls, people with mild cognitive impairment, and people with Alzheimer's disease. They extracted features from the audio recordings. They achieved to classify into healthy controls and mild cognitive impairment with an accuracy of 79%, healthy controls and Alzheimer's disease with an accuracy of 87%. Recently and achieving higher accuracy, Fraser et al [5] studied the potential of using language features to identify Alzheimer's disease. They used speech recordings along with their manually transcribed files derived from the DementiaBank dataset. They selected 240 speech recordings with a set of 370 speech and language features. Then, they applied two machine learning algorithms and obtained a highest accuracy of 92% in distinguishing between healthy controls and Alzheimer's disease.

There have been a small number of papers examining facial expression of dementia as we surveyed so far. For example, the ability to exhibit facial expressions was studied in four patients with severe dementia of the Alzheimer's disease, by means of the Facial Action Coding System (FACS) under pleasant and unpleasant stimulus conditions. The results showed that some types of dementia tend to mute facial expression [2].

This paper proposes a new approach to detect very early stage of dementia automatically. We develop an interactive computer avatar with spoken dialog functionalities that produces natural spoken queries.

3. INTERACTIVE SYSTEM

We used MMDAgent[2] as a computer avatar. The proposed system is a Japanese spoken dialogue integrating speech recognition, dialogue management, text-to-speech, and behavior generation. MMDAgent works as a Windows application. The system was adapted to elderly people by displaying subtitles and slower speaking rate. This development process was conducted in discussion with a professional psychiatrist. We selected an animated female character who is similar to an actual human, as opposed to an animal-like character, but not more realistic human-like one, as we hope that this will make the conversation interesting. When the system recognizes an utterance, after a few seconds the system nods its head. The nodding behavior motions were created by MikuMikuDance[3].

We prepared a total of 6 continuous dialogue procedures to detect very early stage of dementia, and summarize them as follows (in Figure 1):

(a) Self-introduction: The system introduces herself and asks user's name and age. This process is to make users comfort to interact.

(b) Gaze: The system displays a small dot on a computer screen, and users are directed to gaze the moving dot.

(c) Reading: The system displays a document of Wechsler Logical Memory I (immediate) task in the WMS-R [24, 18], and users read aloud the sentence.

(d) Fixed Q&A: The system asks a total of three fixed queries as referring to the MMSE [4]. It consists of 1) What is the date today?, 2) What is your memorable story?, 3) How did you come here today?

(e) Random Q&A: The five queries are randomly produced. A total of 13 questions were prepared for random questions such as "Do you have any appetite?," "Please tell me about Shigeo Nagashima" (who is a famous baseball player in Japan), as referring to e.g. [7].

(f) Retelling: The system read aloud a document of a different part of the Wechsler Logical Memory, and users retell the sentence.

During interaction with a computer avatar, the system records user's video and audio using built-in camera and microphone. The system waits 10 seconds after a user's final utterance for closed / easy to answer questions, and waits 15 seconds for difficult questions to go on to a next question. Two graduate students evaluated load and difficulty of each task. Then, we ordered tasks according to its load and difficulty (ascending order). A total amount of time to complete all dialogues is around 10-15 minutes.

4. EXPERIMENTAL EVALUATION

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

4.1 User Study

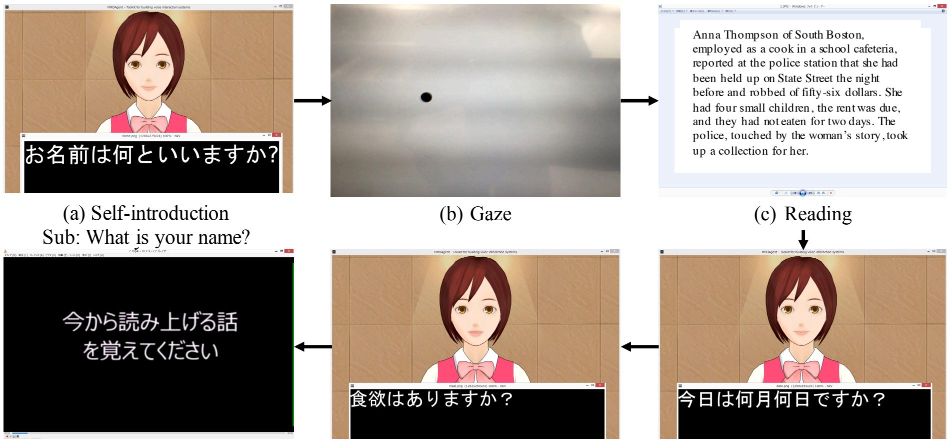
We recruited a total of 20 participants. Each participant gave informed consent before the data recording. 10 participants (9 males and 1 female) were recorded at Osaka University Hospital as dementia group, and other 10 participants (7 males and 3 females) were recorded at Nara Institute of Science and Technology as healthy controls. Participants of the dementia group were diagnosed as very early stage of dementia by expert clinicians at the Osaka University Hospital according to Diagnostic and Statistical Manual of Mental Disorders, 4th. Edition (DSM-IV). We collected information on age and the MMSE score from all participants. We used a laptop (Surface Pro 3) to record interaction, and confirmed 20 participants could complete all procedures of the dialogue as mentioned in the previous section. An amplitude of microphone and distance between users and the laptop is consistent as long as possible in the two separate locations. One person manually transcribed the recorded data.

As shown in Table 1, we finally used audiovisual data of 18 participants. Two participants were removed from

[1] https://talkbank.org/DementiaBank/

[2] http://www.mmdagent.jp/

[3] http://www.geocities.jp/higuchuu4/


お名前は何といいますか?

Anna Thompson of South Boston, employed as a cook in a school cafeteria, reported at the police station that she had been held up on State Street the night before and robbed of fifty-six dollars. She had four small children, the rent was due, and they had not eaten for two days. The police, touched by the woman's story, took up a collection for her.

今から読み上げる話を覚えてください

食欲はありますか？

今日は何月何日ですか？

(a) Self-introduction
Sub: What is your name?

(b) Gaze

(c) Reading

(f) Retelling
Sub: Please remember a story.

(e) Random questions
Sub: Do you have any appetite?

(d) Fixed questions
Sub: What is the date today?

Figure 1: Dialogue procedures. Japanese sentences and subtitles are translated into English (Sub).

the experiment because one in the dementia group did not receive a diagnosis of dementia yet and one in the healthy controls obtained 22 of the MMSE score (below the cut-off score). The MMSE score was significantly different between two groups by using two-tailed Student's t-test ($p < .05$), and age was matched ($p > .05$).

Table 1: Participant demographics with mean and SD values of age and the MMSE score.

Group	N	Age	MMSE
Dementia	9	76.4 (7.4)	21.4 (1.6)
Controls	9	73.9 (4.5)	26.4 (5.7)

4.2 Feature set

We extracted acoustic, linguistic, interactive, and visual features from videos of the answers' speech of the three fixed queries. We averaged each feature of the three answers. We selected user-independent and insightful audiovisual features based on previous works which found differences between psychiatric disorders and early stage of dementia, and healthy controls [23, 9, 1]. Features were compared between groups by using Student's t-test (two-tailed).

4.2.1 Acoustic features

For speech feature extraction, we used the Snack sound toolkit[4]. Here, we considered fundamental frequency, power, speech rate, and voice quality. We did not extract mean values of fundamental frequency because those features are strongly related to individuality. Thus, we extracted statistics of coefficient of variation (F0cov) for fundamental frequency. We extracted mean value of power (Power), and speech rate (SR), which is a feature dividing the number of words by the number of voiced seconds. Voice quality is also computed using the difference between the first harmonic (h1) and the third formant (a3) (h1a3) which defined by [6] as follows.

$$h1a3 = h1 - a3. \tag{1}$$

4.2.2 Language features

We used Mecab[5] for part-of-speech tagging in Japanese utterances. We extracted the number of tokens (Tokens) as well as the ratio of fillers (Fillers) (e.g. "umm," or "eh" in Japanese) from an output of the Mecab. Type-token ratio (TTR) represents the ratio of the total vocabulary to the overall words, and is a simple measure of vocabulary size. We extracted TTR as referring to [3] as follows.

$$TTR = \frac{number\ of\ types}{number\ of\ tokens}. \tag{2}$$

We also computed word difficulty as the ratio of nouns with above intermediate level to the overall nouns (Difficulty). We used word difficulty dictionary[6] to obtain nouns with above intermediate level.

4.2.3 Interactive features

We extracted pauses before new turns (Pause). Here, we denote values of pauses before new turns as a time between the end of the avatar's question (t_q) and the start of the user's answer (t_a). A maximum waiting time is 10 or 15 seconds in the system, and thus the feature ranges between 0 and 15 as follows.

$$Pause = t_a - t_q. \tag{3}$$

4.2.4 Visual features

We extracted the ratio of smiling as a facial expression feature. To analyze the recorded video, fist we extracted a number of facial features by using a constrained local model [20] based face tracker[7]. From a total of 66 feature points,

[4] http://www.speech.kth.se/snack/

[5] http://taku910.github.io/mecab/

[6] http://jreadability.net/

[7] https://github.com/kylemcdonald/FaceTracker

we calculated the features of outer eye-brow height, inner eyebrow height, outer lip height, inner lip height, eye opening, and lip corner distance following Naim et al. [15]. Using these features, we modeled smiling faces using the Japanese female facial expression database [13]. For video, we predicted whether the label belongs to the smiling or neutral class in each frame, and the proportion of smiling frames among all frames was named the ratio of smiling (Smile). To verify that the model generalizes to other speakers and video, we used the NOCOA+ database [22], which contains derisive and friendly videos of four Japanese men. We extracted the ratio of smiling from these videos and confirmed that friendly videos have a significantly larger ratio of smiling compared to derisive videos (p=0.002). Because the system did not record images below the chest, we did not take into account non-facial gestures.

4.3 Classifier

We used two machine learning algorithms for detecting very early stage of dementia from healthy controls. SVM with a sigmoid kernel and logistic regression were compared as classifiers. In this experiment, we used all audiovisual features as an input of the classifiers, and the classifiers trained to predict the label belonging to the dementia group and the healthy controls. We evaluated a classification performance with leave-one-subject-out cross-validation, and plotted ROC curve with areas under ROC curve (AUC) [18].

4.4 Results

First, we sorted each feature by their p-values of t-test as well as Cohen's d values as shown in Table 2.

Table 2: Feature ranking sorted by p-values. The fifth column (Trend) shows the direction of the trend (increasing or decreasing) of the dementia group. Bold fonts indicate significant differences ($p < 0.05$).

Rank	Feature	p-value	Cohen's d	Trend
1	Pause	**0.003**	1.69	↑
2	h1a3	**0.017**	1.31	↑
3	TTR	**0.028**	1.15	↑
4	Tokens	**0.031**	1.17	↓
5	Smile	0.067	0.99	↑
6	F0cov	0.076	0.92	↑
7	Power	0.079	0.88	↓
8	Difficulty	0.195	0.64	↓
9	Fillers	0.383	0.42	↓
10	SR	0.867	0.08	↓

The pause was significantly different between two groups (p=0.003), indicating patients with very early stage of dementia tend to delay responses to the system than the healthy controls. The other features such as the h1a3, the TTR, the Tokens were also significantly different (p=0.028). Here, the TTR was not consistent with the previous work, which showed that most of the healthy controls had lexically richer speech than dementia participants [3]. This was caused because our data was limited in terms of the number of tokens, especially in the dementia group. Rest of features were not significantly different (p > .05), but have possibilities to contribute to further classification. These results showed that our small number of features were effective to distinguish between the dementia group and the healthy controls.

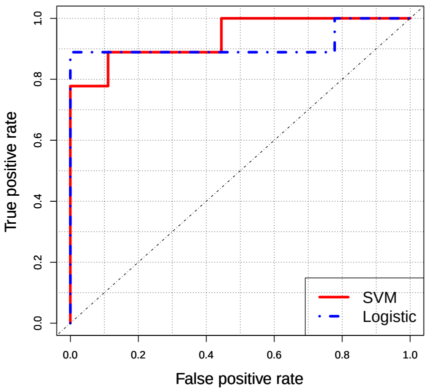


Figure 2: ROC curve. Red and blue lines show true positive rate and false positive rate of SVM and Logistic regression respectively.

Figure 2 shows ROC curve, and we confirmed that machine learning algorithms classified two groups with 0.94 for SVM and 0.91 for logistic regression as measured by AUC. This result indicates a high detection performance compared to previous works [12, 5, 18]. The classifiers could not correctly detect one person in the dementia group. This is because the person spoke the largest number of tokens and was relatively early to answer the questions in the dementia group.

5. CONCLUSION

We developed a computer avatar with spoken dialog functionalities that produces natural spoken queries. As the initial analysis, 18 participants were recorded and audiovisual features were extracted. The results of the analyses showed that several features were effective to distinguish between very early stage of dementia and healthy controls. It was also confirmed that SVM can classify two groups with 0.94 detection performance as measured by AUC. We found that the system has possibilities to detect very early stage of dementia through spoken dialog with a computer avatar.

This work is a first attempt to detect dementia through multimodal interaction with computer avatars. However, the several modules are not adapted to elderly people. For example, to achieve more precise prediction of facial expressions, we have to train facial points model using data of elderly people. Also, this work is limited in terms of 1) the number of participants, 2) feature set, and 3) question set. First, we plan to collect a large number of patients at the Hospital as well as age-matched controls. Second, multimodal integration should be considered such as eye gaze tracking, actual user responses, and richer linguistic features (e.g. syntax complexity measures [17]). Third, we should take other question sets or actual user responses into consideration incorporating with automatic speech recognition (ASR). For other future directions, we plan to examine different types of dementia [8].

6. ACKNOWLEDGMENTS

This work was supported by Yanmar Lab 2112.

7. REFERENCES

[1] E. Aramaki, S. Shikata, M. Miyabe, and A. Kinoshita. Vocabulary size in speech may be an early indicator of cognitive impairment. *PloS one*, 11(5):e0155195, 2016.

[2] 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. *International Journal of Geriatric Psychiatry*, 6(8):599–606, 1991.

[3] 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*, 14(1):71–91, 2000.

[4] 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*, 12(3):189–198, 1975.

[5] K. C. Fraser, J. A. Meltzer, and F. Rudzicz. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422, 2015.

[6] M. Hanson. *Glottal characteristics of female speakers. Harvard University.* PhD thesis, Ph. D. dissertation, 1995.

[7] R. M. Harden and F. Gleeson. Assessment of clinical competence using an objective structured clinical examination (osce). *Medical education*, 13(1):39–54, 1979.

[8] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, pages 27–36, 2014.

[9] 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 *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5813–5816. IEEE, 2013.

[10] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert, et al. Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):112–124, 2015.

[11] C. Laske, H. R. Sohrabi, S. M. Frost, K. López-de Ipiña, P. Garrard, M. Buscema, J. Dauwels, S. R. Soekadar, S. Mueller, C. Linnemann, et al. Innovative diagnostic tools for early detection of alzheimer's disease. *Alzheimer's & Dementia*, 11(5):561–578, 2015.

[12] M. Lehr, E. T. Prud'hommeaux, I. Shafran, and B. Roark. Fully automated neuropsychological assessment for detecting mild cognitive impairment. In *INTERSPEECH*, pages 1039–1042, 2012.

[13] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205. IEEE, 1998.

[14] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, 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*, 7(3):263–269, 2011.

[15] I. Naim, M. I. Tanveer, D. Gildea, et al. Automated analysis and prediction of job interview performance. *arXiv preprint arXiv:1504.03425*, 2015.

[16] 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 *Proceedings of the 1st Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pages 78–87, 2014.

[17] B. Roark, M. Mitchell, and K. Hollingshead. Syntactic complexity measures for detecting mild cognitive impairment. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 1–8. Association for Computational Linguistics, 2007.

[18] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye. Spoken language derived measures for detecting mild cognitive impairment. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2081–2090, 2011.

[19] K. S. Santacruz and D. Swagerty. Early diagnosis of dementia. *Am Fam Physician*, 63(4):703–713, 2001.

[20] J. M. Saragih, S. Lucey, and J. F. Cohn. Face alignment through subspace constrained mean-shifts. In *IEEE 12th International Conference on Computer Vision*, pages 1034–1041, 2009.

[21] V. Taler and N. A. Phillips. Language performance in alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of Clinical and Experimental Neuropsychology*, 30(5):501–556, 2008.

[22] H. Tanaka, S. Sakti, G. Neubig, T. Toda, and S. Nakamura. Nocoa+: Multimodal computer-based training for social and communication skills. *IEICE Transaction on Information and Systems*, E98-D(8):1536–1544, 2015.

[23] H. Tanaka, S. Sakti, G. Neubig, T. Toda, H. Negoro, H. Iwasaka, and S. Nakamura. Automated social skills trainer. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 17–27. ACM, 2015.

[24] D. Wechsler. *WAIS-III: Administration and scoring manual: Wechsler adult intelligence scale.* Psychological Corporation, 1997.