Multimodal speech-based dialogue for the Mini-Mental State Examination

Alexander Prange
Mira Niemann
DFKI
Saarbrücken, Germany
alexander.prange@dfki.de
mira.niemann@dfki.de

Antje Latendorf
Anika Steinert
Charité - Universitätsmedizin Berlin
Berlin, Germany
antje.latendorf@charite.de
anika.steinert@charite.de

Daniel Sonntag
DFKI
Saarbrücken, Germany
daniel.sonntag@dfki.de

ABSTRACT

We present a system-initiative multimodal speech-based dialogue system for the Mini-Mental State Examination (MMSE). The MMSE is a questionnaire-based cognitive test, which is traditionally administered by a trained expert using pen and paper and afterwards scored manually to measure cognitive impairment. By using a digital pen and speech dialogue, we implement a multimodal system for the automatic execution and evaluation of the MMSE. User input is evaluated and scored in real-time. We present a user experience study with 15 participants and compare the usability of the proposed system with the traditional approach. Our experiment suggests that both modes perform

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3299040

equally well in terms of usability, but the proposed system has higher novelty ratings. We compare assessment scorings produced by our system with manual scorings made by domain experts.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI)*; • **Applied computing** → *Health informatics*;

KEYWORDS

HCI; Multimodal Speech-Based Dialogue; Cognitive Assessments; Digital Pen; Mini-Mental State Examination; MMSE;

ACM Reference Format:
Alexander Prange, Mira Niemann, Antje Latendorf, Anika Steinert, and Daniel Sonntag. 2019. Multimodal speech-based dialogue for the Mini-Mental State Examination. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'19 Extended Abstracts), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3290607.3299040

BACKGROUND & RELATED WORK

The internationally used Mini-Mental State Examination (MMSE) [6] is a 30-point questionnaire, which is extensively used in medicine and research to measure cognitive impairment. It is administered by a trained professional, who leads the subject through the questionnaire, while taking notes. Afterwards he manually evaluates the results, based on his notes and a predefined scoring scheme. Two tasks of the questionnaire require the subject to perform handwriting or sketching. Administration takes on average 5 to 10 minutes. Due to its standardization, validity, short administration period, and ease of use, the MMSE is widely applied as a screening tool for dementia [7].

In this case study we present an automated system for the administration and evaluation of the MMSE using a multimodal speech-based dialogue system. We discuss lessons learned while implementing this system and present results of a user experience study comparing our system with the traditional version. We evaluate cognitive impairment automatically by utilizing inexpensive, off-the-shelf consumer hardware to capture speech and handwriting input. For the speech dialogue we employ the Google Dialogflow[1] framework in conjunction with our dialogue manager, whereas tasks involving handwriting are captured with a digital pen which streams the user's input to our server. Based on the traditional scoring method we adopted an automated version that scores user input in real-time. We analyze both, the spoken as well as the sketched parts of the MMSE.

We selected the MMSE based on feedback from domain experts and a recent market analysis of existing, most widely used, cognitive assessments conducted by Niemann et al. [8]. Another cognitive assessment tool, the Clock Drawing Test (CDT) has been digitized only recently in a first version with

Mini-Mental State Examination questions and tasks include, for example:

- "What year is it?"
- "Which country are we in?"
- "On which floor are we?"
- "I will show you items, please name them." (wristwatch and pencil)
- "Repeat the following 3 words: Lemon, Key, Ball."
- "Write a sentence on this piece of paper."
- "Use the pen to copy these figures below on the piece of paper." (overlapping pentagons, as depicted in figure 2)
- ...

Sidebar 1: Examples of tasks and questions from the Mini-Mental State Examination (MMSE).

[1]https://dialogflow.com/

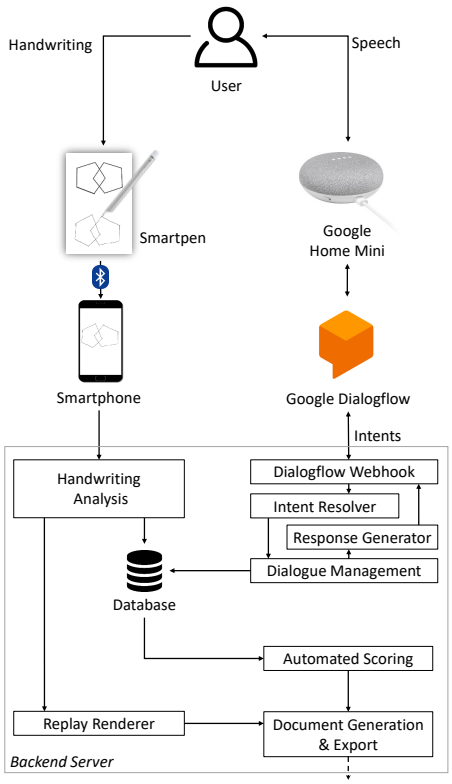


Figure 1: Architecture of the multimodal system.

[2]https://home.google.com/

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

[4]https://www.myscript.com/

a digital pen [4]. Both these tests can assess dementia and screen for it in the community, general practice and general hospital settings [10].

Taking into account multiple sensor inputs is the next step to improve neurocognitive testing by taking advantage of multimodal, multisensor interfaces [9]. We concentrate on evaluating sensor input from speech and handwriting to automatically assess cognitive impairment [13]. Automatically scoring cognitive assessments has several benefits, e.g., automatic assessments are potentially more objective than human assessments, they can include analysis of new features (such as digital pen input) and allow clinicians to shift their attention to other aspects. In this work we focus on the question whether we can conduct tests in an automatic way and thereby reduce the caregiver's time spent on administering and evaluating the assessment. We conduct a user experience study with 15 participants, comparing the traditional version to our novel, automated system and compare the assessment scores produced by the system to manual scoring results made by domain experts. Our experiment suggests that both modes perform equally well in terms of usability, but the automated system receives higher marks in novelty. The comparison of assessment scorings shows that the automated system produces very similar scores to the ones provided by human experts. Our system is currently being deployed for further evaluation in the geriatric daycare clinic of a large hospital.

SYSTEM ARCHITECTURE

The setup is based on our multimodal framework described in Niemann et al. [8] and consists of 3 main components: a speech dialogue system, a smartpen, and a backend server. In order to allow for an easy deployment and large-scale usage, we decided to use the Google Home Mini[2] as a hardware device and frontend for the speech dialogue. Our custom backend service is registered inside our Google Dialogflow application, enabling us to manage the dialogue and directly evaluate the given answers in real-time. Since not all tasks of the MMSE are speech based (some contain handwriting and sketching), we connected a Neo N2 smartpen[3] to our system. This allows us to implement the entire test very close to its original form.

An overview of our system's architecture is depicted in figure 1. The smartphone and the Google Home Mini device are connected to a local wireless network with internet access. A special, almost invisible pattern is printed on the sheets of paper, allowing the camera inside the digital pen to track its movements and stream the handwriting input via Bluetooth to the smartphone. Once received at the backend server the handwriting input is analyzed depending on the current MMSE task. For the analysis of text and sentences we use a state-of-the-art handwriting recognition engine by MyScript[4]. Symbols, such as the pentagrams are analyzed using a multitude of methods, e.g., a sliding window approach is used to determine points of interest, such as corners and crossings. Individual segments are analyzed using the MyScript shape recognizer and additional properties, such as angles are calculated between line segments. Overall, we devised a set of algorithms and classifiers that try to

cover imprecise requirements from the original scoring system, such as "All 10 angles must be present and two must intersect." and create unbiased scores based on the raw handwriting input.

In addition, we take into account multiple sensor data provided by the smartpen, which currently cannot be measured easily by the human observer. Information about velocity, pressure, air-time and other features can bring vital insights about the patient's cognitive state and handwriting behavior. We are currently in the process of extending our system to provide such feedback to clinicians, based on more than one hundred distinct handwriting features [11].

Speech Dialogue

For the speech dialogue system, we use the automatic speech recognition provided by the Google Dialogflow framework, where we registered our backend server as a webhook. Callbacks, that we receive, contain so called *intents*, which are a representation of a specific action that the user can invoke by using one of the defined terms in the Dialogflow console. These intents are analyzed in the *Intent Resolver* (e.g., relevant parameters/keywords are extracted and pre-processed) and passed on to the *Dialogue Management*. Based on the analysis of user utterances and current task the dialogue manager generates responses, which are passed back to the hardware device using the webhook.

In our experience, a long linear dialogue such as the MMSE, cannot be easily implemented by using only the Dialogflow framework. Provided dialogue examples are mostly concerned with shorter question answering interaction pairs, such as "Ok Google, how is the weather?", leaving not enough room for developers to implement more complex dialogues. For each question or task of the assessment, we created a Dialogflow intent using the Dialogflow console, leaving the actual dialogue management to our backend service. In total, we represent the MMSE using the 23 intents listed in sidebar 2.

Intents may also include parameters/entities, which helps in checking if the given answer fulfills a specific type (number, season, month, ...) or a specific keyword ("lemon", "wristwatch", ...). In total, we have manually added 14 entities via the Dialogflow console, because certain necessary types were not provided out-of-the-box:

- *Season/month/day:* the MMSE allows for only very limited possibilities for each answer (four seasons, twelve months, and seven days); we included all of them with a few variations, e.g., "late summer".
- *Country/state/city-names:* we added a few states and nearby country/city names for our study, as the pre-defined system produced many recognition errors for out-of-vocabulary words.
- *3 specific keywords* which the user needs to repeat and recall: "lemon", "key", and "ball".
- *2 depictive keywords* that the patient needs to identify and name: "wristwatch" and "pencil".
- *Clarification entities:* "What (did you say)?" / "(Can you) please repeat?" / "What was the task (again)?" / "I did not understand." etc.

19 questions/tasks of the MMSE
Year, season, month, day of week, country, state, city/town, location, floor, repeating words, spelling (backwards), recall words, name object "wristwatch", name object "pencil", repeat the phrase, read and follow the instruction, fold a paper, write a full sentence, copy the figure.
Default fallback intent
If the transcribed content includes no words or only noise, Dialogflow will ask the user to repeat what he/she said.
Custom fallback intent
If the input does not match any intent our server will respond based on the current state with a clarification question.
Repeat intent
When the user asks, "What did you say?" our dialogue manager reacts based on the current state and, e.g., repeats the question or task.

Sidebar 2: The Dialogflow intents that we created for the MMSE.

Figure 2: Experimental setup: Google Home Mini, smartpen, smartphone and a stack of prepared paper

Our multimodal speech-based dialogue system allows three kinds of turns:

- *Simple instruction/question:* year, season, month, etc. Even if the user said something unrelated or answered wrong, the next instruction followed. Only if the user asked something like "I did not understand" / "can you repeat that?" the instruction was repeated.
- *Complex instruction:* read and follow the instruction, fold a paper, write a full sentence, copy the figure. Users had three options: (1) move to the next task by saying "finished" / "done". (2) ask what the instruction was, then the instruction was repeated. (3) say something else. The system would then repeat the last instruction.
- *Specific instruction:* the user was asked to repeat three specific words. Up to three tries (exactly like in the original MMSE) are permitted. If he does not succeed or successfully repeats the three words, the system continued with the next instruction or task.
- *Control keywords:* Some tasks can only finish as soon as the user says "Done.". As there are several synonyms for done (for instance, "finished") we added an extra entity.

USER EXPERIENCE STUDY

In order to compare the traditional version of the Mini-Mental State Examination with our speech dialogue system, we conducted a user study with 15 participants and 3 different experimenters to prevent human bias. Among the test group were 5 female and 10 male subjects, ranging from 19 to 78 years of age. Five of the participants were recruited from a pool of voluntary elderly people of a geriatric clinic (age >65), while the rest were co-workers and students of varying age (age <60). We admitted undergraduate, graduate and post-doctoral subjects from varying domains and fields of work, including retired and active workers. The system has been implemented in German and all participants were native speakers without cognitive impairment.

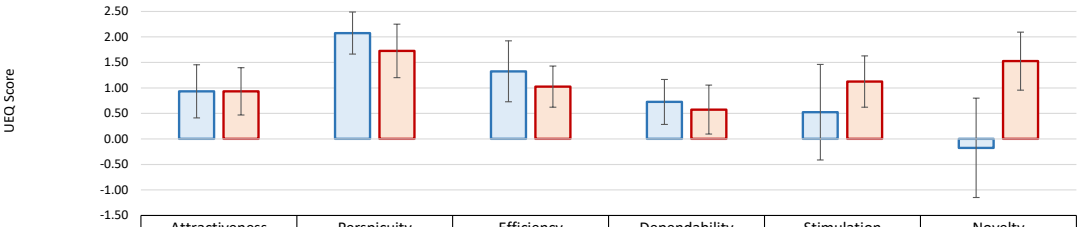
Procedure

Our experimental setup consists of two modes, a traditional administration of the MMSE with a human test administrator and a session of the automated speech dialogue system. Out of the 15 participants we selected 7 at random to start with the automated version and 8 to start with the human variant. After each mode, the subjects were asked to fill out a System Usability Scale [2] and a User Experience Questionnaire [12]. Directly after the first mode the experimenter explained the second mode, and in the end, subjects were given the opportunity to discuss the experiment and give remarks. As required by the MMSE instructions, the experiment took place in a quiet, distraction-free room, with the participant sitting comfortably on a chair at a table, opposite to the experimenter.

Task Design

The original MMSE by Folstein et al. [6] was administered without changes and as required by the instructions. A trained human expert leads the subject through the questionnaire while taking notes on performance for later grading. The MMSE is divided into two sections. During the first part, only speech-based responses are required from the participant. Questions include an assessment of orientation to place and time, after which aspects of memory and attention are tested. In this part subjects can reach a maximum of 21 points. The second part tests the ability to name, follow verbal and written commands, write a sentence spontaneously, and copy overlapping pentagons (see figure 2) using pen and paper [6]. Participants can score up to 9 points in the second part. The maximum score of the MMSE is 30 points, which is the sum of all sub-tasks. To prevent human bias (e.g., the length and intonation of instructions) 3 different testers were chosen, of whom each did 5 assessments.

For the automated speech dialogue the table was prepared as depicted in figure 2, with the smartphone out of reach as it was not required to interact with it. A prepared stack of papers was



	Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
Human	0.93	2.08	1.33	0.73	0.53	-0.18
Speech Dialogue	0.93	1.73	1.03	0.58	1.13	1.53

Figure 3: Results of the User Experience Questionnaire (UEQ); mean values and 5% conf. intervals.

placed upside down on the table and the subject was instructed not to touch the stack unless instructed by the system. The system would eventually ask the user to turn certain pages, some of which contain items that have to be named (wristwatch and pencil), while others contained written tasks ("Close your eyes.") or empty space for copying the overlapping pentagons drawing (see figure 2). In the backend our system scored the transcribed content of the speech recognition component automatically, by checking the given answers for correctness. Input from the digital pen was also automatically scored by our sketch recognition component as described above. Transcribed content and scores from individual tasks were presented to the test administrator in a structured PDF format, including the total score, which was automatically calculated.

RESULTS

Measuring the task completion time of both modes across all participants, we found that patients completed the traditional version (M=04:18 min, SD=0.013) on average in a similar time as the automated speech dialogue (M=04:11 min, SD=0.018). Our elderly subjects finished both modes on average just as fast as the younger participants. However, because the speech dialogue system can be performed without a physician present and the results are automatically scored in real-time, the administration time and manual scoring time drops from 5 to 10 minutes (based on the experience of the tester) down to 0, resulting in a noticeable time saving for the test administrator.

Evaluation of the User Experience Questionnaire (see figure 3) revealed no clear differences in attractiveness, perspicuity, efficiency, dependability and stimulation, but a significant difference in novelty ($t(9)=3.65$; $p=0.005$). Based on the System Usability Scale, which grants up to 100 points, users rated the person-to-person mode (avg. score: 78.75) and the speech dialogue (avg. score: 73.25) both as *good* (SUS grading scale). In terms of usability both systems perform equally well. A total of 80% of

participants reported that they were more stressed during their first mode than during the second. As we could not correlate their answers to the specific order in which the tests were conducted, we assume that this is due to the learning effect and uncertainty what to expect during the first run.

Several subjects reported that they adjusted their pronunciation and emphasis due to their past experience with speech dialogue systems, they expected the system to understand them better this way. All participants reported that the system's voice was clearly understood, except for the pronunciation of the word "Ball" (German for "ball"), which was misunderstood by subjects on three separate occasions. With the absence of a direct indication that questions could be asked at any point during the automated speech dialogue, two thirds of the users did not know that they could have asked for clarification, such as "can you repeat that?". One third of the participants used some form of clarification question and later expressed the opinion that the system answered adequately.

Regarding the comparison of manual and automated scorings we find that most of our participants scored very high marks in both modes (N=14). We had to dismiss one sample due to technical issues. The system awarded on avg. 28.29 points in total (SD=1.73), whereas human raters averaged 29.10 points (SD=1.49). In 21.4% (3 out of 14) of cases both systems scored exactly the same result and in 42.9% (6 out of 14) of cases results only differed by 1 point. The most frequent source of discrepancies was the spelling task. The employed automatic speech recognition component failed in identifying individually spelled letters and instead often wrongly recognized other words. We encountered 7 individual answers where the Dialogflow framework got stuck and input was either not recognized at all or no result was transmitted to our backend server.

DISCUSSION & FUTURE WORK

In this case study we presented a multimodal speech-based dialogue system for administering the Mini-Mental State Examination and automatically evaluating the results in real-time. Such systems have a high impact as they enable inexpensive, time-efficient and unbiased standardized data collection tasks and large-scale dementia screening programmes. Instead of conducting time-intensive problem modelling tasks with domain experts and potentially developing new cognitive assessments, we succeeded in digitalizing an already existing, popular and validated testing method. Our user experience study showed that the proposed method has satisfactory usability among the target audience, and that the assessment scores produced by the automated system are comparable to the ones created by human experts. As the task design of other popular tests such as the Montreal Cognitive Assessment (MoCA) only differs slightly, we are confident that our system can be adapted easily to other settings, as well as to other languages.

One limitation of our system is the inability to react to the subject's state of mind. Where human testers would adjust the speed, intonation or reassure a patient, the system cannot currently do so. We are working towards extending the existing system with additional sensor inputs to provide improved

Acknowledgements
This research is part of the Intera-KT project, which is supported by the Federal Ministry of Education and Research (BMBF) under grant number 16SV7768.

feedback of patient's cognitive state. In our setup the smartphone could be used to capture the video signal and analyze facial expression for further diagnosis of medical conditions [3] and state of mind using frameworks like OpenFace [1]. Continuing in that direction we plan to analyze the already captured audio data using openSMILE to detect emotions [5] and react to signs of distress.

REFERENCES

[1] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*. 59–66. https://doi.org/10.1109/FG.2018.00019

[2] John Brooke. 2013. SUS: A Retrospective. *J. Usability Studies* 8, 2 (Feb. 2013), 29–40.

[3] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. 2009. Detecting depression from facial actions and vocal prosody. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. 1–7. https://doi.org/10.1109/ACII.2009.5349358

[4] Randall Davis, David J Libon, Rhoda Au, David Pitman, and Dana L Penney. 2014. THink: Inferring Cognitive Status from Subtle Behaviors. , 2898–2905 pages.

[5] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In *Proceedings of the 21st ACM International Conference on Multimedia (MM '13)*. ACM, New York, NY, USA, 835–838. https://doi.org/10.1145/2502081.2502224

[6] M. F. Folstein, S. E. Folstein, and P. R. McHugh. 1975. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* 12, 3 (Nov 1975), 189–198.

[7] Lindy E. Harrell, Daniel Marson, Anjan Chatterjee, and Jo Ann Parrish. 2000. The Severe Mini-Mental State Examination: A New Neuropsychologic Instrument for the Bedside Assessment of Severely Impaired Patients With Alzheimer Disease. *Alzheimer Disease & Associated Disorders* 14, 3 (2000).

[8] Mira Niemann, Alexander Prange, and Daniel Sonntag. 2018. Towards a Multimodal Multisensory Cognitive Assessment Framework. In *31st IEEE International Symposium on Computer-Based Medical Systems, CBMS 2018, Karlstad, Sweden, June 18-21, 2018*. 24–29. https://doi.org/10.1109/CBMS.2018.00012

[9] Sharon Oviatt, Björn Schuller, Philip R Cohen, Daniel Sonntag, Gerasimos Potamianos, and Antonio Krüger. 2017. Introduction: Scope, Trends, and Paradigm Shift in the Field of Computer Interfaces. In *The Handbook of Multimodal-Multisensor Interfaces*. Association for Computing Machinery and Morgan & Claypool, New York, NY, USA, 1–15.

[10] D. Palsetia, G. P. Rao, S. C. Tiwari, P. Lodha, and A. De Sousa. 2018. The Clock Drawing Test versus Mini-mental Status Examination as a Screening Tool for Dementia: A Clinical Comparison. *Indian J Psychol Med* 40, 1 (2018), 1–10.

[11] Alexander Prange, Michael Barz, and Daniel Sonntag. 2018. A categorisation and implementation of digital pen features for behaviour characterisation. *CoRR* abs/1810.03970 (2018). arXiv:1709.01796 https://arxiv.org/abs/1810.03970

[12] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2014. Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios. In *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience*, Aaron Marcus (Ed.). Springer International Publishing, Cham, 383–392.

[13] Daniel Sonntag. 2017. Interakt - A Multimodal Multisensory Interactive Cognitive Assessment Tool. *CoRR* abs/1709.01796 (2017). arXiv:1709.01796 http://arxiv.org/abs/1709.01796