New features for on-line aphasia therapy

AnnaMaria Pompili
Instituto Superior Técnico

Resumo Aphasia is a particular type of communication disorder caused by the damage of one or more language areas of the brain affecting various speech and language functionalities. A frequent syndrome is the difficulty to recall names or words. Typically, such problems can be treated through word naming therapeutic exercises. VITHEA (Virtual Therapist for Aphasia Treatment) is an on-line platform developed for the treatment of aphasic patients, incorporating recent advances of speech and language technologies to provide word naming exercises to individuals with lost or reduced word naming ability. The focus of this work is to investigate the feasibility of incorporating additional functionalities that may enhance the VITHEA system. These features aimed at both extending the usability of the system and strengthening its performance, and thus involve several heterogeneous areas of the project.

1 Introduction

Aphasia is a particular type of communication disorder caused by the damage of one or more language areas of the brain affecting various speech and language functionalities. Cerebral vascular accidents are one of the most common causes. A frequent syndrome is the difficulty to recall names or words. Typically, such problems can be treated through word naming therapeutic exercises. In fact, frequency and intensity of speech therapy is a key factor in the recovery, thus motivating the development of automatic therapy methods that may be used remotely.

VITHEA (Virtual Therapist for Aphasia Treatment) is an on-line platform developed for the treatment of aphasic patients, incorporating recent advances of speech and language technologies to provide word naming exercises to individuals with lost or reduced word naming ability. The project is used daily by patients and speech therapists and has been awarded from both the speech and the health-care communities. This success motivated the research on additional features which could extend its functionalities and robustness. The new features have concerned a heterogeneous domain of enhancements that includes, among others, the evaluation of new approaches to improve the recognition quality and the recording process, the development of new interfaces to improve user experience, and the experimental implementation of new types of exercises.

Overall, the results of user interaction satisfaction questionnaires and the automatic evaluations have provided encouraging feedback on the outcome of the developed improvements.
Due to space constraints, in this document we focus on the description of the solutions adopted for each one of the targeted functionalities in this thesis. For a detailed description of the state of the art and related work, please refer to Section 2 of the MsC thesis document Pompili (2013). In the following, Section 2 reports on the architecture and the design followed to develop a new version of the system supported by mobile devices. Section 3 details the options chosen for the implementation of a Voice Activity Detection (VAD) approach for the design of a hands-free interface. In Section 4, instead, the focus is on improving the management of the data of the system, by providing an advanced search feature. Section 5 explains the concept of evocation exercises and how a specific subclass, the animal naming, has been implemented. Then, Section 6 reports on the integration of an alternative lexicon to improve recognition performance. Finally, Section 7 presents the conclusions and future work.

2 Content adaptation for mobile devices

The client side of the VITHEA platform exploits Adobe®Flash® technology to record patients’ answers. This module, unfortunately, limits the extension of the system to mobile devices, such as tablets and smartphones. This is due to a limitation in the API provided by Adobe®. In fact, the microphone class, used to acquire speech input, is not supported by the Flash® player running in a mobile browser. Thus a specific version for these device has to be specifically designed and implemented.

2.1 Architectural overview of the implemented prototype

The final prototype has been built exploiting a Service Oriented Architecture (SOA). RESTful web services have been provided to expose the relevant functionalities to mobile client, adhering to the constraints that this architectural style requires.

Particularly, the stateless constraint requires that no persistent information is stored on the server side, meaning that no client context data is maintained between following requests. Each request from any client should contain all of the information necessary to service the request, and session state is held in the client. This implies important consequences for the application logic, among which having to change the authentication modality. In fact, the VITHEA system currently is only accessible after the user has authenticated properly into the application.

There are several solutions for handling authentication in a REST context. Some possible options are HTTP basic authentication over HTTPS and a dedicated login service. In the current prototype the authentication has been implemented following a hybrid approach close to basic authentication.

In the current implementation, the user digits its credential at the time of accessing the system, and these data are then stored in the client application
for the whole execution time. In each following requests, these informations are encrypted with the MD5 message-digest algorithm, added to the Authorization header field, and sent to the server together with the request over HTTPS. At the server side, data received is verified through the support provided by Spring for basic authentication. The access restriction to a given resource is performed directly at the configuration file level. When the request from the client is received the authorization header is checked for user credentials. When found, data present here is compared with the encrypted version of the same data that exists in the server persistent storage support. If the credentials are correct, the user is granted the access to the given request, otherwise access is denied.

2.2 Client application

The prototype developed in this work has to be considered a proof of concept of the feasibility of the mobile version and, thus, integrates only the features that are important for a correct, complete interaction. These include, of course, the integration of the recognition process, of the virtual therapist character, and the application logic that regulates the exercises flow, thus including listing and navigation, video and audio reproduction.

The virtual therapist’s representation is achieved through the 3D game engine Unity. This environment provides also the possibility to directly build plugins for Android. In this way, the native module of the therapist is exported and then integrated into an Android application easily. Thus, similarly to the standard version, in the mobile version the virtual character guides and interacts with the user by providing audio-visual feedbacks.

The speech recognition process is performed remotely by the in-house speech recognizer AUDIMUS. The audio is acquired through Android microphone, to this purpose Android API provides the AudioRecord class. This class is delegated at the management of the audio resources needed for Java applications to record audio from the input hardware of the platform. During acquisition, read data is stored into an internal buffer of the AudioRecord class. When the recording stops, the audio is sent to the server side through a REST-ful POST request. Here, the in-house speech engine processes the file and the result of the recognition is returned to the user. It is expected that the microphone quality of tablet devices could be of poor quality, thus degrading the quality of the recognition.

2.3 User experience evaluation

A user experience evaluation based on collected questionnaires has been conducted with 16 users, with different ages, varying from 23 to 60 years. Users selected for the evaluation have different background knowledge, ranging from computer science, linguistic and accounting. Overall, the evaluation provided good results, achieving an average score of 4.14 in a likert scale of 5 points (1 to 5). The items related with the usability of the system have an average score of 4.3, the ones related with the robustness achieved 4.7, while the items that compares
the appreciation of the mobile version with respect to the online browser version achieved 3.7.

3 Hands-free speech

In situations where arm mobility is affected, the support for a hands-free interface will possibly improve the overall usability of the user experience. In the particular case of the VITHEA project, since the user of the system is affected from a language disorder, hands-free computing could not be interpreted as an alternative way of interaction, instead it will be selectively applied to automate the process of recording the users answers, and thus, provide additional benefits to people experiencing disabilities. The implementation of an hands-free interface will strongly rely on VAD techniques.

3.1 Voice activity detection task

At a very general level, VAD is the binary classification process that tags each speech segment as containing voice or silence. In the most basic VAD approach, the signal is first sliced into contiguous frames, then a real value parameter is associated with each frame. If this parameter exceeds a certain threshold, the frame is classified as containing speech, otherwise it is classified as containing silence. Here we will follow the same methodology, however the algorithm has been adapted in order to keep into account specific application logic constraints. The measure used to establish if a frame is possibly containing speech is the energy of that frame. If the length of a frame is of $k$ samples, and being $x(i)$ the $i^{th}$ sample, then energy for a given frame $j$ of the input signal is computed as:

$$E_j = \frac{1}{k} \sum_{i=0}^{k} x^2(i)$$  \hspace{1cm} (1)

In the application scenario, VAD computation should not start until voice is detected, and should end after 3 seconds of silence. The first constraint is achieved verifying that a minimum number of frames have been classified as voice. When this is verified, we could define the status of the VAD algorithm as active. The second, to be satisfied, should verify that the VAD status is active and that, since the last voice frame, a minimum number of frames have been classified as silence. When this is achieved, the end of the speech is detected.

3.2 Experimental evaluation

The assessment of the performance of the VAD algorithm has been evaluated through offline tests. To this purpose, a speech corpus has been derived from daily patient’s recordings stored into the system. The evaluation process has been performed in the Matlab environment, simulating the same conditions of the Flash Player environment.
The VAD algorithm described above is aimed at detecting both the start and the end of speech. However, the accurate detection of the two boundary informations is not equally important for the purposes of our application. In fact, regarding the start of speech detection, the exact location of this boundary is not relevant for our application purposes, since the system starts to record automatically. A failure in the algorithm would be caused at this regard only if the start of speech is not detected or if it is very prematurely detected causing an error in the end of speech detection. In other words, the most important boundary to be detected is the end of speech. However, not all the errors in the detection of the end of speech are equally important: a premature detection of this boundary will have a more dramatic impact in the word naming performance that a delay on its detection.

There are several parameters that may influence the performance of the algorithm. In order to determine the best configuration, a non-exhaustive search has been performed for four of the most important variables. The algorithm, with this baseline configuration, has been assessed with the generation of a newly generated corpus. Results have shown reasonably good results. Particularly, the detection rate for the start of the speech is around 0.85%, which is a quite promising achievement. On the other side, we saw that results obtained for the detection of the end of the speech worsen slightly. An analysis of the detection errors in this case, shows that the general trend in the causes of error is, most of the times, due to a delay in the detection of the end of the speech, which is in fact is good news since this type of error is not expected to affect speech recognition results.

4 Exploiting IR for improved search functionality

The objective of the clinician module is to allow the management of patient data as well as of the collection of exercises and resources associated to them. The management of the exercises data, which now exceeds one thousand of stimuli, only provide a listing functionality, missing the option to search for a given stimuli. Considering the amount of data stored in the system, the lack of a search feature strongly affects the daily usage of the module. Besides, it should be noted that the data constituting the exercises and the stimuli is somehow peculiar in its format. In fact, most of the time, this is represented by a single keyword (i.e.: the title of a document). This means that if the therapist does not remember the exact term he/she is looking for, the search will probably fail. For these reasons, it is important that the search functionality provides extended search capabilities.

4.1 Extended search functionality

In information retrieval, full-text search refers to techniques for searching a match of the query terms throughout the content of each stored document in a
collection. Query expansion is the process of reformulating a query to improve retrieval performance. It involves evaluating the search terms and expanding them with additional, related information to match additional results. Query expansion involves techniques such as finding synonyms of a key term and searching for the synonyms as well or finding the various morphological forms of a key term by stemming and including them into the final results. In this work, a hybrid approach was investigated: additional metadata has been generated through the support of semantic resources and exploited through the query expansion of the search terms. Then a full-text search engine indexes and manages this data to improve matched results.

Additional metadata information has been generated with the support of two thesaurus-based, lexical resources: MWN.PT\textsuperscript{1} (MultiWordnet of Portuguese) and PAPEL\textsuperscript{2} (Palavras Associadas Porto Editora – Linguateca). With their support, the relations of synonymy, hypernymy, and part-of have been extracted. Apache Lucene\textsuperscript{TM} is a search engine that efficiently supports full-text indexing, ranking and searching features. Lucene has been integrated into the VITHEA system, to exploit the full-text search functionalities.

Three indexes have been created: on the answers of the questions exploiting, besides the original answer, its synonyms, hypernym, and part-of relation; on the title of a document, considering the synonym relation; and on the category of a document, exploiting both the part-of relation to build a semantic hierarchy and the information on the title to refine the search.

4.2 Experimental evaluation

The strategy described in the previous Section has been implemented into the VITHEA system and, thus, is available through a web interface which allows both to exploit search features and visualization of the results. This interface has been used to query the system and evaluate the returned results in terms of precision and recall. Free-text searching and query expansion are likely to retrieve many documents that are not relevant to the intended search question. In fact, by expanding a search query to search for the synonyms of the user entered term, more documents are matched as the alternate word are matched as well, increasing the total recall. This comes at the expense of a reduced precision of the result returned. However, in this particular context, where the amount of available data is limited to a well defined domain, we consider that the reduced precision is a tolerable price that can be addressed and even compensated by the value added from the extended retrieval.

Here, these measures have been computed for each of the indexes generated, namely on the question’s answers, on the document title, and on the document category. Results confirmed the expectations, in fact both for the answers of a question and for the title of a document, recall has improved at the expense of precision. In most of the queries, the system either correctly provided an

\textsuperscript{1} http://mwnpt.di.fc.ul.pt/index.html
\textsuperscript{2} http://www.linguateca.pt/PAPEL
extended result set with relevant results, or returned the direct match of the query - in the same way of a relational database - when no relevant data have been found for that query. An interesting example is the search for the key term “alimento” (food). The returned item set is not closely related with the intended meaning of the search term, yet, the result provided is relevant in this context. Also, without the extended search capability this query would not have returned any result, since the searched term does not even exist into the database.

5 New automatic evocation exercises for therapy treatment

There are several naming tasks for assessing the patient’s ability to provide the verbal label of objects, actions, events, attributes, and relationship. Currently, the VITHEA system supports exercises based on visual confrontation, automatic closure naming, and responsive naming. The integration of automatic serial naming or semantic category naming exercises would be of valuable help for patients in recovering from aphasia. Animal Naming is a semantic fluency task that consists of naming as many animals as possible within a one-minute interval. The score of the task corresponds to the sum of all admissible words, where names of extinct, imaginary or magic animals are considered admissible, while inflected forms and repetitions are considered inadmissible. The automatization of this task raises several challenges, namely due to the disfluencies that are present in spontaneous speech. It is expected that hesitations, filled pauses and repetitions will be common in the recorded speech. For this reason, the same keyword spotting based approach adopted in Pompili et al. (2012) has been followed, extending it to address the animal naming task.

5.1 Keyword spotting

Keyword spotting techniques aim at detecting a certain set of words of interest in the continuous audio stream. This method is based on the acoustic matching of speech with keyword models in contrast to a background or garbage model. In this extension, the list of keywords also plays a fundamental role, containing the names of admissible animals that will be accepted by the speech recognition system. The size of this list may have a significant impact on the outcome of the recognizer. In fact, if a keyword is missing in the list, it will never be detected, on the other hand we also expect that a longer list will result in an increase of the perplexity of the keyword model.

5.2 Keyword model generation

To automatically build an adequate keyword model for the animal naming task, an existing lexical resource has been used as a baseline, consisting of an extensive list of animal names. It contains 6044 animal names, grouped, classified and
labelled with its semantic category. In order to take into account that some names in this extended list will be much more likely to be said than others, we tried to compute the likelihood of the different target terms, as it is commonly done in n-gram based language modelling. To this purpose, the total number of results provided by any web search engine for a particular query can be a useful information, meaningful of the term’s popularity. However, the homonymy presented by some terms may lead to an incorrect count, related to alternative meanings of the term. Therefore, a more refined retrieval strategy has been implemented, which took into account the semantic information associated to each key term. The search query, is therefore composed of the bigram: \(<animal\ name<> <category>\), i.e.: beta peixe. In this work the Bing Search API\(^3\) has been used to obtain the count data.

5.3 Experimental Evaluation

Once obtained an initial list that fulfils the desired requirements, several phases of experimental tests were performed in order to determine the best compromise between the length of the list and its content. To this purpose, a corpus includes recordings from 31 healthy adults (16 females and 15 males), native Portuguese, was specifically collected. The evaluation process followed the scientific method, wherein several phases composed from hypothesis formulation, measurements, tests and hypothesis adjustments, have been alternated. These have led to important discoveries, such as an error in the rule-based pronunciations that were used in the lexicon, that overall have improved the outcome of the recognition.

Overall, experiments have conducted to encouraging results, leading to a Word Error Rate (WER) of 14.66\%, thus showing the feasibility of the animal naming task.

6 Exploiting sylable information in word naming recognition of aphasic speech

During preliminary experiments evaluating the performance of the word naming recognition task within the VITHEA system, an analysis of word detections errors have been performed Abad et al. (2013). From these results emerged that some characteristics of aphasic patients that sometimes causes keywords to be missed are both pauses in between syllables or mispronounced phonemes

6.1 Syllabification task

Syllabification is the process consisting of the identification and delineation of syllable boundaries in a word. Contrary to what could be expected, syllabification is not a task of simple resolution and can be addressed from different perspectives, namely by considering the orthographic or the phonetic form of

\(^3\) http://datamarket.azure.com/dataset/bing/search
the word. Here a rule-based approach based on the Maximal Onset Principle for European Portuguese Candeias and Perdigão (2008, 2009) has been used. Rules were driven with a lexicon of almost 400K words, syllabification was performed according with the orthographic form.

To integrate the generated syllables into the version of AUDIMUS customized for the VITHEA system, it is necessary to alter the lexicon used by the recognizer. In practice, for each keyword entry a new alternative phonetic transcription is generated which consists of the same original phonetic string with short pause units inserted in between the syllable boundaries. In this way, for each pronunciation provided by the automatic grapheme-to-phoneme module, an alternative “syllabified” version of the canonical pronunciation is generated.

6.2 Experimental evaluation

The assessment of the performance of the recognition process provided with this alternative syllabified lexicon has been evaluated through automated tests. In particular, in order to measure the achieved results in terms of overall improvements, the same set of experiments carried out during the VITHEA project to evaluate the word naming task have also been replicated here Abad et al. (2013). The corpus used in the evaluation is composed of 16 patients, native Portuguese speakers with different types of aphasia. The corpus has been collected in two different therapy centres in two different sessions, thus these sets are referred as APS-I and APS-II, respectively. Surprisingly, the results obtained for the APS-I corpus have shown that the usage of the new augmented pronunciations does not lead to any significant improvements in terms of overall speech recognition performance. Instead, we note that for some patients the Word Verification Rate (WVR) worsens. On the other hand, the results obtained for the APS-II corpus showed encouraging improvements in term of WVR. A detailed analysis of the original audio transcription, confirmed for some patients the general trend to slow down the rhythm of a word in correspondence with syllable boundaries. This phenomena was sometime associated with the hesitations shown by some patients in uttering a word.

7 Conclusions and Future Work

This work addressed the development of new features for the VITHEA system, a virtual therapist for the recovery from a language disorder named aphasia. The project is currently distributed to almost 160 users among speech therapists and patients. Recently, the project has collected several awards from both the speech and the health-care communities. Also, the work that I have carried out in the context of the VITHEA project and of this thesis have led to the following publications: 1 journal paper, 2 conference papers, 2 demo sessions, 2 presentations in speech therapists convention.
The success of the system motivated the research on additional features which could extend its functionalities and robustness. These extensions have been the objectives of the present work and have concerned with many aspects of the VITHEA platform, from its architecture to one of the main components of the system: the speech recognition engine.

Probably, the main contribution of this thesis has been the development of a mobile version of the client module that has shown the feasibility of this kind of systems on such devices. User experience evaluation provided remarkably good results. A custom approach for enabling a hands-free interface, has been designed and implemented. The administration platform of the project has been provided with an advanced search capability that will enhance the usability of the application. Techniques from Information Retrieval have allowed to obtain high values of recall in the results retrieved from the system. Another important achievement for the project has been the implementation of a new category of exercise. To conclude, an alternative pronunciation lexicon has been exploited in order to improve the robustness of the speech recognizer.

Possible directions for future works include the idea of improving with more functionalities the version of the system dedicated to mobile devices. For what concerns the hands-free interface, a future extension that aims at improving the quality of the speech/non-speech detection, may consider the hypothesis of exploiting the in-house AUDIMUS speech segmentation system for performing this task in the server-side. Another important area of the project which deserves further research is the one that concerns to the current set of available exercises. Besides the future integration of the animal naming task, the exercise itself could be extended to other types of stimulations, such as automatic serial naming or picture description.

Referências


