Speech and language technologies for the automatic monitoring and training of cognitive functions

Anna Pompili, Cristiana Amorim, Alberto Abad, Isabel Trancoso

L^{2}F - Spoken Language Systems Lab, INESC-ID Lisboa
IST - Instituto Superior Técnico, University of Lisbon
{anna,cristiana.amorim,alberto,imt}@l2f.inesc-id.pt

Abstract

The diagnosis and monitoring of Alzheimer's Disease (AD), which is the most common form of dementia, has been the motivation for the development of several screening tests such as Mini-Mental State Examination (MMSE), AD Assessment Scale (ADAS-Cog), and others. This work aims to develop an automatic web-based tool that may help patients and therapists to perform screening tests. The tool was implemented by adapting an existing platform for aphasia treatment, known as Virtual Therapist for Aphasia Treatment (VITHEA). The tool includes the type of speech-related exercises one can find in the most common screening tests, totalling over 180 stimuli, as well as the Animal Naming test. Its great flexibility allows for the creation of different exercises of the same type (repetition, calculation, naming, orientation, evocation, ...). The tool was evaluated with both healthy subjects and others diagnosed with cognitive impairment, using a representative subset of exercises, with satisfactory results.

1. Introduction

Alzheimer's Disease (AD) is a neurodegenerative disease which represents 60 to 70% of the dementia cases in Portugal [1, 2, 3]. However, its first signs can go unnoticed [1, 2, 4, 5]. Typically, AD is known to cause alterations of memory and of spacial and temporal orientation [3, 4, 6]. Furthermore, AD increases dramatically with age and it has no cure. Nevertheless, an early diagnosis may slow down its progression by enabling a more effective treatment [5]. For this purpose, several neuropsychological tests exist in the literature, each targeting different cognitive domains and capabilities. The Mini-Mental State Examination (MMSE) [7] and the AD Assessment Scale - Cognitive subscale (ADAS-Cog) [7] are two of the most popular tests used in Portugal for screening cognitive performance and tracking alterations of cognition over time. They involve the assessment of different capabilities, such as orientation to time and place, attention and calculus, language (naming, repetition, and comprehension), or immediate and delayed recall. Another type of test also commonly applied by therapists in the diagnosis of AD is the Verbal Fluency test [7]. In this test, the patient should produce as many words as he can beginning with a particular letter (phonemic fluency test) or belonging to a particular category, e.g. fruits (semantic fluency test), during 60s. This test is used both in assessing the verbal initiative ability and executive function such as the inhibition ability, the difficulty in switching among tasks, and the perseverance attitude [7]. Typically, the most commonly used versions for the Portuguese population consider the letter "P" for the phonemic version and the "Animal" category for the semantic category. Other tests not so frequently adopted for this population are the Wechsler Adult Intelligent Scale - III (WAIS-III) [7], which provides a measure of general intellectual function in older adolescent and adults, and the Stroop test [7], which is a measure of cognitive control, evaluating how easily a person can maintain a goal in mind while suppressing habitual responses.

Most of these tests include a verbal component provided in response to a visual or spoken stimulus solicited by a therapist. Thus, due to their nature, and the need to continuously monitor the cognitive decline over time, these tests lend themselves naturally to be automated through speech and languages technologies (SLT). A tool including the digitized version of these tests with the possibility of an immediate evaluation through automatic speech recognition could be of valuable support in health care centres. The therapist will have access to an organized archive of tests which could be administered in the traditional way, or remotely, when the physical dislocation of the subject is hampered by logistic constraints or physical disabilities. Recordings and evaluations will be stored and made available for later consultation. On the other hand, research has shown that cognitive skills, which can fade without stimulation as we age, can be improved by playing games that stimulate brain activity [8]. An automated tool for the monitoring of AD could be easily extended to support exercises and brain games for the daily training of cognitive capabilities such as short-time memory, attention, calculus, reasoning ability and many others.

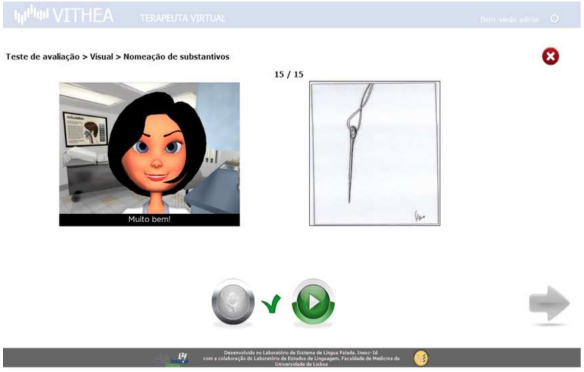
Up to our knowledge, there are few works in the literature that exploit SLT to automate certain types of neuropsychological tests. Some of the most relevant are the kiosk system designed to use at home as a prevention instrument for early detection of AD described in [5], the end-to-end system for automatically scoring the Logic Memory test of the WAIS-III presented in [9], and the system that implements a modified version of the MMSE based on the IBM ViaVoice recognition engine of [10]. These works show the recent increasing interest on this area, but also the long road ahead to support the large variety of existing neuropsychological tests (e.g., some of them are not fully automated).

This work makes a step towards filling this gap by introducing a set of neuropsychological tests for AD intended for the Portuguese population, which were integrated into an automatic web-based system [11]. The system presented in this work extends an on-line platform named VITHEA [12] used for aphasia treatment that incorporates SLT to provide word naming exercises. For this to be possible, the system resorts to a keyword spotting technique which consists of detecting a certain set of words by using a competing background model with the keywords model [13]. This platform is used daily by patients and speech therapists and has received several awards from both the speech and the health-care communities. The success of this platform and its flexibility, that allows to create different ex-

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 103–109,
Dresden, Germany, 11 September, 2015. ©2015 The Association for Computational Linguistics

ercises, have motivated its use as a foundation for this work. Our first step was the automation of the exercises in MMSE and ADAS-COG that involve speech. The second step was the implementation of the semantic fluency test, starting with the Animal category, also known as Animal Naming test. As explained in the next sections, the automation of such tests have raised several technological challenges, both for the automatic speech recognition and text-to-speech synthesis technologies.

In the following, Section 2 briefly presents the VITHEA platform that was used as a foundation for this work, while Section 3 describes the extended system resulting from the implementation of the selected neuropsychological tests into the VITHEA platform. Section 4 reports how each type of test was concretely implemented. Then, in Section 5 the focus is on the experiments, both detailing the automatic speech recognition module, the speech corpus used for evaluation and the experimental results. Finally, Section 6 presents the conclusions and future work.



Desenvolvido no Laboratório de Sistema de Língua Falada, Inesc-Id com a colaboração do Laboratório de Estudos de Linguagem, Faculdade de Medicina da Universidade de Lisboa

Figure 1: A caption of the VITHEA platform during the presentation of an exercise.

2. The VITHEA platform

VITHEA (Virtual Therapist for Aphasia Treatment) is a web-based platform developed with the collaboration of the Spoken Language Processing Lab of INESC-ID (L^2F) and the Language Research Laboratory of the Lisbon Faculty of Medicine (LEL). The system aims at acting as a ”virtual therapist”, allowing the remote rehabilitation from a particular language disorder, aphasia. For this to be possible, the platform comprises two specific modules, dedicated respectively to the patients, for carrying out the therapy sessions, and to the clinicians, for the administration of the functionalities related to them (e.g., manage patient data, manage exercises, and monitor user performance). In this way, speech therapy exercises created by speech therapists through the clinician module, can be later accessed by aphasia patients through the patient module with a web-browser. During the training sessions, the role of the speech therapist is taken by a ”virtual therapist” that presents the exercises and is able to validate the patients answers.

The overall flow of the system can be described as follows: when a therapy session starts, the virtual therapist shows to the patient, one at a time, a series of exercises. These may include either the presentation of images, the reproduction of short videos or audios, and textual information. The patient is then required to respond verbally by naming the contents of the object or action that is represented. The utterance produced is recorded, encoded and sent via network to the server side. Here, a web application server receives the audio file which is processed by the ASR system, generating a textual representation of the patient's answer. This result is then compared with a set of predetermined textual answers (for the given question) in order to verify the correctness of the patient's input. Finally, feedback is sent back to the patient with the correctness of the answer provided. Figure 1 illustrates the use of the VITHEA platform.

3. Extending VITHEA for neuropsychological screening

Extending VITHEA for including neuropsychological tests involved important alterations in the original platform, both on the patient and the clinician modules. However, the flexibility of VITHEA allows for the easy addition of new categories of exercises. These can then be combined in multiple ways by the clinician to form new tests, and to create different exercises of the same type. According to the original system, in order to answer the question presented by the virtual therapist, the patient needs to manually interact with the system to start and stop the recording of his answer, and to advance among different stimuli.

The usability of this interface has been adapted to meet the needs of an ageing population, with cognitive impairments. In particular, we considered important to implement the following updates:

- To simplify the interaction with the tool and make the evaluation process more fluid, we minimized the use of the mouse. The interface now automates part of the recording process and the progression between stimuli. An action from the patient is only required to stop the recording process.
- Since cognitive impairments and ageing often results in a limited auditory capability, the speech rate of the therapist has been tuned until finding the best compromise between a more understandable but still natural voice.

Also, the feedback from the neurologists involved in this work provided us important guidelines regarding the presentation of the tests. Following their advices, we introduced the alterations listed below:

- To make the interaction with the system more natural the virtual therapist now provides a random feedback to the patient when the evaluation switches among different classes of stimuli.
- Optional instructions have been added for the more complex questions.
- For some stimuli, the virtual therapist now provides a semantic hint if the patient has not provided an answer after a given amount of time.

The platform now allows also to store some additional personal information of the profile of the patient that are needed for the assessment of some sub-tests (i.e. place of birth, age, etc.), and the result of the assessment in terms of test score obtained. During the application of a neuropsychological test, the scores are individually calculated for each question. After the answer has been processed by the automatic speech recognition (ASR) system, the platform computes both the maximum score allowed for the current question, and the actual score obtained by the patient. These results are stored in the database. At the end of the test, both the maximum scores and the obtained scores for each question are summed separately to obtain a global score in the form $score/maxScore$ (e.g., a score of 18/22). This result can then be consulted by the patient. In order to follow the patient's progress, each time an evaluation test is performed, the platform sends an e-mail with a summary of the patient's performance to the therapist assigned to him/her.

Overall, these alterations contributed to building a simplified interface, suited for aged people, especially if cognitively impaired.

4. Automated tests

Since the selected neuropsychological tests comprise common or similar questions, we may approach its concrete implementation organized by type of question and the underlying technology with which they were implemented, rather than per test. Each type of question has set different challenges, each of which has been addressed individually with ad-hoc solutions. Overall, a total of 185 stimuli belonging to different types of tests have been selected for their implementation in the platform.

4.1. Naming objects and fingers

This type of stimuli belongs both to the MMSE and the ADAS-cog tests and evaluates a person's naming ability. Similarly to the exercises used in aphasia treatment, it consists of naming a series of objects that are shown in pictures, one at a time. These stimuli were implemented following a keyword spotting approach. A maximum score of 1 is given for each correct answer. The major innovation relative to the VITHEA exercises was the introduction of an optional semantic cue for some of the questions. This was implemented by adding a timer in the component responsible for the answer's recording and by making the virtual therapist to speak the cue after 20 seconds if no answer is detected. For this to be possible, both the clinician module and the internal structure of the database had to be extended for managing and storing the additional information. In fact, since the recording process is started from the beginning, the semantic cue is also recorded together with the patient's answer. Consequently, the logic of the patient module had also to be updated in order to remove the semantic cue spoken by the virtual therapist.

4.2. Repetition

The repetition question is part of the MMSE test and consists of repeating the following sentence: "*O rato roeu a rolha*" (the mouse gnawed the stopper). This question could be easily implemented with a keyword/key-phrase spotting approach, just like the ones for aphasia treatment. The maximum score is 1, which corresponds to a sentence correctly repeated.

4.3. Attention and calculation

This type of question belongs to the MMSE test, the idea is to successively subtract 3 beginning on 30 until 5 answers are given. In our first approach, we created a set of 5 different stimuli, each one asking separately for a specific calculation. These questions were also implemented with a keyword spotting approach. A score of 1 is given for each stimulus that corresponds to a correct answer, for a maximum score of 5.

4.4. Orientation to time, place and person

These type of stimuli are part both of the MMSE and the ADAS-cog, though some questions differ. They comprise stimuli intended to evaluate a person's orientation ability, asking the patient to report the current year, day, month, his name, the country and the town he lives in, among others. These are dynamic questions in the sense that there is not a universal answer to each question as it changes depending on the time, place and person. The solution was to provide several pre-compiled language models that were carefully structured so that, at any time, the platform knows which is the right model to chose. For the questions of orientation to person, the necessary information is acquired at the time of the creation of the user profile and then it is used to automatically generate the corresponding language models. The majority of these questions were implemented based on a standard keyword spotting approach. However, for the day and hour, it was necessary to create dedicated rule-based grammars. A correct answer is always scored with 1 point, while an incorrect answer scores 0.

4.5. Word recognition

The word recognition stimuli belong to the ADAS-Cog test and consist of presenting the patient a list with 12 words to learn, one at a time. Words are written in block letters on white cards. The learning process is made by asking the patient to read each word aloud and try to remember it. Then, a new list with 24 words is shown in the same way. This new list contains the 12 original words of the learning list, plus 12 new distracting words that are carefully chosen in terms of phonetic similarity and semantic meaning. For each word, the patient is then asked to indicate whether it was on the learning list or not. This whole process is repeated in 3 trials. Just like the day and hour questions, rule-based hand crafted recognition grammars for positive, negative or neutral answers were built. For the word recognition task itself, each presented word is individually scored. Specifically, a correct answer corresponds to a maximum score of 1, which yields a total score of 72 for the word recognition sub-test (i.e., 24 for each trial).

4.6. Evocation

Generally speaking, an evocation question consists of recalling a series of words, whether they have been previously learned or if they are subject to compliance with certain requirements. In either cases, the spoken answers produced for this kind of stimuli are commonly followed by filled pauses, i.e. hesitation sounds. For this reason, we adopted a keyword spotting approach that incorporates an ad-hoc model to deal with filled pauses. In terms of score, the calculation is processed by considering the number of keywords that are correctly produced, without repetitions.

For MMSE, the evocation question consists of the immediate and delayed recall of 3 words. This was implemented with an auditory stimuli for the immediate recall task and with a textual stimuli for the delayed recall task. The presentation of the evocation question that belongs to the ADAS-cog is very similar to the word recognition question. Basically, it consists of the immediate recall of a list with 10 words that were previously learned, the whole process is repeated in 3 trials.

4.7. Verbal Fluency

The animal naming question, which belongs to the Verbal Fluency test, is the most challenging among the evocation tests. This is explained by the fact that, contrarily to the other cases, which are based on a limited domain vocabulary tasks, this question comprises a more extended domain composed of the name of all known species of animals. Theoretically, the language model should cover all the known species of animals, however, in practice this is infeasible. Moreover, the size of the language model directly impacts the output of the ASR system. In fact, while a shorter list will cause the missing keywords to never be recognized, a longer list will increase instead the perplexity of the task.

The automatic creation of an ad-hoc language model for this type of question is detailed in [14], and is briefly reported here. The starting point consisted of an existing list of animal names [15] that included 6044 animal names, grouped, classified, and labelled with its semantic category, without inflected forms. Since some names are more likely and common than others, the initial list was used to build a probabilistic language model that exploited this information. The likelihood of each term was computed considering the total number of results that is returned by querying a web search engine. The retrieval strategy had to be refined several times in order to find the optimal approach, in fact initial queries have led to incorrect counts due to homonyms of some terms. The final approach consisted in using the animal name and the semantic category associated. Finally, the likelihood associated with each term also allows to sort the list numerically and thus to reduce its size by filtering out less popular terms. After several experiments, the language model that achieved the best results contained the 802 most popular animal names.

5. Experimental set-up

5.1. ASR/KWS system

The monitoring tool integrates the in-house ASR engine named AUDIMUS [16, 17], a hybrid recognizer that follows the connectionist approach [18]. The baseline system combines three MLP-based acoustic models trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), log-RelAtive SpecTrAl features (RASTA, 13 static + first derivative) and Modulation SpectroGram features (MSG, 28 static). These model networks were trained with 57 hours of downsampled Broadcast News data and 58 hours of mixed fixed-telephone and mobile-telephone data in European Portuguese [19]. The number of context input frames is 13 for the PLP and RASTA networks and 15 for the MSG network. Neural networks are composed by two hidden layers of 1500 units each one. Monophone units are modelled, which results in MLP networks of 39 soft-max outputs (38 phonemes + 1 silence) [12].

In order to support a keyword spotting approach, the baseline ASR system was modified to incorporate a competing background speech model that is estimated without the need for acoustic model re-training. In fact, while keyword models are described by their sequence of phonetic units provided by an automatic grapheme-to-phoneme module, the problem of background speech modelling must be specifically addressed. Here, the posterior probability of a background speech unit is estimated as the mean probability of the top-6 most likely outputs of the phonetic network at each time frame. In this way, there is no need for acoustic network re-training.

5.2. Portuguese cognitive impaired speech corpus

To evaluate the feasibility of the monitoring tool, we collected an ad-hoc speech corpus. This includes recordings of 5 people diagnosed with cognitive impairments and 5 healthy control subjects. All the participants are Portuguese native speakers.

Recordings took place in different environments with different acoustic conditions. In fact, healthy subjects were recorded in a quiet, domestic environment, while patients were recorded at CHPL, the Psychiatric Hospital of Lisbon. No particular constraints were imposed over background noise conditions. Each session consisted approximately of a 20 to 30-minutes recording. The data was originally captured with the platform at 16 kHz, and later down-sampled to 8 kHz to match the acoustic models sampling frequency. The collection of the patients data, besides being emotionally demanding, it is a valuable resource which implied logistic difficulties.

Table 1: Speech corpus data, including gender, age, education and diagnosis. B.E.: Basic Education, S.E.: Secondary Education, MCI: Mild Cognitive Impairment, AD: Alzheimer Disease, PTD: Post-traumatic Dementia

User	Gender	Age	Education	Diagnosis
1	M	86	B.E. - 1st Cycle	MCI
2	F	71	B.E. - 1st Cycle	AD
3	M	60	B.E. - 1st Cycle	PTD
4	F	79	Illiterate	AD
5	M	80	S. E.	MCI
6	F	67	B.E. - 1st Cycle	Healthy
7	F	72	B.E. - 1st Cycle	Healthy
8	M	76	B.E. - 1st Cycle	Healthy
9	F	74	B.E. - 1st Cycle	Healthy
10	M	76	B.E. - 1st Cycle	Healthy

5.3. Evaluation results

Due to the extensiveness of the ADAS-cog test, it was infeasible to evaluate all the implemented neuropsychological tests. In fact, we estimated that the total duration of the evaluation would have been more than two hours, which was considered unacceptable. Thus, only a representative subset of all the tests has been selected, comprising a total of 41 stimuli. We have considered different individual evaluation metrics, depending on the type of automated tests and a global evaluation focused on the targeted final application.

5.3.1. Evaluation of KWS-based tests

The Word Verification Rate (WVR) was used to assess the performance of the automatic evaluation module in the tests based on keyword spotting (KWS). This metric provides a measure of the reliability of the platform as a verification tool. In order to compute it, both manual and automatic transcriptions are processed to indicate, for each utterance, if the expected keyword has been said or not. Then, the WVR is computed for each speaker as the number of coincidences between the manual and automatic result (either true acceptances or true rejections) divided by the total number of exercises. Thus, a result closer to 1 is desirable. Table 2 presents the WVR computed for each speaker on all the tasks based on keyword spotting. Results are provided separately for those tests that rely on simple KWS (word-lists) and those based on rule-based grammars with competing background model (i.e.: hours, date, yes/no, etc.). In general, we can consider these results quite promising. In fact, they are comparable to those reported in [13], in an evaluation with aphasia patients. In this case, the average verification rates are considerably better with healthy users, which was expected due to the more challenging characteristics of the patients' data. Nevertheless, the performance achieved with cognitive impaired users is still quite promising. On the other hand, no significant differences can be observed regarding the KWS strategy (word-list vs. rule-based grammar).

Table 2: WVR by speaker for keyword spotting exercises.

Patients

User	KWS (word-list)	KWS (rule-based)
1	0.78	0.79
2	0.78	0.93
3	0.74	0.71
4	0.91	0.50
5	0.65	0.79
Avg. WVR	0.77	0.74

Healthy

User	KWS (word-list)	KWS (rule-based)
6	0.91	0.86
7	0.91	0.93
8	0.91	0.93
9	0.87	0.86
10	0.83	0.86
Avg. WVR	0.89	0.89

5.3.2. *Evaluation of evocation tests*

The evocation exercises differ from the keyword spotting exercises in the sense that the answers are not evaluated as right or wrong, but instead the number of terms correctly recalled is counted. For this reason, we started by evaluating the Word Error Rate (WER) between the reference (manual) and the hypothesis (automatic) users' answer. Evocation exercises are divided into two categories: they may contain a limited number of words to recall or, contrarily, they may consider an open domain of possible answers complying to a specific semantic domain (e.g., Animal Naming test). Thus, the evaluation was processed separately for the two categories. The average WER computed for patients and control group in the class of evocation exercises with a closed domain was 20.00% and 8.16%, respectively. However, the average WER computed for patients and control group on the Animal Naming test was much higher, 65.12% and 46.48%, respectively. After a closer analysis, we noticed that the substitutions were the main source of error. This may be explained by the poor language model used in this type of question, since this is based on an extensive list of unigrams. Basically, the size of the list impacts greatly the performance of the ASR system by increasing its perplexity. Moreover, the list comprises uncommon animal names and some of them are quite short, which implies that even a small sound may be detected as an animal. It is interesting to notice, however, that although the results in terms of WER are clearly unsatisfactory, and demand further research, the number of animals recognized is not so different from the reference number of animals actually said.

5.3.3. *Global evaluation*

An evaluation analysis of the automatic tests closer to the final targeted application is necessary to better assess their possible applicability as part of an automatic screening platform. For a sub-set of the tests, a straightforward evaluation method consists of comparing the total manual and the automatic scores achieved by the user according to the scoring values described in section 4 for each type of test. In particular, the Mean Absolute Error (MAE) and the Mean Relative Absolute Error (MRAE) is used to measure the differences between the overall scores computed manually and automatically. The scores were calculated according to the traditional assessment that is performed when applying a neuropsychological test. Table 3 reports the MAE and MRAE for the previously reported subsets of stimuli and the maximum possible score for each test set, which corresponds to the maximum error achievable, in addition to the results for two specific screening tests: the MMSE and the Animal Naming test. For these two tests, the scores achieved by each speaker are also shown in Figures 2 and 3.

Table 3: MAE and MRAE (in brackets) by type of question and by neuropsychological test.

Question type / Test	Max. Score	Patients	Healthy
KWS (word-list)	23	3.00 (26%)	2.60 (12%)
KWS (rule-based)	14	2.80 (37%)	1.60 (15%)
Evocation (w/o animals)	11	0.80 (23%)	0.80 (11%)
MMSE	22	2.20 (21%)	2.80 (14%)
Animal Naming	∞	2.60 (24%)	1.80 (17%)

In general, the achieved results were better for healthy people than for patients. This is an expected result due to the impaired condition of the patients, which are reflected on the quality and coherence of their speech. The most common symptoms, even in less impaired subjects, are a reduced intensity, a reduced pitch, and a hoarse voice. Besides, quite often during the evaluation, patients started talking of general topics of their interest not related with the question under evaluation. It was also the case that sometimes the subject uttered his answer when the virtual therapist was still explaining the stimulus, thus resulting in overlapped speech. Finally, differently from healthy subjects, patients sometimes changed their mind while they were answering a question. This may increase the perplexity of the ASR result, especially when dealing with rule-based keyword spotting questions, due to the added complexity of the language models.

The MAE error reported for question type and test ranges from 0.80 to 3.00 for the patients, and from 0.80 to 2.80 for the control group. In relative terms, the mean relative error with respect to the manual scores (MRAE) ranges from 21% to 37% for the patients group, and from 12% to 17% for the control group. Notice that, since the scores achieved by healthy users are generally higher and since there are few differences between the two groups in terms of MAE, the MRAE for patients is considerably higher. Alternatively, it is worth comparing the MAE with the maximum possible score. This value depends on the number of stimuli selected for each test. For the Animal Naming test, the maximum score could not be computed since the number of elements a subject is able to name in the given time is unknown. In general, it can be observed that the difference between the automatic and the manual evaluation is relatively small compared to the maximum score. For instance, we can observe that the MAE for the questions based on keyword spotting is 3.00 out of 23 for the patients, which corresponds to 13%, and 2.60 out of 23 for control subjects, which corresponds to 11.3%. Overall, we consider these results a quite good performance, suggesting that the platform may be useful and reliable as a monitoring tool.

5.4. Discussion about the platform

The conducted evaluation and data collection also allowed us to collect important feedback about the platform itself. In fact, we

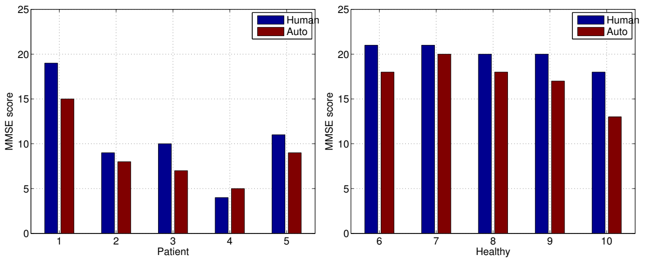


Figure 2: On the left side, MMSE scores of the human and automatic evaluation for the patient speakers. On the right side, MMSE scores of the human and automatic evaluations for the healthy speakers.

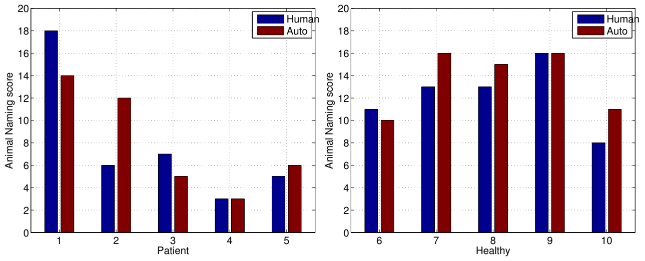


Figure 3: On the left side, Animal Naming scores of the human and automatic evaluation for the patient speakers. On the right side, Animal Naming scores of the human and automatic evaluations for the healthy speakers.

noticed that an advanced impaired condition may render more difficult the use of the system, specially when combined with deafness and with computer illiteracy, two factors that are associated with ageing. Patients with a more pronounced cognitive impairment or with auditory impairments, may have difficulties in understanding the question being asked. The computer illiteracy, however, may no longer be a problem in the not so distant future. Nevertheless, we expect that this tool will have adhesion for its usefulness and relevance. In fact, during test application, both the patients and healthy people demonstrated their appreciation for the platform, indicating that this is an interesting and appealing system. Moreover, they showed their interest in repeating the tests and using the platform regularly. Particularly, some of the patients were captivated by the animated virtual character, they liked its cartoon nature and the fact that it interacted with them verbally. This factor, together with the flexibility of the platform, let us think that in a near future the platform could be successfully turned in an environment useful both for training and monitoring cognitive skills. In fact, the kind of exercises that were adapted in the current version of the platform could be easily extended to the kind of games that are useful for stimulating brain activity, such as attention, memory etc. Further, the platform also allows to store recordings and evaluation results of each patient and make them available in an organized way, which can be useful for later consultation and comparison both by patients and by clinicians.

Finally, this platform raises interesting questions of ethical nature, i.e. whether such an automated tool should directly provide patients with a diagnosis similar to the one given by a clinician, or whether medical diagnosis should rather be produced exclusively by human doctors. One key related question is that diagnosis of mental disorders should always keep into account also normative data related with the language and education level of the patient. While we envision the possibility to incorporate the evaluations of such factors in future versions of the platform, these are not currently encompassed by our system. Finally, another important ethical question is whether patients should *always* be presented with the results of the automated tests. One may indeed argue that, in particular in presence of negative outcomes, the sensitivity of patients may be hurt and that, in such situations, it may be advisable to avoid exposing directly the tests' results to the patient and contact, instead, his/her relatives.

6. Conclusions and Future Work

In this work we developed an automatic web-based tool with SLT integration which could be used for monitoring cognitive impairments. The platform automates a set of neuropsychological tests that are commonly applied by therapists to assess the cognitive condition of a person. As far as we know, it is the only platform of this type implemented for the Portuguese population. The system has been assessed both with healthy subjects and patients. The mean absolute error between the manual and the automatic evaluation was relatively small, showing the feasibility of such type of system. We believe that this platform could be helpful for therapists and patients in the diagnosis of the disease. Its flexibility also allows the very easy creation of new exercises of the same type, with different stimuli. Besides, it could be easily extended to include different types of exercises that can be used for the daily training of cognitive abilities. For these reasons, we think that this tool could be an added value for society, helping in the prevention and in the early diagnosis of AD and mild cognitive impairments.

As future work we wish to remove completely the mouse interaction with the platform during the test application, automatically detecting when to stop the recording through silence detection technique. As the test currently already advances on its own when the recording is stopped, implementing this modification would enable to perform a complete test without any interaction, in a more agile way. Also, we plan to address the verbal fluency task, for which the preliminary results of the baseline system show much room for improvement. Finally, as mentioned in Section 5.4, medical diagnosis of dementia should keep into account also normative data related with the language and education level of the patient. In this sense, one open research question is how to automate the evaluation of these factors and incorporate them in the final diagnosis emitted by the system. Further, it would be desirable to extend the platform to incorporate intelligent filters aimed at identifying critical/negative outcomes, whose disclosure to the patient may risk hurting his/her sensitivity.

7. Acknowledgements

The authors of this work want to express their deep gratitude to Dr. José Salgado, clinical director of the Psychiatric Hospital of Lisbon Julio de Matos, for having made this work possible and for having provided us constant and full support. We also want to thank Dr. Inês Cunha and Dr. Vitorina Passão for their kind availability and precious feedback. This work was supported by national funds through – Fundação para a Ciência e a Tecnologia (FCT), under Grants SFRH/BD/97187/2013 and with reference UID/CEC/50021/2013.

8. References

[1] B. Nunes, "A demência em números," in *A Doença de Alzheimer e Outras Demências em Portugal*, A. Castro-Caldas and A. Mendonça, Eds. LIDEL, 2005.

[2] P. Moreira and C. Oliveira, "Fisiopatologia da doença de alzheimer e de outras demências," in *A Doença de Alzheimer e Outras Demências em Portugal*, A. Castro-Caldas and A. Mendonça, Eds. LIDEL, 2005.

[3] I. Santana, "A doença de alzheimer e outras demências - diagnstico diferencial," in *A Doença de Alzheimer e Outras Demências em Portugal*, A. Castro-Caldas and A. Mendonça, Eds. LIDEL, 2005.

[4] J. Barreto, "Os sinais da doença e a sua evolução," in *A Doença de Alzheimer e Outras Demências em Portugal*, A. Castro-Caldas and A. Mendonça, Eds. LIDEL, 2005.

[5] R. Coulston, E. Klabbers, J. Villiers, and J. Hosom, "Application of speech technology in a home based assessment kiosk for early detection of alzheimer's disease," in *Proc. Interspeech*, 2007.

[6] M. Guerreiro, "Avaliação neuropsicolgica das doenças degenerativas," in *A Doença de Alzheimer e Outras Demências em Portugal*, A. Castro-Caldas and A. Mendonça, Eds. LIDEL, 2005.

[7] E. Strauss, E. Sherman, and O. Spreen, *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*, 3rd ed. Oxford University Press, 2006.

[8] J. A. Anguera, J. Boccanfuso, J. L. Rintoul, O. Al-Hashimi, F. Faraji, J. Janowich, E. Kong, Y. Larraburo, C. Rolle, E. Johnston, and A. Gazzaley, "Video game training enhances cognitive control in older adults," *Nature*, vol. 501, pp. 97–101, 2013.

[9] M. Lehr, I. Shafran, and B. Roark, "Fully automated neuropsychological assessment for detecting mild cognitive impairment," in *In Interspeech*, 2012.

[10] S. S. Wang, P. D, J. Starren, and P. D, "A java speech implementation of the mini mental status exam."

[11] C. Amorim, "Automatic tool for screening of cognitive impairments," Master's thesis, Instituto Superior Tcnico, June 2014.

[12] A. Abad, A. Pompili, A. Costa, I. Trancoso, J. Fonseca, G. Leal, L. Farrajota, and I. P. Martins, "Automatic word naming recognition for an on-line aphasia treatment system," *Computer Speech & Language*, vol. 27, no. 6, pp. 1235 – 1248, 2013, special Issue on Speech and Language Processing for Assistive Technology.

[13] A. Abad, A. Pompili, A. Costa, and I. Trancoso, "Automatic word naming recognition for treatment and assessment of aphasia," in *Proc. Interspeech*, 2012.

[14] H. Moniz, A. Pompili, F. Batista, I. Trancoso, A. Abad, and C. Amorim, "Automatic recognition of prosodic patterns in semantic verbal fluency tests - an animal naming task for edutainment applications," in *18TH INTERNATIONAL CONGRESS OF PHONETIC SCIENCES*, 2015.

[15] N. J. Mamede, J. Baptista, C. Diniz, and V. Cabarrao, "String: An hybrid statistical and rule-based natural lan- guage processing chain for portuguese," in *International Conference on Computational Processing of Portuguese Propor*, 2012.

[16] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "AUDIMUS.Media: a Broadcast News speech recognition system for the European Portuguese language," in *Proc. International Conference on Computational Processing of Portuguese Language (PROPOR)*, 2003.

[17] H. Meinedo, A. Abad, T. Pellegrini, I. Trancoso, and J. Neto, "The L^2F Broadcast News Speech Recognition System," in *Proc. Fala2010*, 2010.

[18] N. Morgan and H. Bourlad, "An introduction to hybrid HMM/connectionist continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 25–42, 1995.

[19] A. Abad and J. Neto, "Automatic classification and transcription of telephone speech in radio broadcast data," in *Proc.International Conference on Computational Processing of Portuguese Language (PROPOR)*, 2008.