COMPUTATIONAL COGNITIVE ASSESSMENT: INVESTIGATING THE USE OF AN INTELLIGENT VIRTUAL AGENT FOR THE DETECTION OF EARLY SIGNS OF DEMENTIA

Bahman Mirheidari[1] *Daniel Blackburn*[2] *Ronan O'Malley*[2] *Traci Walker*[3]
Annalena Venneri[4] *Markus Reuber*[5] *Heidi Christensen*[1,6]

[1]Department of Computer Science, [2]Sheffield Institute for Translational Neuroscience (SITraN)
[3]Department of Human Communication Sciences, [4]Department of Neuroscience, Royal Hallamshire Hospital
[5]Academic Neurology Unit, Royal Hallamshire Hospital
[6]Centre for Assistive Technology and Connected Healthcare (CATCH), University of Sheffield, Sheffield, UK
{bmirheidari2, heidi.christensen}@sheffield.ac.uk

ABSTRACT

The ageing population has caused a marked increased in the number of people with cognitive decline linked with dementia. Thus, current diagnostic services are overstretched, and there is an urgent need for automating parts of the assessment process. In previous work, we demonstrated how a stratification tool built around an Intelligent Virtual Agent (IVA) eliciting a conversation by asking memory-probing questions, was able to accurately distinguish between people with a neuro-degenerative disorder (ND) and a functional memory disorder (FMD). In this paper, we extend the number of diagnostic classes to include healthy elderly controls (HCs) as well as people with mild cognitive impairment (MCI). We also investigate whether the IVA may be used for administering more standard cognitive tests, like the verbal fluency tests. A four-way classifier trained on an extended feature set achieved 48% accuracy, which improved to 62% by using just the 22 most significant features (ROC-AUC: 82%).

***Index Terms*—** clinical applications of speech technology, speaker diarisation, automatic speech recognition

1. INTRODUCTION

Dementia is a disorder affecting cognitive skills including memory, and caused by different pathological processes like Alzheimer's disease (AD). Dementia can affect a person's speech, language and interaction capabilities. The number of people developing dementia is increasing drastically, thus the early diagnosis of dementia is of great clinical importance, and there is a need for automatic, easy-to-use, low-cost and a accurate stratifiaction tools.

Recent studies using the qualitative methodology of conversation analysis (CA) demonstrated that communication problems may be picked up during conversations between patients and neurologists and that this can be used to differentiate between patients with neuro-degenerative disorder (ND) and functional memory disorder (FMD; exhibiting problems with memory not caused by dementia) [1, 2]. However, conducting manual CA is expensive and difficult to scale up for routine clinical use. We have therefore developed a fully automatic system based on analysing a person’s speech and language as they speak to an Intelligent Virtual Agent (IVA). The IVA asks a series of memory-probing questions that have been found to be cognitively demanding to answers. These questions are mimicking the style of questions often using during the *history taking* part of a normal face-to-face consultation. A number of features routed in conversation analysis were extracted and high accuracy levels were achieved when evaluating the system in a real memory clinic on patients with ND and FMD [3, 4].

The use of IVAs has recently become more prevalent in healthcare applications. An IVA is a talking head animation displayed on a screen which might be accompanied by other speech/video technologies such as text-to-speech (TTS), pre-recorded audio/video and automatic speech recognition (ASR) embedded in a form of spoken dialogue system that conducts conversations with users or provide different services for them (e.g. motivating them to go for a walk). Applications include use by people with mental health problems [5, 6, 7, 8], mild cognitive impairment (MCI) [9], Alzheimer's Disease (AD) [10, 11], and the healthy elderly [12]. Nakatani *et al.* developed a 3D virtual agent from a photo of a familiar face, such as a family member, to communicate with people with dementia and provide “person centred care” [13]. IVAs have been used for *detecting* dementia as well. Tanaka *et al.* designed an IVA with spoken dialogue for detecting the early signs of dementia [14]. Although that system was based on standard cognitive tests (MMSE and Wechsler logical memory), in line with our findings, it demonstrated encouraging results for the use and acceptability of an IVA-based, automatic interactional system for patients with memory concerns.

In general, an interface based around *conversation* is often preferred over other modes of interaction with computers (keyboards or touch screens) as it is seen as more natural and easy to use. It is sometimes even preferred over interaction with human; for example, the disclosure of potentially embarrassing information to a computer may be easier than to a human being, especially if the talking head is perceived to be supported by Artificial Intelligent (AI) [15].

The initial objective of introducing the IVA was to assess the feasibility of eliciting conversations with people with memory problems. That is, the IVA acted as a neurologist (a *virtual doctor* and asked similar questions to those asked in a real assessment situation. This paper further explores the applicability of an IVA in the diagnostic process by augmenting the initial conversation-based assessment to include more standard test procedures, namely by administering *verbal fluency tests*. In addition we have expanded our diagnostic categories to also include healthy elderly controls and MCI, that is, better reflecting the variety of conditions seen in practice.

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

2. VERBAL FLUENCY TESTS

A verbal fluency test is one of the standard cognitive tests used for assessing people at risk of developing dementia, and comes in two main varieties: *semantic* (naming from a category e.g. animal or fruit) and *phonemic* (naming words beginning with a letter e.g. “P”). Impaired verbal fluency is common amongst people with dementia. For example, people with AD show more deficiency in the 1-minute fluency semantic test comparing to the 1-minute fluency phonemic test [16]. Forbes *et al.* reported that compared to HCs, people with AD i) produced fewer words, ii) tended to use words acquired earlier in life, i.e. words with a lower Age of Acquisition (AoA)[1] [17], iii) use words with a higher occurrence frequency and as well as more typical examples (e.g. thinking of ”lion” faster than ”kangaroo”).

Pakhomov *et al.* claimed that the performance of the fluency semantic test is dependent on the efficiency of clustering the related items in a category by the examiner [18]. They used a Latent Semantic Association (LSA) approach to automatically determine the category of the words and calculate the mean of semantic clusters for all words, as well as the mean of semantic clusters in the neighbouring words. However, they could not find a significant correlation between their automatic features and the manual scores. Later, they extended the features to the density of repeated words and semantic and lexical diversity. On a longitudinal study of people with dementia and HC, they found that the later features showed a much more significant decline in MCI and AD patients, while they almost stayed the same for HC [19].

Verbal fluency tests are routinely administered as part of diagnosing and automating this as well as the scoring would free up valuable time for the clinicians.

3. IVA FOR DETECTING DEMENTIA

The automatic dementia detection system that we have developed consists of a speaker diarisation module (to identify speaker and segment the recording), an automatic speech recogniser (ASR) (to transcribe the segmented audio files), and a feature extraction unit (to extract a number of features from the audio files as well as the outputs produced by the diarisation and ASR modules). The features, are used for training a classifier to determine the diagnostic category of the input recording (for more details please refer to [20, 3]). The Kaldi diarisation[2] and ASR [21] toolkit recipes were used for training the diarisation and ASR modules of the system respectively. The IVA software used for this study was based on the Botlibre[3] library. Based on feedback from end-users, we chose to replace the synthesized speech with recordings of human speech.

The IVA asks a question when the participant clicks on a button. Since the participants were mostly old people, who were less familiar with computers and the use of a PC mouse, as a further simplification, they were directed to use just two keys: ‘enter’ (play) and‘space bar’ (repeat). A laptop was used to run the IVA application. The audio was recorded using a distant microphone, Tascam^{TM} DR-40, which was placed on the table, as well as two other attached microphones close to the subjects. The laptop's built-in camera was used for video capturing the participant's face (and/or that of the accompanying person) from the front, while another camera (Hue) was located on a table near to participants, and recorded the session from a side angle. This allowed us to capture extra movements of the patient which may not be captured by the front camera. In the current study, we have not used the videos.

[1]The age we normally learn a word for the first time.

[2]https://github.com/kaldi-asr/kaldi/tree/master/egs/callhome_diarization

[3]https://www.botlibre.com

Table 1. Samples of the questions. Conv.= Conversational question, Flu.= Verbal fluency test.

Type	Question
Conv.	Where have you come in from today and what are you hoping to find out? Tell me what problems you have noticed with your memory?
Flu.	Please name as many animals as you can in one minute. Please name as many words as you can that begin with the letter “P”, except for names of people such as Peter or names of countries such as Portugal.

Table 2. Demographic information of the participants. FMD: Functional Memory Disorder, ND: Neuro-degenerative Disorder, MCI: Mild Cognitive Impairment, HC: Healthy Control.

Class	Number	Age	Female
FMD	10	56.4 (+/- 6.22)	60.0%
ND	19	69.8 (+/- 8.24)	36.8%
MCI	18	62.2 (+/- 9.52)	33.3%
HC	14	69.4 (+/- 8.22)	57.1%
All	61	65.3 (+/- 9.62)	44.3%

The participants were asked a total of 10 conversational questions and encouraged to take part in two 1-minute verbal fluency tests (*”name as many animals...”*, and *”name as many words starting with the letter P...”* . Table 1 shows samples of the questions and fluency test prompting. Note that these questions are very similar to those being asked by the neurologists in our initial data set (the Hallamshire data set) [22, 20]. The two verbal fluency tests are the standard screening tests known as the ‘fluency semantic’ and ‘fluency phonemic’ tests.

4. EXPERIMENTAL SETUP

The IVA has enabled us to evaluate our dementia detection system in a real clinical setting, and we have collected data during the summers of 2016, 2017 and 2018. The data (hereafter called the IVA data set) was collected by three MSc students at the Department of Neurology, University of Sheffield based at the Royal Hallamshire Hospital.

Of the total number of 78 participants, 61 were chosen for the study (the rest were diagnosed as having non-common memory problems). Table 2 shows the demographic information of the participants in the study. The partients were grouped into four classes: FMD, ND, MCI, and HC. Table 3 shows the information of the three data sets used: HALLAM (45 neurologist-patient interactions), SEIZURE (261 neurologist-patient interactions, like HALLAM but for seizure patients), IVA (61 patient-IVA recordings). The HALLAM and SEIZURE data sets were used for training the i-vector based diarisation module (similar to the recipe for the CALL_HOME data set using the Probabilistic Linear Discriminant Analysis (PLDA; [23]) and the Bidirectional Long Short Term Memory/Time-Delay Neural Network (BLSTM)-TDNN based ASR.

For the 18 recordings of the IVA data set, where the manual transcripts were available, the Diarisation Error Rate (**DER**) was around **11%**, and the Word Error Rate (**WER**) was **59%**. Note that here we did not use the leave-one-out cross validation approach for train-

Table 3. Data set info including Length: the total length in hours/mins, Utts.: number of utterances, Spks.: number of speakers, and Avg. Utts.: Average utterance length in seconds.

Data set (No)	Length	Utts.	Spks.	Avg. Utts.
HALLAM (45)	12h	8,970	117	4.8s
SEIZURE (241)	50h 16m	28,000	597	6.3s
IVA (61)	4h 20m	1,944	85	8.0s

ing the diarisation and the ASR modules (i.e. we could add $n-1$ recordings from IVA data set to HALLAM and SEIZURE to improve the performance of the diarisation and the IVR). This allowed us to evaluate our system on a totally unseen data set (**'blind evaluation'**).

We extracted 72 features from the conversations from both the patients ('Pat') and the accompanying persons ('Aps'), if present. These features are categorised in four different feature types: CA-inspired, lexical-only, acoustic-only and word vector features. Table 4 lists the names of the features with a brief description (for more details, please refer to [3, 4, 20, 22]). In addition to the features extracted from the conversations, 6 more features were extracted automatically from the verbal fluency tests (3 for the animal naming and 3 for the P-words test). We counted the number of correctly named items, as well as the average and standard deviation of the AoA for the naming (see Table 5).

5. RESULTS

The extracted features were passed to a 'Logistic Regression' classifier to determine the category of each audio recording. The k-fold cross validation approach with $k = 10$ was used for training the classifier. In addition to the 4-way classifier accuracy (ND/FMD/MCI/HC), 6 more binary classifiers were trained as well, one for each possible pair of diagnostic categories.

5.1. Classification accuracy

Table 6 shows the accuracy for the 4-way classifier as well as for the 6 binary classifiers. Using the conversational only features (Conv.), the accuracy of the 4-way classifier was 43%. Similarly, the accuracy when using just fluency test features (Flu.) was 39%; both above the 25% chance level. However, combining the features (Conv.+Flu.) boosted the accuracy of the classifier to 48% with even higher accuracy gains seen for the binary classifiers. The best accuracy for the binary classifiers was achieved by the ND/HC classifier which reached 85%. The FMD/ND classifier gained 79% accuracy. [4]

Using the Recursive Feature Elimination (RFE) [24] feature selection, the 22 most significant features were selected. This further increased the accuracy of the 4-way classifier to 62% using the most significant features. ND/HC achieved 94% accuracy, while the accuracy of FMD/ND remained the same.In the set of the most significant features are some CA-inspired, some word vector and some fluency test features (with 10 and 6 and 4 features respectively).

[4]We previously reported an accuracy of 91% however that was based on using substantially more training data, as we used leave-one-out cross validation on 12 recordings of IVA and 30 more for training from the manual transcripts of HALLAM [3]. Here, we have reported results using k-fold cross validation ($k = 10$) to split the data into training and test partitions.

Table 4. List of the extracted features from the conversations. Prefixes:'Pat' (patient), 'Aps' (accompanying person(s).

Type	Feature
CA-inspired	number of turns (**APsNoOfTurns**, **PatNoOfTurns**); average length of turn (**APsAVTurnLength**, **PatAVTurnLength**); number of unique words in a turn (**APsAVUniqueWords**, **PatAVUniqueWords**); patient answers "me" for question "who's most concerned" (**PatMeForWhoConcerns**); patient recalls memory failure features (**PatFailureExampleEmptyWords**, **PatFailureExampleAVPauses**, **PatFailureExampleAllTime**); patient replies 'dunno for the expectation question (**PatDunnoForExpectations**); average number of filler, empty, unique and low-frequency words (**PatAVFillers**), **PatAVEmptyWords**, **PatAVUniqueWords PatAVAllWords**); average number of repeated questions (**AVNoOfRepeatedQuestions**); average number of topics discussed (**AVNoOfTopics**)
Lexical-only	average number of verbs, nouns, adjectives, adverbs, pronouns, wh_words(e.g, who), determiner, conjunctions, cardinals, existential(e.g., there is), prepositions etc(**PatAvgVerb**, **PatAvgNoun**, **PatAvgAdjective**, **PatAvgAdverb**, **PatAvgPronoun**, **PatAvgWh_word**, **PatAvgDeterminer**, **PatAvgConjunction**, **PatAvgCardinal**, **PatAvgExistential**, **PatAvgPreposition**, **PatAvgOtherPOS, APsAvgVerb, APsAvgNoun, APsAvgAdjective, APsAvgAdverb, APsAvgPronoun, APsAvgWh_word, APsAvgDeterminer, APsAvgConjunction, APsAvgCardinal, APsAvgExistential, APsAvgPreposition, APsAvgOtherPOS**)
Acoustic-only	average overall intonation, pitch, duration and silence(**PatAvgIntonation**, **PatAvgPitch**, **PatAvgDuration PatAvgSil**,**APsAvgIntonation**, **APsAvgPitch**, **APsAvgDuration APsAvgSil**); difference between the first harmonic and the harmonic close to the first, second and third formants(**PatAvgH1-A1**, **PatAvgH1-A2**, **PatAvgH1-A3**,**APsAvgH1-A1**, **APsAvgH1-A2**, **APsAvgH1-A3**); difference between the two first harmonics (**PatAvgH1-H2, APsAvgH1-H2**); local jitter and shimmer(**PatAvgGitterLocal**, **PatAvgShimmerLocal**, **APsAvgGitterLocal**, **APsAvgShimmerLocal**); harmonics-to-noise and noise-to-harmonics ratios(**PatAvgMeanHNR**, **PatAvgMeanNHR**,**APsAvgMeanHNR**, **APsAvgMeanNHR**)
Word V.	Word vector features dimensions were reduced to 7 using the (Principal component analysis) PCA (**WV_col1**, ..., **WV_col7**)

Table 5. Verbal fluency test' features. Sem: fluency semantic test (animal naming), Phn: fluency phonemic test (p-words).

Feature	Description
PatSemCount/ PatPhnCount	Number of unique animals/P-words correctly uttered.
PatSemAVGAoA/ PatPhnAVGAoA	Average AoA for the fluency semantic/phonemic test.
PatSemSTDAoA/ PatPhnSTDAoA	Standard deviation AoA for the fluency semantic/phonemic test.

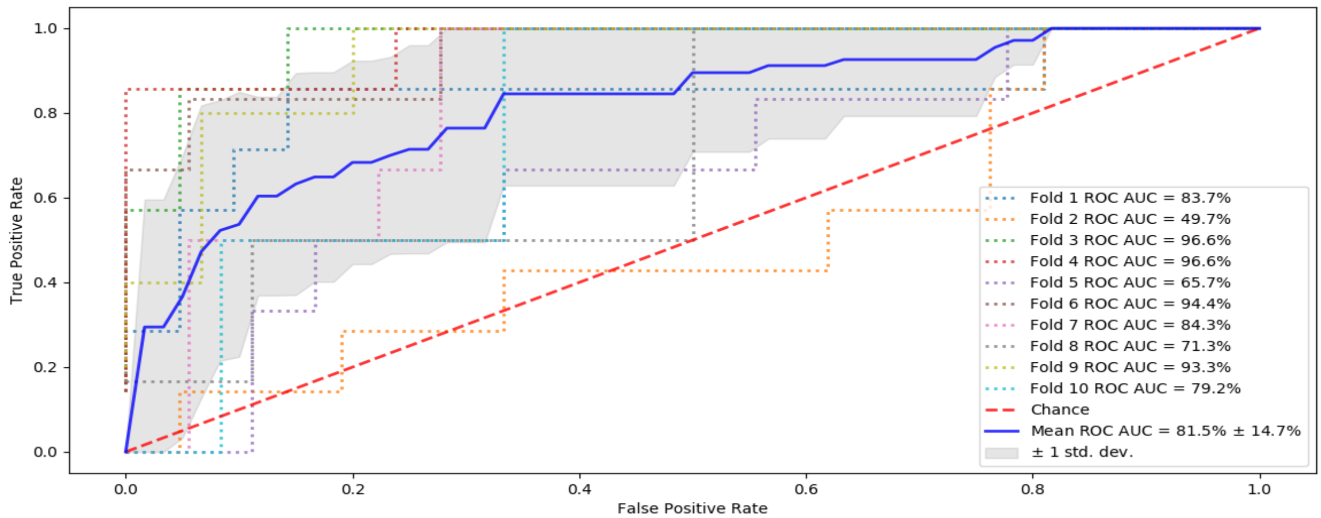

Fig. 1. *ROC-AUC for both the conversations and the verbal fluency tests (the most significant features).*

Table 6. Classification accuracy. Conv.: Conversational question, Flu.: Verbal fluency test, *:the most significant features.

Features (No)	**FMD/ND/MCI/HC**	**FMD/ND**	**FMD/MCI**	**ND/MCI**	**FMD/HC**	**ND/HC**	**MCI/HC**
Conv.(72)	43%	**79%**	68%	51%	54%	**88%**	**78%**
Flu.(6)	39%	69%	**71%**	51%	50%	73%	69%
Conv.+Flu.(78)	**48%**	**79%**	**71%**	**57%**	**58%**	85%	72%
Conv.+Flu.(22)*	***62%***	***79%***	***75%***	***68%***	***63%***	***94%***	***88%***

5.2. Receiver Operating Characteristic curve

In addition to the accuracy, it is important for a classifier to have a high sensitivity and specificity[5]. The Receiver Operating Characteristic (ROC) curve shows the true positives against the false positives for different settings (thresholds) of a classifier. Figure 1 shows the ROC Area Under Cover (AUC) for the 4-way classifier using the most significant features. The blue line is the average ROC-AUC for 10 folds and the grey area shows the error range. The average ROC-AUC for this classifier is around 82% with 15% errors. This indicates a relatively robust classifier regarding both the sensitivity and the specificity of the classification.

6. CONCLUSIONS

Building on previous successful evaluations of our IVA-based dementia detection system, we explored the feasibility of using it for administering standard dementia screening tests such as the verbal fluency tests. Extracting features based on Age of Acquisition and number of correctly named items, we showed that the IVA is effective for eliciting patient responses, and that adding the fluency test improves accuracy for the 4-way classification (HC/MCI/FMD/ND) achieving 62%, when feature selecting is also applied. The ND/HC was the easiest binary classification (94%), while ND/MCI (the two groups with the most clinical overlap) was harder (68%).

[5]Often these are referred to as recall and precision respectively

Table 7. The most significant (22) features.

Rank	**Feature**	**Feature type**
1	ApsAvgSil	Acoustic-only
2	PatAVPauses	CA-inspired
3	PatAvgSil	Acoustic-only
4	PatSemSTDAoA	Fluency semantic test
5	ApsAVUniqueWords	CA-inspired
6	APsAVTurnLength	CA-inspired
7	PatAVFillers	CA-inspired
8	PatSemCount	Fluency semantic test
9	WV_col5	Word vector
10	PatPhnAVGAoA	Fluency phonemic test
11	PatSemAVGAoA	Fluency semantic test
12	WV_col4	Word vector
13	WV_col7	Word vector
14	APsNoOfTurns	CA-inspired
15	WV_col1	Word vector
16	WV_col3	Word vector
17	PatFailureExampleEmptyWords	CA-inspired
18	PatAVUniqueWords	CA-inspired
19	WV_col2	Word vector
20	PatAVTurnLength	CA-inspired
21	PatAVAllWords	CA-inspired
22	PatNoOfTurns	CA-inspired

7. ACKNOWLEDGEMENTS

This research has been partly supported under the European Union's H2020 Marie Skodowska-Curie programme TAPAS (Training Network for Pathological Speech processing; Grant Agreement No. 766287)

8. REFERENCES

[1] C. Elsey, P. Drew, D. Jones, D. Blackburn, S. Wakefield, K. Harkness, A. Venneri, and M. Reuber, "Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics," *Patient Education and Counseling*, vol. 98, pp. 1071–1077, 2015.

[2] D. Jones, P. Drew, C. Elsey, D. Blackburn, S. Wakefield, K. Harkness, and M. Reuber, "Conversational assessment in memory clinic encounters: interactional profiling for differentiating dementia from functional memory disorders," *Aging & Mental Health*, vol. 7863, pp. 1–10, 2015.

[3] B Mirheidari, D Blackburn, K Harkness, T Walker, A Venneri, M Reuber, and H Christensen, "An avatar-based system for identifying individuals likely to develop dementia," *Proc. Interspeech*, pp. 3147–3151, 2017.

[4] B. Mirheidari, D. Blackburn, A. Venneri, M. Reuber, T. Walker, and H. Christensen, "Detecting signs of dementia using word vector representations," in *Proc. Interspeech*. ISCA, 2018.

[5] M. Rus-Calafell, J. Gutiérrez-Maldonado, and J. Ribas-Sabaté, "A virtual reality-integrated program for improving social skills in patients with schizophrenia: a pilot study," *Journal of behavior therapy and experimental psychiatry*, vol. 45, no. 1, pp. 81–89, 2014.

[6] J. Leff, G. Williams, M. Huckvale, M. Arbuthnot, and A. Leff, "Avatar therapy for persecutory auditory hallucinations: what is it and how does it work?," *Psychosis*, vol. 6, no. 2, pp. 166–176, 2014.

[7] M. Huckvale, J. Leff, and G. Williams, "Avatar therapy: an audio-visual dialogue system for treating auditory hallucinations.," in *INTERSPEECH*, 2013, pp. 392–396.

[8] M. Hayward, A. Jones, L. Bogen-Johnston, N. Thomas, and C. Strauss, "Relating therapy for distressing auditory hallucinations: A pilot randomized controlled trial," *Schizophrenia research*, vol. 183, pp. 137–142, 2017.

[9] M. Morandell, A. Hochgatterer, S. Fagel, and S. Wassertheurer, "Avatars in assistive homes for the elderly," in *Symposium of the Austrian HCI and Usability Engineering Group*. Springer, 2008, pp. 391–402.

[10] E. Carrasco, G. Epelde, A. Moreno, A. Ortiz, I. Garcia, C. Buiza, E. Urdaneta, A. Etxaniz, M. González, and A. Arruti, "Natural interaction between avatars and persons with alzheimers disease," in *International Conference on Computers for Handicapped Persons*. Springer, 2008, pp. 38–45.

[11] M. Tran, P. Robert, and F. Bremond, "A virtual agent for enhancing performance and engagement of older people with dementia in serious games," in *Workshop Artificial Compagnon-Affect-Interaction 2016*, 2016.

[12] E. Cyarto, F. Batchelor, S. Baker, and B. Dow, "Active ageing with avatars: a virtual exercise class for older adults," in *Proceedings of the 28th Australian Conference on Computer-Human Interaction*. ACM, 2016, pp. 302–309.

[13] S. Nakatani, S. Saiki, and M. Nakamura, "Integrating 3d facial model with person-centered care support system for people with dementia," in *International Conference on Intelligent Human Systems Integration*. Springer, 2018, pp. 216–222.

[14] H. Tanaka, H. Adachi, N. Ukita, M. Ikeda, H. Kazui, T. Kudo, and S. Nakamura, "Detecting dementia through interactive computer avatars," *IEEE journal of translational engineering in health and medicine*, vol. 5, pp. 1–11, 2017.

[15] A. Rizzo, G. Lucas, J. Gratch, G. Stratou, L. Morency, K. Chavez, R. Shilling, and S. Scherer, "Automatic behavior analysis during a clinical interview with a virtual human.," in *MMVR*, 2016, pp. 316–322.

[16] SJ. Duff Canning, L. Leach, D. Stuss, L. Ngo, and SE. Black, "Diagnostic utility of abbreviated fluency measures in alzheimer disease and vascular dementia," *Neurology*, vol. 62, no. 4, pp. 556–562, 2004.

[17] K. Forbes-McKay, A. Ellis, M. Shanks, and A. Venneri, "The age of acquisition of words produced in a semantic fluency task can reliably differentiate normal from pathological age related cognitive decline," *Neuropsychologia*, vol. 43, no. 11, pp. 1625–1632, 2005.

[18] S. Pakhomov and L. Hemmy, "A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the nun study," *Cortex*, vol. 55, pp. 97–106, 2014.

[19] S. Pakhomov, L. Eberly, and D. Knopman, "Characterizing cognitive performance in a large longitudinal study of aging with computerized semantic indices of verbal fluency," *Neuropsychologia*, vol. 89, pp. 42–56, 2016.

[20] B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Toward the automation of diagnostic conversation analysis in patients with memory complaints," *Journal of Alzheimer's Disease*, , no. Preprint, pp. 1–15, 2017.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[22] B. Mirheidari, D. Blackburn, M. Reuber, T. Walker, and H. Christensen, "Diagnosing people with dementia using automatic conversation analysis," in *Proc. Interspeech*. ISCA, 2016, pp. 1220–1224.

[23] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision. IEEE 11th International Conference*, 2007, pp. 1–8.

[24] F. Pedregosa and G. Varoquaux, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.