VITHEA: On-line word naming therapy in Portuguese for aphasic patients exploiting automatic speech recognition

Anna Pompili, Pedro Fialho and Alberto Abad

L2F - Spoken Language Systems Lab, INESC-ID Lisboa, Portugal
\{anna.pompili, pedro.fialho, alberto.abad\}@l2f.inesc-id.pt
http://www.l2f.inesc-id.pt/

Abstract. Aphasia is an acquired communication disorder that affects speech and language functionalities at varying degrees. The recovery of lost communication functionalities is possible through frequent and intense speech therapy sessions. The aim of the VITHEA -Virtual Therapist for Aphasia Treatment- project is to exploit speech and language technology (SLT) to facilitate the recovery process of Portuguese aphasic patients, more concretely to recover lost word naming abilities. The proposed system has been designed to behave as a virtual therapist that simulates ordinary speech therapy sessions and, by means of automatic speech recognition (ASR) technology, validates patients performance. In addition to ASR, the system integrates several technological components including animated virtual characters and text-to-speech synthesis (TTS). At the IberSPEECH2012 demo session, the current prototype consisting of two web-based applications will be demonstrated.

Keywords: speech disorder, aphasia, speech recognition, keyword spotting, word naming, virtual therapy, on-line recovery

1 Introduction

Aphasia is a communication disorder caused by the damage of one or more language areas of the brain affecting various speech and language functionalities, including hearing comprehension, speech production, and reading and writing fluency. 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 [1]. Several studies about aphasia have demonstrated the positive effect of speech-language therapy activities for the improvement of social communication abilities [2]. Moreover, it has been shown that the frequency and intensity of these therapy sessions positively influences speech and language recovery in aphasic

* This work was partially funded by the Portuguese Foundation for Science and Technology (FCT), through the project VITHEA (RIPD/ADA/109646/2009) and the project PEst-OE/EEI/LA0021/2011.
patients [3]. Typically, word retrieval problems can be treated through semantic exercises like “naming objects” or “naming common actions”. The approach usually followed is to subject the patient to a set of stimuli in a variety of tasks.

VITHEA is the first prototype of an on-line platform that incorporates SLT for treatment of Portuguese speakers with aphasia [4]. The system 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. Thus, the platform makes available word naming exercises to aphasia patients from their homes at any time, which will allow an increase in the number of training hours, and hopefully a significant improvement in the rehabilitation process.

2 VITHEA: The proposed system

VITHEA is a web-based platform that permits speech therapists to easily create speech therapy exercises that can be later accessed by aphasia patients using a web-browser. During the training sessions, the role of the speech therapist is taken by a “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 encoding the patient’s answer and passes it to the ASR system which generates a textual representation of it. 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.

2.1 Automatic Speech Recognition

The ASR module is the backbone of the system: it is responsible for receiving the patient’s speech answer and validating the correctness of the utterance for a given therapeutic exercise. Consequently, it strongly determines the usability of the whole therapeutic platform. 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. Keyword spotting is an adequate solution to deal with unexpected speech effects typical in aphasic patients, such as hesitations, doubts, repetitions, descriptions and other speech disturbing factors. In the current version of the system, acoustic based keyword spotting is applied for word verification. In order to do so, our in-house ASR engine named AUDIMUS [5], that has been previously used for the development of several ASR applications, was modified to incorporate a competing background speech model that is estimated without the need for acoustic model re-training similarly to [6].
**The baseline speech recognizer** AUDIMUS is a hybrid recognizer that follows the connectionist approach [7]. 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 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 [8]. 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 softmax 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.

**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. In this work, the posterior probability of the background unit is estimated based on the posterior probabilities of the other phones. In practice, instead of estimating the posterior, we compute the likelihood of the background speech unit as the mean likelihood of the top-6 most likely classes the phonetic network at each time frame. In this way, there is no need for acoustic network re-training. Further details on this approach and on the complete word verification module can be found in [9].
2.2 Virtual character animation and speech synthesis

The virtual therapist’s representation to the user is achieved through a tri-dimensional (3D) game environment with speech synthesis capabilities. Being a web based platform, the game environment is essentially dedicated to graphical computations (performed locally in the user’s computer), while synthesized speech generation occurs in a remote server, thus assuring proper hardware performance. The game environment is based on the Unity\(^1\) game engine and contains a low poly 3D model of a cartoon character with visemes and facial emotions (defined as morph targets), which receives and forwards text (dynamically generated according to the system’s flow) to the TTS server. Upon server reply, the character’s lips are synchronized with synthesized speech.

**Speech synthesis** The system integrates our in-house TTS engine named DIXI\(^{[10]}\), configured for unit selection synthesis with an open domain cluster voice for European Portuguese. The server uses DIXI to gather SAMPA phonemes\(^{[11]}\), their timings and raw audio signal information, which is lossy encoded for usage in the client game. The phoneme timings are essential for a visual output of the synthesized speech, since the difference between consecutive phoneme timings determines the amount of time a viseme should be animated.

**Animation** Lip synchronization is based on a custom phoneme to viseme mapping, specific to the visemes available in the 3D model. Due to its simplicity (and cartoon nature), this model includes visemes only for \(e\), \(a\) and \(o\) sounds, therefore it’s not suitable for lip reading. Apart from lips, independent idle animations for eyes and head/body were developed, for a more engaging experience.

2.3 The patient and clinician modules

The system comprises two specific web applications: the patient and the clinician modules. They are 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 patient module** The patient exercise interface has been designed to cope with the functionalities needed for automatic word recalling therapy exercises, which includes besides the integration of an animated virtual character (the virtual therapist), text-to-speech synthesized voice \(^{[10]}\), image and video displaying, speech recording and playback functionalities, automatic word naming recognition and exercise validation and feedback prompting. 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

\(^{1}\) http://unity3d.com/
chosen, using big icons for representing our interface. Figure 2 illustrates some screen-shots of the patient module.

The clinician module The clinician module is specifically designed to facilitate the management of both patients information and therapy exercises by the therapist. The module is composed of three sub-modules: User management, Exercise editor, and Patient tracking. These 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.

![Screen-shots of the VITHEA patient application.](image)

**Fig. 2.** Screen-shots of the VITHEA patient application.

### 3 Demo details

During the demo session, the patient and clinician modules will be shown. Since both modules are remote web applications, one or more computers will be provided, although any computer with a built-in or external microphone, Internet connection and a web-browser can be used to access the system. Portuguese speaking attendees will be able to directly test the system: for the patient module they will be able to experience word naming therapy exercises and for the clinician module they will be able to easily create and edit new exercises.

2 [http://utente.vithea.org](http://utente.vithea.org)

3 [http://terapeuta.vithea.org](http://terapeuta.vithea.org)
References