Automatic word naming recognition for an on-line aphasia treatment system

Alberto Abad\textsuperscript{a,*}, Anna Pompili\textsuperscript{a,b}, Angela Costa\textsuperscript{a,c}, Isabel Trancoso\textsuperscript{a,b}, José Fonseca\textsuperscript{d}, Gabriela Leal\textsuperscript{d}, Luisa Farrajota\textsuperscript{d}, Isabel P. Martins\textsuperscript{d}

\textsuperscript{a}Spoken Language Systems Lab, INESC-ID, Lisbon, Portugal
\textsuperscript{b}Instituto Superior Técnico, Lisbon, Portugal
\textsuperscript{c}Centro de Linguística da Universidade Nova de Lisboa, Portugal
\textsuperscript{d}Language Research Laboratory, Lisbon Faculty of Medicine, Portugal

Abstract

One of the most common effects among aphasia patients is the difficulty to recall names or words. Typically, word retrieval problems can be treated through word naming therapeutic exercises. In fact, the frequency and the intensity of speech therapy are key factors in the recovery of lost communication functionalities. In this sense, speech and language technology can have a relevant contribution to the development of automatic therapy methods. In this work, we present an on-line system designed to behave as a virtual therapist incorporating automatic speech recognition technology that permits aphasia patients to perform word naming training exercises. We focus on the study of the automatic word naming detector module and on its utility for both global evaluation and treatment. For that purpose, a database consisting of word naming therapy sessions of aphasic Portuguese native speakers has been collected. In spite of the different patient characteristics and speech quality conditions of the collected data, encouraging results have been obtained thanks to a calibration method that makes use of the patients word naming ability to automatically adapt to the patients speech particularities.

Keywords: aphasia, word naming, speech disorder, virtual therapy

*Alberto Abad is the corresponding author. 
Tel: +351 213 100 232 Fax: +351 213 145 843 
\textit{Email address:} alberto.abad@12f.inesc-id.pt (Alberto Abad)

Preprint submitted to Special Issue on SLPAT July 13, 2012
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, including auditory comprehension, speech production, oral expression, and reading and writing fluency. There are several causes of brain injuries affecting communication skills, such as brain tumours, brain infections, severe head injuries, and most commonly, cerebral vascular accidents (CVA). The number of individuals that suffered from a CVA has dramatically increased in the last decades, with approximately 600,000 estimated new cases each year in the EU. Typically, a third of these cases present communication deficiencies [1]. Among the effects of aphasia, the difficulty to recall words or names is the most common disorder presented by aphasic individuals. In fact, it has been reported in some cases as the only residual deficit after rehabilitation [2]. Several studies about aphasia have demonstrated the positive effect of speech-language therapy activities for the improvement of social communication abilities [3]. Moreover, it has been shown that the intensity of therapy positively influences speech and language recovery in aphasic patients [4].

The use of computers for aphasia treatment has been explored since the early eighties [5, 6] due to its importance for overcoming resource limitations inherent in the provision of rehabilitation services. Recent commercial products and research laboratory prototypes are currently available for different kinds and degrees of aphasia treatment and in some cases they incorporate speech and language technology (SLT), mainly the use of speech output and virtual agents. For instance, the Aphasia Tutor\(^1\) proposes to the patient a series of training exercises showing a letter, word, picture, or spoken sentence and the patient responds by choosing or typing an answer. SentenceShaper\(^2\) is a program that helps people with aphasia to communicate in their own voices. Lingraphica\(^3\) is a complete communication device that provides communication facilities by using a vocabulary of icons that can be combined to create personalized phrases and simple sentences that are synthesized with a text-to-speech (TTS) system. Sentactics and AphasiaScripts\(^4\) are research

\(^{1}\text{http://www.bungalowsoftware.com/}\)
\(^{2}\text{http://www.sentenceshaper.com/}\)
\(^{3}\text{http://www.aphasia.com/}\)
\(^{4}\text{http://www.bitek.com/virtual-teachers-and-therapists/}\)
laboratory prototypes aimed at improving sentence comprehension and production abilities of aphasic individuals and at training conversational speech, respectively. In both cases, an animated agent or virtual therapist is designed to behave like a clinician and to interact with and provide help to the users. In the particular case of the rehabilitation of word recalling difficulties or anomia, there has been a recent proliferation of studies to examine the most useful way to use computers [7, 8], in some cases based on a commercial system named MossTalk Words\(^5\) [9, 10]. However, the few studies among these works that stimulate oral naming always use the active participation of a speech-language therapist in the evaluation of the answer. Up to our knowledge, there is no therapeutic tool for word recalling training addressed to aphasic patients that makes use of speech recognition technology. In Portuguese language, the only software available related to aphasia therapy is LISLING [11], which is designed for global language stimulation.

Recently, we have presented the first prototype of an on-line platform that incorporates SLT for treatment of Portuguese speakers with aphasia [12]. The proposed system —named VITHEA (Virtual Therapist for Aphasia treatment)— aims at acting as a “virtual therapist”, asking the patient to recall the content represented in a photo or picture shown. By means of the use of automatic speech recognition (ASR) technology, the system processes what is said by the patient and decides if it is correct or wrong. One of the advantages of such an automatic system is that computer delivery of stimulation permits intense and repeated access to therapy exercises, while human agents get bored relatively quickly in delivering such stimulation inputs. Consequently, aphasia patients using VITHEA can access to word naming exercises from their homes at any time, which will hopefully cause an increase in the number of training hours, and consequently a significant improvement in the rehabilitation process. One particular issue with a fully automatic system that may prevent it from becoming a truly effective therapy tool is the responsiveness of the system to the users. In an effort to increase this responsiveness and to introduce stronger “social learning elements” [13], a virtual animated character was integrated. The agent is able to provide synthesized responses to the user in a natural way, which is expected to partially overcome this responsiveness issue and to enhance the users therapy experience, in a similar way to virtual tutors for e-learning

\(^5\)http://www.mosstalkwords.com/
applications [14]. Nonetheless, the major challenge to ensure the utility of VITHEA for therapeutic purposes is the ability to produce consistent and reliable responses to the patients’ oral productions. The presence of feedback is a key aspect in most methods of aphasia treatment. It is essential for the patients to realize when they perform correctly or not the tasks presented in order to develop compensatory strategies that help them to overcome their own difficulties. For this reason, the accurate performance of the automatic word naming recognition module is crucial.

In this work, we focus on the detailed description and assessment of our built-in automatic word naming recognition module and on its role within the VITHEA system. The proposed system validates or rejects patients’ answers based on a keyword spotting approach that makes use of a model of background speech in competition with the expected target keyword model. A corpus consisting of word naming exercises during conventional therapy sessions of aphasia patients has been collected to evaluate the word naming detector. The reliability of the word naming detector for both global evaluation and training purposes is then investigated. Finally, a calibration method to adapt the VITHEA system to the type of speech, acoustic conditions and the recovery stage of each patient is proposed. This document is organized as follows. In section 2 the VITHEA system is introduced. Section 3 is devoted to a detailed description of the automatic recognition approach developed for word naming exercises. The speech corpus of aphasia patients collected for experimental assessment is described in section 4, followed by word naming verification experiments in section 5. Finally, the paper ends with the concluding remarks in section 6.

2. VITHEA: An on-line system for virtual treatment of aphasia

The on-line system described in [12] is the first prototype for aphasia treatment resulting from the collaboration of the Spoken Language Processing Lab of INESC-ID (L2F) and the Language Research Laboratory of the Lisbon Faculty of Medicine (LEL), which has been developed in the context of the activities of the Portuguese national project VITHEA\(^6\). It consists of a web-based platform that permits speech-language therapists to easily create therapy exercises that can be later accessed by aphasia patients using a web-browser. During the training sessions, the role of the therapist is taken by a

\(^6\)http://www.vithea.org
“virtual therapist” that presents the exercises and that 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 visual or auditory stimuli. The patient is then required to respond verbally to these stimuli 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 and processes it by an ASR module, which generates a textual representation. 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. Figure 1 shows a comprehensive view of this process. In practice, the platform is intended not only to serve as an alternative, but most importantly, as a complement to conventional speech-language therapy sessions, permitting intensive and inexpensive therapy to patients, besides providing to the therapists a tool to assess and track the evolution of their patients.

2.1. Therapeutic focus: Recovery of word naming ability

Approaches for aphasia rehabilitation can be typically classified into three groups: disorder-oriented treatment, functional treatment and participation-oriented treatment [15]. In the first case, the aim is to restore linguistic processing by providing linguistic exercises. This group includes the majority of recovery aphasia studies. The functional treatment approach aims at
emphasizing the achievement of communicative activities in everyday life. It may include the use of residual linguistic skills as effectively as possible, and/or use of augmentative and alternative communication (AAC) strategies to compensate the linguistic deficits. Participation-oriented treatment directly targets the domain of social participation. Within this third approach, the emphasis is on living with the consequences of aphasia [16].

This work deals exclusively with disorder-oriented treatment approaches. Particularly, the focus of the VITHEA system is on the recovery of word naming ability for aphasic patients. The improvement of word retrieval ability is of particular importance given the widespread difficulty of recalling words or names among aphasia patients, and there is a considerable number of studies on this topic [17, 18, 19, 20]. Naming ability problems are typically treated with semantic exercises like “naming objects” or “naming common actions”. The approach usually followed is to subject the patient to a set of exercises comprising a set of stimuli in a variety of tasks. The stimuli are chosen based on their semantic content and the patient is asked to name the subject that has been shown. The VITHEA system is designed to provide this type of exercises to Portuguese speaking aphasia patients.

2.2. The patient and the clinician applications

The system 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. The two modules adhere to different requirements that have been defined for the particular class of user for which they have been developed. Nonetheless they share the set of training exercises, that are built by the clinicians and performed by the patients.

2.2.1. Patient application module

The patient module is meant to be used by aphasic individuals to perform the therapeutic exercises. Figure 2 illustrates some screen-shots of the Patient Module.

Exercise protocol Following the common therapeutic approach for treatment of word finding difficulties, a training exercise is composed of several semantic stimuli items. Stimuli may be of several different types (text, audio, image and video) and they are classified according to
themes, in order to immerse the individual in a pragmatic, familiar environment. Like in ordinary speech-language therapy sessions, once the patient is logged into the system, the virtual therapist guides him/her in carrying out the training sessions, providing a list of possible exercises to be performed. When the patient chooses to start a training exercise, the system presents target stimuli one at a time in a random way and he/she is asked to respond to each stimulus verbally. After the evaluation of the patient’s answer by the system, the patient can listen again to his/her previous answer, record an utterance in case of invalid answer or skip to the next exercise.

**Exercise interface** The exercise interface has been designed to cope with the functionalities needed for automatic word recalling therapy exercises, which includes among others the integration of an animated virtual character (the virtual therapist), text-to-speech synthesised voice, image and video displaying, speech recording and play-back functionalities, automatic word naming recognition and exercise validation and feed-back prompting, besides conventional exercise navigation options. Additionally, the exercise interface has also been designed to maximize simplicity and accessibility. First, because most of the users for whom this application is intended suffered a CVA and they may also have some sort of physical disability. Second, because aphasia is a predominant disorder among elderly people, which are more prone to suffer from visual impairments. Thus, we carefully considered the graphic elements chosen, using big icons for representing our interface.

### 2.2.2. Clinician application module

The clinician module is specifically designed to allow clinicians to manage patient data, to regulate the creation of new stimuli and the alteration of the existing ones, and to monitor user performance in terms of frequency of access to the system and user progress. The module is composed of three sub-modules:

**User management** This module allows the management of a knowledge base of patients that can be edited by the therapist at any time. Besides basic information related to the user personal profile, the database also stores for each individual his/her type of aphasia, his/her aphasia
severity (7-level subjective scale) and aphasia quotient (AQ) information from the Western Aphasia Battery.

Exercise editor This module allows the clinician to create, update, preview and delete stimuli from an exercise in an intuitive fashion similar in style to a WYSIWYG editor. In addition to the canonical valid answer, the system accepts for each stimulus an extended word list comprising the most frequent synonyms and diminutives. Besides the management of the set of stimuli, the sub-module also provides an interface to manage the database of multimedia resources used by the exercises.

Patient tracking This module allows the clinician to monitor statistical information related to user-system interactions and to access the utterances produced by the patient during the therapeutic sessions. The statistical information comprises data related to the user’s progress and to the frequency with which users access the system. On the one hand, all the attempts recorded by the patients are stored in order to allow a re-evaluation by clinicians. This data can be used to identify possible weaknesses or errors from the recognition engine. On the other hand, monitoring the usage of the application by the patients will permit the speech-language therapist to assess the effectiveness of the platform and its impact on the patients’ recovery progress.
2.3. Platform architecture overview

An ad hoc multi-tier framework that adheres to the VITHEA requirements has been developed by integrating different heterogeneous technologies. The back-end of the system relies on some of the most advanced open source frameworks for the development of web applications: Apache Tiles, Apache Struts 2, Hibernate and Spring. These frameworks follow the best practice and principles of software engineering, thus guaranteeing the reliability of the system on critical tasks such as databases access, security, session management etc. The back-end side also integrates our in-house speech recognition system (AUDIMUS, [21, 22]), text-to-speech synthesizer (DIXI, [23]) and virtual face animation engine (FACE, [24]). The ASR component is the backbone of the system and it is responsible for the validation or rejection of the answers provided by the user. TTS and face animation technologies allow the virtual therapist to “speak” the text associated with a stimuli and supply positive reinforcement to the user. The client side also exploits Adobe® Flash® technology to support rich multimedia interaction, which includes audio and video stimuli reproduction and recording and play-back of patients’ answers. Finally, a recent feature introduced into the system implements a data architecture that allows handling groups of speech-language therapists and groups of patients. Thus, a user may belong to a specific group of patients and this group can be assigned to a therapist or to a group of therapists. Therapists who belong to the same group share the clinical information of the patients, the set of therapeutics exercises, and also the set of resources used within the various stimuli. In this way patients with the same type and/or degree of severity of aphasia can be clustered together and take advantage of exercises and stimuli that are tailored to their specific disorder, thus improving the benefits resulting from a therapeutic training session.

3. Automatic word naming recognition

The automatic word naming recognition module is the component in charge of receiving the patients’ speech answer and validating the correctness of the utterance for a given therapeutic exercise. Consequently, it is an essential component of the virtual therapy system and it strongly determines the usability of the whole therapeutic platform. It is worth noting that the targeted users in VITHEA are assumed to be patients with word
naming difficulties but without (or very low) articulatory or speech production impairments. Hence, in contrast to other communication disorders such as dysarthria, there is no need for acoustic model re-training or adaptation and general purpose models can be used. In fact, in contrast to the case of aphasia, there are several studies on the use of ASR systems with dysarthric patients [25, 26]. This fact may be partially explained by the totally different nature of both deficits, since dysarthria is a motor speech disorder that is more likely to be characterized in terms of acoustic modeling, while aphasia is a disorder that affects language and it is characterized by the inconstancy of the errors and the unpredictability of their occurrence.

3.1. Definition of word verification task

The targeted task for automatic word naming recognition consists of deciding whether a claimed word \( W \) is uttered in a given speech segment \( S \) or not. We refer to this task as word verification. In the simplest case, a true/false answer is provided, but a verification score might be also generated. Notice that we name it word verification, although we actually refer to term verification, since a keyword may in fact consist of more than one word (e.g. rocking chair).

3.2. Word verification based on keyword spotting

Several approaches exist based on speech recognition technology to tackle the word verification problem. Given that word \( W \) is known, forced alignment with an automatic speech recognition system could be one of the most straightforward possibilities. However, the use of synonyms, diminutives, or any other alternative valid way of answering to the proposed question introduces a high degree of uncertainty in the generation of the reference text. Moreover, we expect that speech from aphasic patients will contain a considerable amount of hesitations, doubts, repetitions, descriptions and other speech disturbing factors that are known to degrade ASR performance [27], and consequently, this will further affect the alignment process. These issues led us to consider the forced alignment approach inconvenient for the word verification task. Alternatively, keyword spotting methods can better deal with unexpected speech effects. The object of keyword spotting is to detect a certain set of words of interest in the continuous audio stream. In fact, word verification can be considered a particular case of keyword spotting (with a single search term) and similar approaches can be used.
Keyword spotting approaches can be broadly classified into two categories [28]: based on large vocabulary continuous speech recognition (LVCSR) or based on acoustic matching of speech with keyword models in contrast to a background model. Methods based on LVCSR search for the target keywords in the recognition results, usually in lattices, confusion networks or n-best hypothesis results since they allow improved performances compared to searching in the 1-best raw output result. The training process of an LVCSR system requires large amounts of audio and text data, which may be a limitation in some cases. Additionally, LVCSR systems make use of fixed large vocabularies (>100K words), but when a specific keyword is not included in the dictionary, it is never detected. Acoustic approaches are very closely related to Isolated Word Recognition (IWR). They basically extend the IWR framework by incorporating an alternative competing model to the list of keywords generally known as background, garbage or filler speech model. A robust background speech model must be able to provide low recognition likelihoods for the keywords and high likelihoods for out-of-vocabulary words in order to minimize false alarms and false rejections when continuous speech recognition is performed. Like in the IWR framework, keyword models can be word-based or phonetic-based (or sub-phonetic). The latter allows simple modification of the target keywords since they are described by their sequence of phonetic units. Preliminary experiments were conducted on a telephone speech corpus to choose the best approach for this task [12]. According to the results obtained, it was considered that acoustic based approaches were more adequate for the type of problem addressed in the on-line therapy system.

3.3. Keyword spotting with AUDIMUS

The in-house ASR engine named AUDIMUS [21, 22], that has been previously used for the development of several ASR applications, has been integrated into the VITHEA system. In order to do so, the baseline ASR system was modified to incorporate a competing background speech model that is estimated without the need for acoustic model re-training.

3.3.1. The baseline speech recognizer

AUDIMUS is a hybrid recognizer that follows the connectionist approach [29]. It combines the temporal modelling capacity of Hidden Markov Models (HMMs) with the pattern discriminative classification of multilayer perceptrons (MLP). A Markov process is used to model the basic temporal nature of the speech
signal, while an artificial neural network is used to estimate posterior phone probabilities given the acoustic data at each frame. The baseline system combines three MLP outputs 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). The AUDIMUS decoder is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition \[30, 31\]. A block diagram of the speech recognition system is shown in Figure 3.

The version of AUDIMUS integrated in VITHEA uses an acoustic model trained with 57 hours of downsampled Broadcast News data and 58 hours of mixed fixed-telephone and mobile-telephone data in European Portuguese \[32\]. 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). For the word naming detection task, an equally-likely 1-gram language model formed by the possible target keywords and a competing background model is used.

3.3.2. Background speech modelling in a HMM/MLP speech recognizer

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. The most common method consists of building a new phoneme classification network that in addition to the conventional phoneme set, also models the posterior probabil-
ity of a background speech unit representing “general speech”. This is usually done by using all the training speech as positive examples for background modelling and requires re-training the acoustic networks. Alternatively, the posterior probability of the background unit can be estimated based on the posterior probabilities of the other phones [33]. We followed the second approach, estimating the posterior probability of a background speech unit 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. The minimum duration for the background speech word is fixed to 250 msec.

3.3.3. Background penalty for word spotting tuning

In order to balance the weight of the background speech competing model with respect to the keyword models in the decoding process, we have introduced a background penalty term $\beta$. This term, that is a power term in the probability domain (multiplicative in the log-likelihood domain), permits adjusting the word naming detection system to penalize the background speech model ($\beta > 1$) or to favour it ($\beta < 1$). In this way, it is possible to make the system more prone towards keyword detections (and possibly false alarms) or towards keyword rejections (and possibly miss detections). The default value for the penalty term is $\beta = 1$.

4. Aphasia Portuguese Speech corpus

A corpus consisting of a series of nomination tests from native Portuguese speakers with different types of aphasia has been collected for the assessment of automatic word naming recognition. This database is named Aphasia Portuguese Speech (APS) corpus.

4.1. Types of aphasia in the APS corpus

It is possible to distinguish two different types of aphasia on the basis of the fluency of the speech produced: fluent and non-fluent aphasia. The speech of someone with fluent aphasia has normal articulation and rhythm, but is deficient in meaning. Typically, there are word-finding problems that affect mostly nouns and picturable action words. Non-fluent aphasic speech is slow and laboured, with short utterance length. The flow of speech is more or less impaired at the levels of speech initiation, the finding and sequencing of articulatory movements, and the production of grammatical sequences. Speech is choppy, interrupted, and awkwardly articulated. While
other criteria may be used for aphasia classification, such as localization or comprehension, the fluency of speech is expected to be the most determinant factor affecting automatic word naming recognition. Hence, we distinguish our patients as belonging to one of these two types of aphasia: fluent (F) and non-fluent (NF). In the case of fluent aphasia, we distinguish the following sub-types: Wernicke’s aphasia (WA), anomic aphasia (AA), conduction aphasia (CA) and transcortical sensory aphasia (TSA). We classify non-fluent aphasia speech in: global aphasia (GA), Broca’s aphasia (BA), mixed transcortical aphasia (MTA) and transcortical motor aphasia (TMA). Additionally, we classify as after-effects (AE) and not belonging to any of these two large groups those patients that do not present word naming alterations, but still present aphasia-related deficits in more complex tasks.

4.2. Nomination test definition

Data collection was performed during regular speech-language therapy sessions. Each of the sessions consisted of naming exercises with pictures of objects presented at intervals of at most 15 seconds. The objects and the presentation order were the same for all patients. The pictures adopted in the nomination exercises were selected from a standardized set of 260 pictures described in [34]. They are black-and-white line drawings executed according to a set of rules that provide consistency of pictorial representation. The pictures have been standardized on four variables of central relevance to memory and cognitive processing: name agreement, image agreement, familiarity, and visual complexity. The tests used a subset of 103 pictures selected as the ones for which a name agreement over 85% was reached in an experiment with 300 healthy subjects, and for which no differences in gender, age and education level were found.

4.3. Recording set-up and data collection

Recordings took place in two different therapy centres in two phases. In both phases, the same common set-up was used for recording, which consists of a lap-top computer and two inexpensive microphones: a built-in headset microphone and a table-top microphone. Inexpensive microphones were preferred to high-quality ones in order to better resemble the actual speech recordings of potential users. With the same intention, background noise conditions were not particularly controlled and different sources of noise are in some cases present in the recordings, such as babble noise or noises produced by the patient, the therapists or other people that were present in the room.
during the sessions. Data originally captured at 44.1 kHz was downsampled to 8 kHz to match the acoustic models sampling frequency. Segmentation and word-level transcriptions were manually produced for each session. From the complete manual transcriptions, only the excerpts that correspond to patient’s answers to naming exercises were extracted and used for evaluation.

**APS-I** The first phase of data collection campaign was carried out in February and March of 2011. It includes speech from 8 aphasia patients recorded in a small office room of wooden walls (see left side of Table 1). The total number of extracted segments for evaluation is 1004. Each one of these extracted segments are assumed to correspond to an individual word naming exercise. The number of exercises differs from 8*103 since some of them were not answered, were repeated or were removed due to overlapped speech. The complete APS-I corpus has a duration of approximately 1 hour and 30 minutes, which corresponds to an average answer duration of 5.4 seconds.

<table>
<thead>
<tr>
<th>Patient</th>
<th>Gender</th>
<th>Age</th>
<th>Type</th>
<th>WNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>m</td>
<td>57</td>
<td>TSA</td>
<td>0.60</td>
</tr>
<tr>
<td>2</td>
<td>f</td>
<td>74</td>
<td>TSA</td>
<td>0.55</td>
</tr>
<tr>
<td>3</td>
<td>m</td>
<td>65</td>
<td>CA</td>
<td>0.33</td>
</tr>
<tr>
<td>4</td>
<td>m</td>
<td>60</td>
<td>CA</td>
<td>0.22</td>
</tr>
<tr>
<td>5</td>
<td>m</td>
<td>78</td>
<td>CA</td>
<td>0.75</td>
</tr>
<tr>
<td>6</td>
<td>f</td>
<td>52</td>
<td>AA</td>
<td>0.84</td>
</tr>
<tr>
<td>7</td>
<td>f</td>
<td>57</td>
<td>BA</td>
<td>0.63</td>
</tr>
<tr>
<td>8</td>
<td>m</td>
<td>52</td>
<td>TSA</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Table 1: APS-I and APS-II patients information including gender, age, type of aphasia and word naming score (WNS) computed in the APS data-set.

**APS-II** The second data collection was carried out during May and June of 2011. The original recorded set consisted of 18 new patients. Unfortunately, the audio quality was severely affected by the presence of electrical noise, due most probably to the failure of one of the components of the two-microphone recording set-up. This degradation effect was only found during the manual transcription phase. In order to partially reduce this effect, a noise removal processing was applied to the recorded data. Nevertheless, the resulting speech quality was still
poor (much worse than APS-I) and only the recordings of the patients with a subjective good speech quality were kept. Finally, the APS-II data set includes speech from 8 aphasia patients recorded in a larger office room (see right side of Table 1). The total number of extracted segments for evaluation is 850. The complete APS-II corpus has a duration of 63 minutes, which corresponds to an average answer duration of 4.4 seconds.

5. Experimental evaluation

The APS corpus is used to evaluate the automatic word naming recognition system described in Section 3. Two metrics are considered throughout this section: the word naming score (WNS) and the word verification rate (WVR). On the one hand, the WNS is computed for every speaker as the number of positive word detections divided by the total number of exercises. The manual WNS is shown in Table 1. The automatic WNS in contrast to the manual WNS can be considered a measure of the goodness of the word detector as a tool for global evaluation of patients’ word naming ability. On the other hand, the WVR is computed for every speaker as the number of coincidences between the manual and automatic result (true acceptances and true rejections) divided by the total number of exercises. Thus, it is a measure of the reliability of the detector as a verification tool for virtual word naming therapy exercises. In these experiments, only the head-set microphone recordings have been considered.

5.1. Baseline word naming results

The baseline configuration of the automatic detector is used in this first set of experiments. The leftmost plot of Figure 4 shows the WNS for each patient of APS-I (1 to 8) computed by a human evaluator (blue bar) and by the automatic detector (red bar). In this case, the Pearson’s correlation coefficient between human and automatic scores is 0.904. Moreover, the average absolute difference between the human and automatic scores is 0.074. This means that both scoring methods provide similar figures for each patient, besides providing highly correlated scores. The leftmost side of Table 2 shows the WVR for every speaker of APS-I and the average WVR. These results can be generally considered quite promising. The automatic word detector for this data set is able to provide significant global evaluation results,
but also high word verification rates which would permit performing reliable word naming therapy exercises.

On the other hand, the rightmost plot of Figure 4 shows manual and automatic WNS for each patient of APS-II (9 to 16). In contrast to the results obtained with the APS-I patients, there is a strong degradation of the automatic scores with this data set that results in an average absolute difference between human and automatic scores of 0.314. Furthermore, the WVR shown in the rightmost side of Table 2 for every speaker is quite low and the average WVR is reduced to 0.663 (in contrast to APS-I 0.804). Nevertheless, high correlations between manual and automatic WNS can still be observed, and the Pearson’s correlation coefficient is 0.972. These results suggest that the automatic system still provides significant global word naming ability scores for comparing patients of APS-II: patients with higher automatic WNS have in fact a better word naming ability. However, the electrical noise present in APS-II speech has resulted in a generalized higher miss detection rate. In other words, the baseline detector is miscalibrated to the problems present in APS-II data, resulting in a considerable drop of correct word naming detections.

Figure 4: On the left side, average word naming scores of the human and automatic evaluations for the APS-I corpus. On the right side, average word naming scores of the human and automatic evaluations for the APS-II corpus.

5.2. Background penalty experiments

The goal of this section is to validate the use of the background penalty term ($\beta$) described in section 3.3.3 as a method to adjust the word detector
<table>
<thead>
<tr>
<th>APS-I</th>
<th></th>
<th>APS-II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>WVR</td>
<td>Patient</td>
</tr>
<tr>
<td>1</td>
<td>0.78</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>0.80</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>0.73</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>0.84</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>0.73</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>0.91</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>0.70</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>0.91</td>
<td>16</td>
</tr>
<tr>
<td>avg</td>
<td>0.80</td>
<td>avg</td>
</tr>
</tbody>
</table>

Table 2: Word verification rate (WVR) for the patients of APS-I and APS-II data sets and average WVR, using the baseline automatic word verification system.

to the speech production characteristics of each patient and to the particularities of the recorded audio. The leftmost plot of Figure 5 shows the WVR for the APS-I patients for different $\beta$ values. Different trends can be observed for different speakers. In some cases, $\beta$ values lower than the baseline ($\beta = 1$) allow better verification rates, while the opposite is observed for other speakers. The average WVR with the optimum penalty term for each patient (the one that gives a higher WVR) improves to 0.841. These upper-bound detection rates are shown in Table 3. Similarly, this background penalty dependent performance may also be observed in the APS-II data in the rightmost plot of Figure 5. In this case, the lower WVR obtained as a result of the different characteristics of the speech data can be compensated with the appropriate selection of the penalty term. In the optimum case, it is possible to achieve an average word detection performance of 0.819, as it is shown in Table 3.

These results confirm that it is possible to find an operation point of the automatic word detector adapted to the type of speech and to the acoustic characteristics of the data in order to obtain high word verification performances. The remaining problem is how to select this background penalty for a certain set of word naming exercises of a particular patient. As a first possible solution, we hypothesized that patients suffering the same type of aphasia (see Table 1) might be correlated in terms of word naming detection and show similar trends. Thus, it would be possible to select the penalty term $\beta$ depending on the type of disease. However, attending to the results shown in Figure 5 this hypothesis was dismissed. Another hypothesized solution is the possibility of adjusting the word detector according to a parameter that
characterizes the word naming ability of each patient. This parameter could be the manual WNS. In fact, it is a characteristic measured by therapists that can be used as a design parameter for the word detector. Thus, if the $\beta$ term used for each patient is selected in order to minimize the absolute difference between the manual and the automatic WNS, the obtained average WVR increases to 0.824 and 0.803 for the APS-I and APS-II data sets respectively.

5.3. Calibration of automatic word naming verification

The results of the previous section suggest that an automatic computation of the $\beta$ parameter that provides an automatic WNS closer to the manual value might be used as a calibration step. In order to validate this hypothesis, the data from every speaker was randomly split into two halves, in a cross-validation experiment. The first half is used to search for the best $\beta$ parameter on that data sub-set. Then, this $\beta$ penalty term is used to process the second half of the data and the WVR is computed on this second sub-set. Note, however, that if we consider a single partition, a bias may be introduced. To have a more robust measure of the proposed method, we define 10 random partitions and compute the mean WVR performance on them. Mean word verification rates and standard deviation for each patient are shown in Table 4.
Table 3: Word verification rate (WVR) for the patients of APS-I and APS-II data sets and average WVR, using ideal background penalty terms for each patient.

<table>
<thead>
<tr>
<th>APS-I</th>
<th>APS-II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>WVR</td>
</tr>
<tr>
<td>1</td>
<td>0.84</td>
</tr>
<tr>
<td>2</td>
<td>0.84</td>
</tr>
<tr>
<td>3</td>
<td>0.83</td>
</tr>
<tr>
<td>4</td>
<td>0.86</td>
</tr>
<tr>
<td>5</td>
<td>0.75</td>
</tr>
<tr>
<td>6</td>
<td>0.94</td>
</tr>
<tr>
<td>7</td>
<td>0.72</td>
</tr>
<tr>
<td>8</td>
<td>0.94</td>
</tr>
<tr>
<td>Avg</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Table 4: Word verification rate (WVR) for the patients of APS-I and APS-II data sets and average WVR, using automatically calibrated background penalty terms.

<table>
<thead>
<tr>
<th>APS-I</th>
<th>APS-II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>WVR</td>
</tr>
<tr>
<td>1</td>
<td>0.84 (±0.036)</td>
</tr>
<tr>
<td>2</td>
<td>0.84 (±0.020)</td>
</tr>
<tr>
<td>3</td>
<td>0.83 (±0.038)</td>
</tr>
<tr>
<td>4</td>
<td>0.82 (±0.024)</td>
</tr>
<tr>
<td>5</td>
<td>0.70 (±0.024)</td>
</tr>
<tr>
<td>6</td>
<td>0.93 (±0.030)</td>
</tr>
<tr>
<td>7</td>
<td>0.69 (±0.063)</td>
</tr>
<tr>
<td>8</td>
<td>0.92 (±0.032)</td>
</tr>
<tr>
<td>Avg</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Regarding the baseline results of Table 2, one can observe a slightly better average performance in the APS-I corpus using the proposed calibration method, but more importantly, a large improvement is achieved with the APS-II patients. This is an important result for the VITHEA system. If the patients’ word naming ability can be characterized by therapists based on similar exercises to the ones proposed by the virtual therapist, the system can use this characterization to adapt the word detector to any particular patient after a certain number of word naming exercises. Moreover, the calibration method permits adapting to different acoustic environment conditions, such as the ones presented by APS-I and APS-II data. In fact, this quite simple
method can be used to adapt regularly to the word naming ability progress of each patient measured by speech-language therapists, which is expected to evolve as a result of the therapy sessions.

6. Conclusions

Automatic word naming recognition for evaluation and treatment of speakers with aphasia has been investigated in this work. The proposed word recognition module –based on acoustic keyword spotting approaches– is the essential component of an on-line treatment system aimed at acting as a virtual therapist for word naming ability training. Using a new collected database of word naming exercises performed by native Portuguese speakers with aphasia, we have shown that it is possible to achieve highly correlated global word naming scores and high performance word verification rates even for different types of patients and acoustic conditions. This was achieved thanks to a calibration step that makes use of the similarity between manual and automatic word naming scores to tune the penalty term of the keyword spotting system.

In general, we consider the achieved results very promising. Nevertheless, an open remaining question is whether the accuracy achieved is sufficient for a satisfactory user experience and if it positively contributes to the recovery of real users. In this sense, in the near future we plan to evaluate both the patients and therapists satisfaction by means of opinion questionnaires and also to assess the utility of the whole therapy platform in terms of word naming recovery progress of patients. Moreover, the possibility to incorporate both supervised and unsupervised model adaptation strategies to improve the word recognition module will be explored. Also, as part of our future plans, more sophisticated speech tools may be also integrated, such as goodness of pronunciation [35] tools. This would allow a different type of assessment of the pronunciation errors, which may provide useful feedback for the therapist and the patients. It will also potentially facilitate the adaptation of VITHEA to the treatment of other speech and language disorders.

Acknowledgments

This work was supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, under project RIPD/ADA/109646/2009.
References


