

PROJECT APHATAR : AN EXPERIMENT AND IMPLEMENTATION

A Thesis
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
MASTER OF SCIENCE in Computer Science

by
Gregory Teodoro
August 2013

Thesis Approvals:

Dr. Justin Y. Shi, Thesis Advisor, Department of Computer Science

© Copyright 2013

Gregory Teodoro. All Rights Reserved.

ABSTRACT

Aphasia is an acquired communication disorder that affects the ability of a person to speak and understand spoken language. The purpose of the Aphatar project is to create a virtual clinician that will help suffers of aphasia improve their speech in common scenarios. The project will gauge the interaction and quality of this virtual clinician against those of a real clinician. Aphatar will be created using three major systems: (1) KINECT for audio and visual recording, audio input, and future work in reading the client's emotional state using the KINECT 3D Camera system, (2) The Olympus Speech Recognition System, provided by Carnegie-Mellon University which will accept the audio input of the user and translate it from speech to text then provide spoken feedback to the user, and (3) the Avatar display system, which will provide the graphical interface for the former, allowing the user to see the avatar and interact with it.

ACKNOWLEDGEMENTS

There are many people who have helped make this possible. I would like to first thank my parents and girlfriend for helping support me throughout my academic career, and Temple University for offering me this opportunity.

I also would like to thank Dr. Justin Yaun Shi, for without his guidance I would not be where I am today. I want to thank Dr. Emily Keshner for use of the VEPO lab for testing, and offering me the summer opportunity at the lab that set me down this path. I would like to thank Dr. Nadine Martin for her work in helping create and make the Aphatar Project what it is today, as well as affording me the opportunity to help build it and become part of the team.

I would first like to thank the NIH for helping fund this project. Finally, I would like to thank all of the team members who have helped me along the way. Without their work, analysis, and expertise in the fields of speech therapy and neuroscience, this project would not have been possible.

TABLE OF CONTENTS

ABSTRACT.....	iii
ACKNOWLEDGEMENTS.....	iv
LIST OF FIGURES.....	viii
LIST OF TABLES.....	ix
CHAPTER	
1. INTRODUCTION	
1.1. Motivation.....	1
1.2. Problem Statement.....	2
1.3. Methodology.....	3
1.4. Expected Results.....	4
2. APHASIA AND THERAPY	
2.1. Aphasia – A Language Disorder.....	5
2.1.1. Types of Acute Aphasia.....	7
2.2. Aphasia Therapy.....	9
3. SURVEY OF COMPUTER-ASSISTED THERAPY	
3.1. AphasiaScripts Test.....	10
3.2. Sentactics Test.....	11
3.3. Aphatar Project.....	13
4. THE APHATAR PROJECT	

4.1. Overview.....	14
4.2. The Microsoft Kinect.....	16
4.2.1. Depth Imaging.....	17
4.2.2. Skeletal Tracking.....	20
4.2.3. Face Recognition and Expression Tracking.....	23
4.3. Facegen Software Suite.....	25
4.4. Olympus System and Speech Recognition.....	26
4.5. Speech Recognition Component.....	27
4.6. Speech Synthesis Component.....	29
4.6.1. Concatenative Synthesis.....	30
4.6.2. Diphone Synthesis.....	31
4.6.3. Domain-Specific Synthesis.....	31
4.6.4. Formant Synthesis.....	31
4.7. The Flite System.....	32
4.8. Microsoft Speech API.....	35
5. METHODOLOGY	
5.1. Virtual Clinician.....	37
5.2. Aphatar 2.0.....	39
5.3. Olympus Integration and API.....	40
6. EXPERIMENTATIONS	
6.1. Pilot Study.....	42
6.2. Wizard of Oz Study.....	43

6.2.1. Data Gathering Methodology.....46

6.2.2. Results.....48

6.3 Future Study – Autonomous Virtual Clinician Study.....50

7. CONCLUSION AND FUTURE WORK.....51

REFERENCES.....54

LIST OF FIGURES

Figure 1: Areas of Brain affected by aphasia.....	5
Figure 2 : Kinect Hardware Diagram.....	16
Figure 3: The Axis System as used by the Microsoft Kinect.....	18
Figure 4: Overview of the Kinect Pose and Skeletal Tracking system.....	23
Figure 5: A Visual Representation of the Olympus Architecture and System.....	28
Figure 6: A sample representation of the minimized CART tree for Flite.....	34
Figure 7: Results of the pilot study.....	43
Figure 8: The face used for the Wizard of Oz project.....	45
Figure 9: Sample questionnaire for reception of virtual clinician.....	46

LIST OF TABLES

Table 1: Summary of the characteristics of the different forms of acute aphasia.....	7
Table 2: List of MS SAPI visimes and corresponding sound.....	38
Table 3: Results of the first Wizard of Oz study.....	51

CHAPTER 1

INTRODUCTION

1.1 Motivation

Aphasia is a communication disorder that affects the ability to speak and understand spoken languages. Aphasia can occur at all ages, but is most common in the elderly, and thanks to improved health care, more people are living longer with aphasia than in the past. (National Aphasia Association (<http://www.aphasia.org/>)). In-clinic treatments done with a human clinician are limited by time and scope, although they are effective at improving the speech and understanding of the client over the period of the treatment[1-4]. As more people with aphasia reenter the community in response to these treatments, it becomes more important to incorporate more residual language skill development for the use of functional communication[5-7]. This process can begin during in-clinic treatment, but for the best results, aphasia sufferers need to have continual practice in everyday situations.

This form of treatment is usually done through the use of clinician and client role-playing[8-9]. These scenarios are formed through the use of scripts prepared ahead of time by the clinician and client.

The motivation behind this project and study is to create an automated virtual clinician that can allow the aphasia suffers to have continual practice everyday at home, so that therapy can continue and progress even without a human therapist present.

1.2 Problem Statement

Existing aphasia home therapy systems are those based on the repetition of problem words or specific unchanging scripts, such as the AphasiaScripts and Syntactics systems. Though these systems have their merit, they do not accurately attempt to recreate actual scenarios that a patient may encounter in their everyday lives. The Aphatar system separates itself from these projects by attempting to create a real-time, dynamic dialogue between the client and the virtual therapist.

By having a real-time, dynamic dialogue, we can create everyday scenarios that the client will run into, such as ordering food at a diner, or visiting a doctor. The dialogues will also be built to have a large degree of variety of accepted inputs, and potential outputs. A client may start up a diner scenario, and choose between breakfast, lunch, or dinner, create their own orders, and even receive recommendations from the virtual clinician. The end result is, we hope, will be a more immersive experience that is far more similar to an actual dialogue with a person, rather than practicing the same script.

1.3 Methodology

In the context of this project, the role-playing scenarios and script will be performed by the virtual clinician when interacting with the client. Though in the beginning the virtual clinician is driven by a clinician behind the scenes, later on in the project this will be handled through the use of the speech recognition system developed by Carnegie-Mellon University called Olympus.

These early tests, dubbed the “Wizard of Oz” test, are named due to the nature of having a driver behind the virtual clinician. This test is set up to evaluate the efficacy and social validity of the functional communication treatment for aphasia and show that it is comparable to the results that would be obtained with a real clinician.

The first specific aim of this study is to develop or aid in the development of software that is capable of recognizing and identifying the language and speech patterns that are common in aphasia sufferers. This Spoken Dialog System (SDS) will interact autonomously with the client and can be configured behind the scenes with scripts to provide personalized treatment and scenarios. Secondly, client and virtual clinician interactions must be tested through the use of the “Wizard of Oz” test so that we can identify the conversational variations and key words we'll need during implementation of the end product. The third aim will be to compare the quantity and quality of the language production in dialogues between the virtual clinician and client, and of those between a real clinician and client.

The end result of this project is to determine whether or not the use of a virtual clinician is as equally successful in obtaining the therapeutic goals as it would be with a real clinician.

1.4 Expected Results

We have gone into this project with high hopes. Past projects such as AphasiaScripts and Syntactics have shown that virtual therapy provides measurable results when coupled with human therapy and practice. Given the precedence for such projects, we predict that the use of a virtual clinician for role-playing and script therapy will deliver similar results to practicing the same therapy with just a human clinician.

To show this, we will take the measured CIU (Correct Information Unit) scores of the subject before beginning virtual/human therapy, and after virtual/human therapy is completed. We believe that we will find higher CIU scores after therapy, that would be similar to that of patients that worked with only a human therapist.

CHAPTER 2

APHASIA AND THERAPY

2.1 Aphasia – A Language Disorder

Aphasia is a disturbance of the comprehension and formulation of language, which is caused by dysfunction in specific regions of the brain. It is a language disorder that is marked by sufferers having difficulty remembering words, stuttering, inability to pronounce or form words, inability to name objects, excessive use of utterances, paraphasia, and other such symptoms. Aphasia is often linked to brain damage, commonly caused by a stroke[43,44].

Aphasia is usually caused from lesions to the language areas of the frontal, temporal, and parietal lobes of the brain. These areas are located in the left hemisphere of the brain, these lesions can be caused by stroke or brain trauma. Aphasia may also develop through a brain tumor or progressive neurological disease, or hemorrhaging of the brain though these are out of the scope of the Aphatar project.

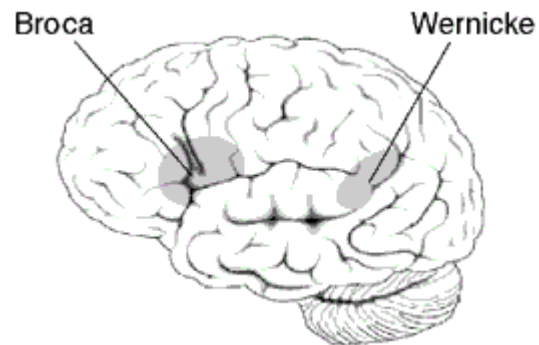


Figure 1: Areas of the brain affected by Broca's (expressive) and Wernicke's (receptive) aphasia. (Source : Wikipedia)

Suffers of acute aphasia can recover some or most of their skills through work with a Speech-Language Pathologist. Treatment for aphasia varies, in part due to the potential range and severity of symptoms, and most treatments are custom-tailored to the individual. Most treatment relies heavily on repetition to address language performance, working on very specific task-based skills. There are two models of treatment, the substitute skill model which uses an aid to help with spoken language, and the direct treatment model, which targets deficits with specific exercises.

There are several treatment techniques; Visual Communication Therapy (VIC), Visual Action Therapy (VAT), Functional Communication Treatment (FCT), and Promoting Aphasic's Communicative Effectiveness (PACE). The Aphatar project can be classified under the FCT treatment, which focuses on improving activities to specific functional tasks, social interaction, and self expression.

The use of computer technology to aid aphasia treatment is relatively new. Computers hold a number of advantages over traditional therapy, often being more intense and achieving better results. Most programs consist of a number of exercises that are done at home to augment face-to-face therapy with a clinician. Computer technology however is limited in communicative settings, due to the limit of a computer system to imitate normal speech and keep up and recognize an aphasic's speech patterns[45]. Our hope is that the Aphatar project will improve speech recovery.

2.1.1 Types of Acute Aphasia

Six of the acute aphasia types are considered major, each with their own characteristics. Some tests take a general approach to treatment, while others target a specific type of aphasia.

Table 1: Summary of the characteristics of the different forms of acute aphasia. (Source: Wikipedia)

<u>Type of Aphasia</u>	<u>Repetition</u>	<u>Naming</u>	<u>Auditory Comprehension</u>	<u>Fluency</u>
Receptive Aphasia	mild – moderate	mild – severe	defective	fluent paraphasic
Transcortical sensor aphasia	good	moderate – severe	poor	fluent
Conduction aphasia	poor	poor	Relatively good	fluent
Anomic aphasia	mild	moderate – severe	mild	fluent
Expressive aphasia	moderate – severe	moderate – severe	mild difficulties	Non-fluent, effortful, slow
Transcortical motor aphasia	good	mild – severe	mild	non-fluent
Global aphasia	poor	poor	poor	non-fluent
Mixed transcortical aphasia	moderate	poor	poor	non-fluent

Receptive aphasia is marked by the sufferer speaking in excessively long sentences. Often these sentences may have no meaning, or add a number of unnecessary or made up words. They often have poor auditory and reading comprehension, but speak and write fluently. Receptive aphasics often have trouble understanding others, and their own speech and may be completely unaware of mistakes they are making. Transcortical sensory aphasia is similar to receptive aphasia, but retain much better repetition skills.

Conduction aphasics have deficits in connections between the speech-comprehension and speech-production areas of the brain. They retain high auditory comprehension and remain fluent with their speech with occasional errors. Their repetition ability, however, is poor.

Anomic aphasia is marked by a difficulty in recalling words or names. Sufferers have trouble naming certain words, often linked by grammatical type or semantic category. They are capable of producing grammatical, but empty speech. Auditory comprehension remains near normal range. Anomic aphasia is linked to Alzheimer's disease.

Expressive aphasia shows the ability to speak short, meaningful phrases with great effort. Expressive aphasics may omit small words such as “is”, “the”, or “I” which can inhibit understanding by others. Expressive aphasics are able to understand the speech of others and are aware of their speech problems.

Transcortical motor aphasia is similar to expressive aphasia, but the repetition ability is intact. Auditory comprehension is often high, but can degenerate quickly as conversation complexity goes up. It is often associated with right hemiparesis, or paralysis of the patient's right face and arm.

Global aphasia is the most extreme form of aphasia, and is characterized by severe communication difficulties and extreme limitation in the ability to speak or comprehend language. In some cases they may become completely non-verbal and communicate only through gestures. It too is associated with right hemiparesis.

Mixed transcortical aphasia is similar to global aphasia but retains repetition ability.

2.2 Aphasia Therapy

There is no singular treatment for all forms of aphasia due to the nature of the disorder and the large variety of symptoms in sufferers. Instead most treatments are tailored to the individual. Aphasia therapy often involves the use of a multidisciplinary team consisting of a doctor, physiotherapist, occupational therapist, speech-language pathologist, and social workers. Most forms of treatment of aphasia rely on heavy repetition in attempts to address the language performance by working on very specific language skills.

There are a number of different forms of therapy. Visual Communication Therapy uses index cards with symbols to represent and work with various components of speech. Visual Action Therapy (VAT) trains individuals to map gestures to objects. Functional Communication Treatment (FCT) is the form of therapy we focus on in which therapy focuses on improving specific tasks, social interactions, and self-expression.

One such method of FCT is the use of role playing scenarios. The goal of these role playing scenarios is to place the client into a scenario in which they will need to either perform a task (order food for dinner), or deal with basic daily social interactions (such as waiting at a bus stop).

CHAPTER 3

SURVEY OF COMPUTER-ASSISTED THERAPY

3.1 AphasiaScripts Test

There have been some instances of computerized treatment of aphasia done in the past, one of which is the use of AphasiaScripts in a test done by Leora Cherney and team[46]. In this test, the program AphasiaScripts is used. The therapy takes place as follows; first the client listens to the entire script. After that, each sentence that is the client's role in the conversation is practiced by them. After practice is done, the client and virtual clinician take turns reading through their roles in the script, simulating an actual conversation.

The scripts themselves are put into the system and prerecorded by a normal speaker. For the sake of consistency of data and ensuring that clients were following the given directions, the script and therapy sessions were closely monitored.

The results were measured for content, grammatical productivity, and rate of speech. Content dealt with the number and percent of the client's speech that was script-related. Grammatical productivity was measured by the total number of script-related morphemes, nouns, verbs, and modifiers. The client's rate was defined as the number of script-related words produced in a given time. Client's were then scored based on these values. There were three participants in total.

The first participant produced more complete sentences, and was able to do

so at a faster rate, the second and third participants had reduced the amount of empty speech and utterances.

The end results of this study showed a positive change overall in the content, grammatical productivity and the rate of script production after training. The positive changes are suggestive of the fact that results will be had whether there is a human or virtual clinician. These results show that the difference between human and virtual clinician will be minimized and not have an influence on the end results or efficacy of therapy. It is our hope to further improve on this study as well, by doing away with fully pre-scripted scenarios, and allowing variations and role playing to occur more freely. It should be noted that the AphasiaScripts system makes no use of speech recognition or speech synthesis.

3.2 Sentactics Test

The Sentactics Test was another such test performed to study the effects of the Sentactics therapy program in acquisition, general production, and comprehension of complex sentences by sufferers of aphasia.

In this test, they used twelve participants with agrammatic expressive aphasia. Twenty-four active sentences were developed using a set of 24 transitive verbs, they are semantically reversible, and contain two animate nouns. No noun or verb exceeded two syllables. Every developed sentence had two black and white line drawings to go along with it, one depicted the target sentence, and the other

depicted the semantically reversed sentence. For each of these sentences, an object relative structure was developed to comprise the sentences in treatment. All target object relative sentences, and any prompts and feedback used in the training was prerecorded by a normal speaker.

A pretreatment test was taken by each participant. The computer would randomly generate a picture pair, the system would produce a sentence based on the picture, and then prompt the participant to produce a similar sentence to describe the one on the right. The differences between the two pictures were highlighted for the user, one picture may have a man, and another a woman.

Following this, treatment started and was done through trials. In each trial, the computer would select a target sentence to practice. It would first present the sentence-picture matching task, then follow it with the sentence production task. Computer feedback was given for the participant responses. The computer would then walk the participant through the TUF training protocol for additional practicing on comprehending and producing the target sentence. These trials would continue for a maximum of 20 days, or until the participant achieved an 80% correct performance for 4 consecutive days. After this was post-treatment testing, which was identical to pretreatment testing.

Human clinician trials followed the same structure as the computer clinician trials. Half the participants worked with the virtual clinician, and the other half worked with the human clinician as a control.

The results of the test showed an improvement in all participants' ability to

produce target object relative sentences and scores were vastly improved from the pretreatment test to the post-treatment test. Comprehension also improved from a mean value of 50% to 82%.

Overall, the results between the human clinician and the virtual clinician do not differ significantly. It can be assumed that a virtual clinician is as much a viable option for TUF treatment as a human clinician. This lends credence to the Aphatar project as it shows that computerized aphasia treatment is equally effective to human based treatment.

3.3 Aphatar Project

The Aphatar project seeks to improve upon both of these virtual therapy applications by using role-playing therapy and conversational variance. AphasiaScripts and Sentactics rely on the use of preset sentences and scripts that are practiced and repeated with no variance. The Aphatar project will improve upon these past works by emphasizing the natural flow and dynamic nature of a conversation through the use of variable role-playing scenarios.

This will be done through the creation of a dialogue system with Olympus, in which the patient will be able to navigate through the dialogue at their own ease, answer questions freely, and will not be bound to a set script, rather they will be placed into a specific scenario with a goal and be allowed to make up their own responses to the questions asked.

CHAPTER 4

THE APHATAR PROJECT

4.1 Overview

The Aphatar project is a multifaceted project involving three separate goals that will be combined into a final product, each goal having its aims and its own method to gather research data. The end goal of which is to fully evaluate client-virtual clinician interactions in terms of quality and quantity of the spoken communications when compared to the quality and quantity between a client and human clinician. Thus we are to develop a virtual clinician and human interaction system, the Aphatar, that can be used independently by people with aphasia to practice and improve functional communication skills through the use of role playing activities. We will then use the data collected by this to evaluate the efficacy of the treatment for aphasia.

For this project there are three aims. Firstly, we had to develop software that could recognize and identify language and speech patterns. A Spoken Dialog System will be used so that the client can interact with the system autonomously, and a dialog tree system that can be configured to provide proper responses and personalized treatment and scenarios modeled on role playing and scenario scripts. The acoustics, language, and lexical models will need to be customized to deal with speech issues present in aphasic clients, and the Spoken Dialog System will need to be integrated with a virtual clinician virtual avatar.

Secondly, the aim will be to test the above product in a controlled environment. This pilot test between the client and virtual clinician will involve the animated virtual avatar, and a human clinician who will be behind the scenes driving it. This has been dubbed the “Wizard of Oz” study, due to having a 'wizard' working behind the scenes controlling the avatar. The purpose of this study will be to identify the conversation variations that may come up in a given scenario and any key words that we will need to add into the above Spoken Dialog System. This will also help measure client adaption and response to the virtual avatar and some early data for gauging efficacy of the therapy.

Thirdly, the final aim will be to compare the quantity and quality of the language production in communication between the human clinician and client, and the virtual clinician and client. The purpose here is to determine if the use of a virtual clinician is equally as successful for reaching the therapeutic goals when compared to a human clinician. This will be done through the use of video recordings to evaluate the language output in both virtual clinician and human clinician tests. Recordings will be analyzed for qualitative and quantitative measures of content, syntax, and morphology. A Likert scale questionnaire will also be given to the clients to gain feedback on the client's reception to interacting with an avatar when compared to a human clinician.

The long term goal of this will be to provide a functional rehabilitation program through an easy to use software suite that will allow therapy to be done at home, while remaining economically feasible and accessible to a broad spectrum of

the population.

4.2 The Microsoft Kinect

The Microsoft Kinect is a motion sensing device originally developed for the Xbox 360 gaming console. It's original intent was to allow the user to control and interact with the Xbox 360 without the use of a game controller. This was done through the use of gestures and spoken commands. The Kinect itself consists of a RGB Camera for reading in a 2D color image, 3D depth sensors for the detection of 3D imagery, and a microphone array capable of reading in audio and pinpointing the source, and a motorized tilt mechanic to move the Kinect[10].

The Kinect hosts a number of advantages over other gesture based systems as it does not require the user to wear any accessory to track movement, all required hardware is housed inside the Kinect itself. (Figure 1)

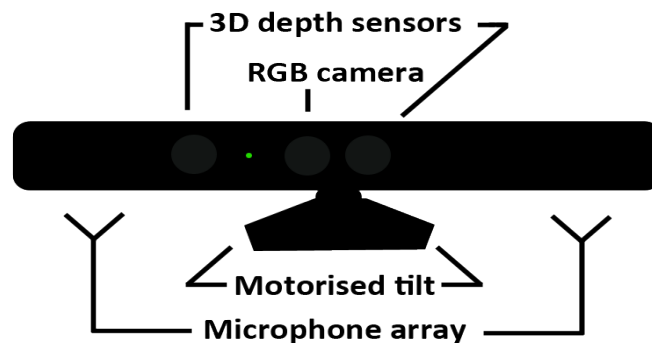


Figure 2 : Kinect Hardware Diagram

The RGB Camera receives the 2D color image that is used in face recognition and recording. The microphone array is located along of the bottom of

the Kinect device and consists of four separate microphones. This enables the Kinect to listen to and recognize speech as well as pinpoint the location of the sound and who is speaking. The microphones have built in ambient noise suppression and echo cancellation to further increase audio quality.

The 3D depth sensors are separated into two parts, an infrared projector and a monochrome CMOS sensor[11-12]. The infrared projector creates a speckled pattern of infrared light on the Kinect's viewing area, this pattern is then reflected and read back into the Kinect. The Kinect translates the deformation of this pattern into a detailed 3D depth image of the scene. This depth image is a gray scale image of the Kinect's view, and contains the Z-Axis data of the scene. This 3D depth image is then used with the RGB image to preform skeleton tracking[11].

4.2.1 Depth Imaging

Depth Imaging in the Kinect is done through the use of an infrared laser emitter, an infrared camera, and an RGB camera. The laser emitter emits a single beam that is split into multiple beams through the use of a diffraction grating to create a pattern of speckles onto the scene. The speckled pattern is captured by the infrared camera and is compared to a reference pattern, which is stored in the memory of the sensor. An object that is closer or farther from the Kinect will have the speckle pattern appear larger or smaller to the infrared camera. This results in a series of shifts to restore the speckle pattern, and yields a disparity image. For each pixel (640x480) the distance to the sensor can be retrieved from the corresponding

disparity figures.

The method to obtain this disparity figure is discussed by K. Khoshelham of University of Twente[41]. The distance of an object at point k , relative to the sensor's reference plane, and the disparity that can be measure by this is d . The Kinect's depth coordinate system has origin at the perspective center of the infrared camera. The axes of the Kinect are organized as follows.

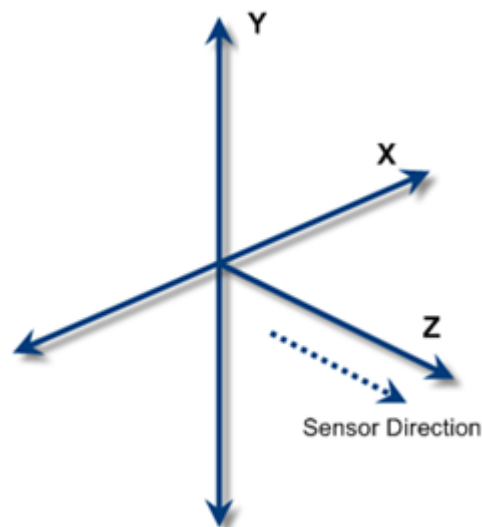


Figure 3: The Axis System as used by the Microsoft Kinect[10]

The assumption is made that any given object is on the reference plane at a distance $Z(o)$ from the sensor, and there is a speckle pattern projected onto that object from the infrared camera. If the object is moved either closer or farther away from the Kinect, the speckle pattern projected on the image will be displaced in the X direction. This is measured as disparity d corresponding to point k . Using the similarity of triangles we have the following :

$$\frac{D}{b} = \frac{Z(o) - Z(k)}{Z(o)}$$

and also :

$$\frac{d}{f} = \frac{D}{Z(k)}$$

Where $Z(k)$ represents the depth of point k of the object, b is the base length, and f is the focal point of the infrared camera. D is the displacement of k , and d is the disparity. Through the use of substitution, we take D from the second equation and place it into the first equation and express $Z(k)$ in terms of the other variables.

$$Z(k) = \frac{Z(o)}{1 + \frac{Z(o)}{fb}d}$$

The above equation is the mathematical model for derivation of depth from the disparity, with $Z(o)$, f , and b provided by the calibration of the Kinect and infrared camera. The Z coordinate of a point, when taken with f , can be used to define the image scale for that specific point. The object coordinates of each point can then finally be calculated from its coordinates in the image, and the scale using the following.

$$X(k) = \frac{-Z(k)}{f} (x(k) - x(o) + \delta x)$$

$$Y(k) = \frac{-Z(k)}{f} (y(k) - y(o) + \delta y)$$

$X(k)$ and $Y(k)$ represent the image coordinates of the point, $x(o)$ and $y(o)$ are the coordinates of the principle point, and the deltas of x and y are corrections for any distortion in the lens.

Once we have the depth points, we can project each 3d point from the point cloud onto the RGB image, giving us a integration of the depth and color maps of the Kinect. This integration of depth map and RGB imaging is used heavily by the Kinect in Skeletal and Facial tracking.

4.2.2 Skeletal Tracking

In the Kinect's skeletal tracking system, a human body is represented by a number of joints and body parts, such as the head, neck, shoulders, and arms. Each joint is represented by a 3D Coordinate. These joints must all be detected and determined correctly in real time to allowed interactivity with a limited computation and time, since gaming performance is the top goal for the Kinect. This problem was found and solved by Jamie Shotton and the Kinect Skeletal Tracking team[42].

The team created a large database of motion captured human actions, based on what they believed people would make in a given entertainment scenario. This database consisted of around 500,000 frames of data, in a hundred different sequences of actions, from driving, dancing, kicking, running, menu navigation, and other such actions that may be taken in a game scenario. In any given pose, it

is only necessary to record the limbs directly involved in the pose, for example; waving does not need the lower limbs to be tested or recorded.

Temporal information is also unused, instead the use of static poses is focused on. Since motion-capture has very small amounts of change from one pose to another that they can be considered insignificant, 'further neighbor' cluster is used to prune the similar frames from the database. Thus no more than a subset of 100,000 poses are used, and no two poses are closer than 5 centimeters in Euclidean distance over the used joints.

This allowed the formation of the intermediate body part representation. In this, the team defined several localized body part labels that cover areas of the body. These parts are defined to directly localized particular skeletal joints, while others are used to fill in gaps or aid in prediction of other joints. There are 31 body parts that are used in total, ranging from head to the lower feet. Left and right parts are considered distinct to aid disambiguation.

Features must be pulled from the depth map, so a simple depth comparison function is used. At any given pixel of x , the features are computed using the following :

$$F_{\theta}(I, x) = d_I\left(x + \frac{u}{d_I(x)}\right) - d_I\left(x + \frac{v}{d_I(x)}\right)$$

Where $d_i(x)$ is the depth at a given pixel x in the image I . The parameter $\Theta=(u,v)$ describes the offsets u and v . Normalization of the offsets makes the features depth invariant. This also guarantees that the features are 3D translation

invariant.

These features provide only a little information about which part of the body the pixel belongs to, so they are ran through a decision forest to further clarify which body part they belong to. The design of the tree is such that each pixel has a limited amount of features that need to be read, and operations done on them. These trees are trained on a different set of synthesized images, about 2000 example pixels from each image is chosen and distributed across all body parts.

When the information from the Kinect is ran through these trees, the information obtained about what pixel belongs to which body part needs to be pooled to generate accurate proposals for the positions of the skeletal joints. This is done through the use of a local mode-finding approach based on mean shift, through the use of a weighted Gaussian kernel.

The density estimator per body part is as follows.

$$f_c(x) \propto \sum_{i=1}^N w_{ic} \exp\left(-\left\|\frac{x-x_i}{b_c}\right\|^2\right)$$

x represents the coordinate in 3D world space, and N is the number of image pixels. w_{ic} is pixel weighting and x_i is reprojection of the image pixel into world space given the depth. b_c is learned per-part bandwidth. Pixel weighting considers both the inferred body part probability at the pixel and world surface area of a given pixel.

The density estimates are made depth invariant. Given the context of the

application, certain parts can be merged, for example, four parts which cover the head can be merged into a single head joint.

All pixels above a learned probability threshold are used as starting points for part c . After all of this a confidence estimate is given as a sum of all pixel weights reaching each mode. Detected modes are on the surface of the body, and need to be pushed back by z offset to produce a final joint position proposal.

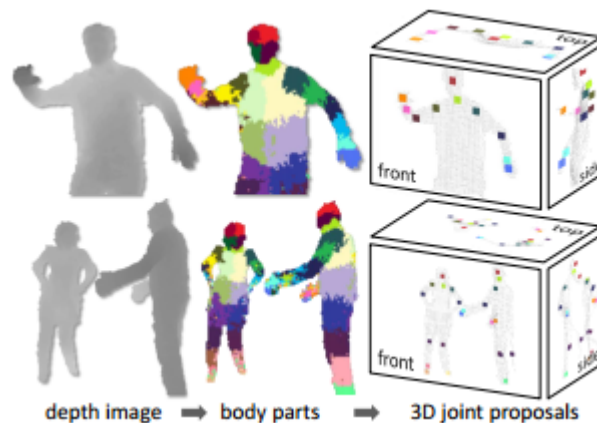


Figure 4: Overview of the Kinect Pose and Skeletal Tracking system[42]

4.2.3 Face Recognition and Expression Tracking

A method of facial-expression recognition has been developed by Wenjun Zeng and Zhengyou Zhang for use with the Microsoft Kinect[39]. The Kinect sensor produces both a 2D color stream and a 3D depth stream at 30 fps. However, it has been observed that the data of a close-up facial region is very noisy. A method of regularized maximum-likelihood deformable model fitting (DMF) algorithm for 3D face tracking was developed to solve this issue[40].

The DMF Algorithm is currently used in Microsoft Avatar Kinect to allow a user to control their avatar's facial expression and head through facial-expression tracking and arm movements through the skeletal tracking system of the Kinect. Though originally for use in Teleimmersive Conferencing, these algorithms can be employed in the Aphantar system to gauge the client's emotional and comfort state on run-time and modify the script as required.

The algorithm works as follows. First multiple neutral face frames are used in model initialization. For each initialization frame, face detection and alignment is done on the image. The alignment algorithm provides 83 main feature points on the face, which are assumed to be consistent across all given frames. These features are put into four categories: first of which is corners of main features, such as eyes or mouths. These points have a clear correspondences to the face model and are consistent across all faces.

Second are points on the eyebrows and upper and lower lips. These points are highly deformable and may not directly correspond to 2D feature points in the alignment section. The third category contains points that surround the face, referred to as silhouette points, once again there is little correspondence here with the alignment results. Lastly the algorithm marks off white points which are unused.

Tracking is done through the use of point-to-point and point-to-plane comparison. The algorithm relies on the feature points detected and tracked earlier. Feature points are detected in the image of the previous frame through the use of

the Harris corner detector. These points are tracked to the current frame by matching areas around the points with cross correlation.

The end result is a tracked face, containing important potential features for reading emotion, such as the location of the eyebrows, and shape of the lips.

4.3 Facegen Software Suite

The Facegen Software Suite consists of a 3D rendering program designed specifically for the creation and animation of faces. The system supports over 50 different morph points for the 3D face rendering, and a large number of options, additional models, and texture packs.

Facegen offers two advantages. First, Facegen allows the user to create a face and modify it based on a number of different sliders for age, sex, and race. Using these sliders, we can custom-tailor a avatar for a particular user. It is our belief that clients are more at ease when dealing with a avatar that is of a similar race and gender to their own.

Secondly, Facegen offers a wide selection of morph points and poses to place the face in. It supports mixing and matching of these morph points as well, allowing us to set the structure and emotional state of a given avatar. The morph points also have a number of preset mouth movements built-in, based on the basic phonemes. We can set the face to appear as though it is saying “Ah” or “Oh”. Using these different positions, it is possible to string together a series of mouth

positions to simulate lip syncing with the speech put out by the synthesizer.

4.4 Olympus System and Speech Recognition

Olympus is an open source framework for the development and research in conversational interfacing[14]. Olympus was designed to be a free alternative to otherwise expensive and time-consuming systems that may hamper long-term research. Olympus is founded on a number of design principles to help further this goal. The system itself is open and all source code is available to the public, and every component of Olympus is documented and detailed. Work went into the framework to allow easy integration into a wide-range of applications, and functions of the program are written and encapsulated within easy-to-use interfaces to promote re-usability.

Olympus is separated into a number of different components, which are connected together in a pipeline architecture. An audio signal from a user is captured through a microphone device, then passed through a speech recognition module that creates what is called a “recognition hypothesis”. This is sent to a language understanding component that will extract keywords from the hypothesis and assign these a hypothesis score. A dialog manager integrates this input into the current context, and then signals the system what the next action is that should be taken. A language generation module is then used to create a string which can be passed to a speech synthesis module, such as Microsoft SAPI, and creates an audio feedback for the user.

Olympus uses the Sphinx decoding engine to recognize speech. Sphinx is a continuous-speech, speaker-independent recognition system. A number of versions of Sphinx have been created and made available. Olympus itself uses the Sphinx-2 and Sphinx-3 features, passing speech data to both for accuracy. Sphinx-2 was developed by Xuedong Huang[15] at Carnegie-Mellon and is released as open source. Sphinx uses a number of different speech recognition techniques.

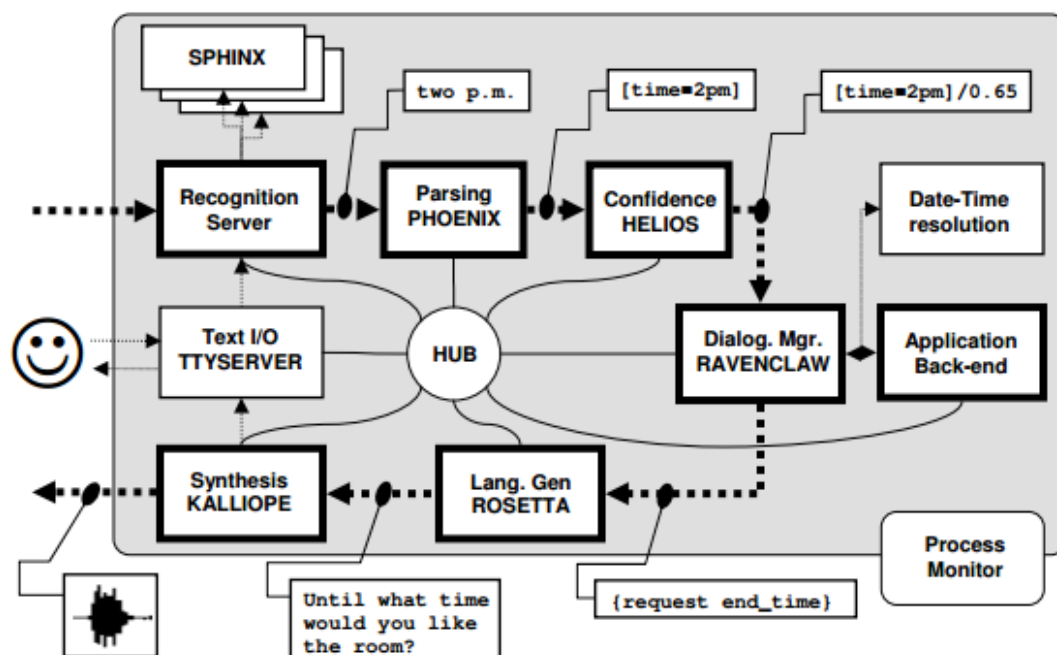


Figure 5: A Visual Representation of the Olympus Architecture and System[14].

4.5 Speech Recognition Component

Speech Recognition is the translation of spoken words into text, it is commonly used in interfaces that require human to computer interaction, but do not allow the user a convenient physical controller such as on a telephone call. It can

also be used for accessibility purposes, allowing someone who cannot use their hands to operate or type into a system such as a computer.

Speech is a stream of audio in which attempts can be made to define similar classes of sounds, known as phones. Words are built out of phones, though the acoustic properties of a waveform corresponding to a phone can vary greatly due to context, different speakers, and speaker accents. Transition between words are known as diphones, or the parts of phones between two consecutive phones. In a Hidden Markov Model system, such as the one used by Olympus' Sphinx[14], there are three specific states that are used in recognition. The preceding phone, the ending phone, and the stable phone in between the two.

As the Aphatar project makes use of highly contextualized environments, we can make use of senones, a phone which has dependence on context, to increase accuracy.

Creation of words is reliant on these phones, as phones build syllables. Syllables are more stable in natural speech environments, and are used to build words either morphologically or phonetically. There are also fillers, and utterances. Utterances are usually separate from the sentences and are detected and ignored by speech recognition systems, though in the Aphatar project and for use of therapy, utterances must be detected and recorded to gauge therapy efficiency. Some examples of utterances are filler words such as “Uhm”, “Uh”, or heavy breathing.

The Olympus system operates on speech recognition by taking the waveform of an audio stream, and then splitting it by utterances and silences.

These bits of audio are analyzed looking for features. These features calculated from the audio stream are divided into frames. Each frame is of a length of 10 milliseconds, and thirty nine features that represent the speech. This becomes the feature vector. These vectors are matched up to a model to correctly attempt to assume what the speech was.

4.6 Speech Synthesis Component

Speech synthesis is the artificial production of human speech by a computer system. The Aphatar project uses two potential text-to-speech systems, which convert normal language text into speech[1]. Synthesized speech is created by joining pieces of prerecorded speech that is stored in a database. Depending upon the design, a database may contain a series of phones and diphones, which are combined to create a word, or it may contain entire spoken words. Quality of the speech synthesizer is dependent on naturalness and intelligibility. Naturalness is a measurement of how closely the output sound mimics human speech, while intelligibility measures how easily understood the speech is by another human.

Text-to-speech systems are comprised of a back-end and a front-end[18]. The front-end handles the text, and will convert any numbers or symbols into written words; '1' will be 'one' for example. This is done in a process called text normalization. The front-end will then attribute phonetic transcriptions to each word, and divide the text into prosodical units; such as individual phrases, clauses, and sentences. The back-end of the system will take this new input and convert it

into sound using the phone/diphone database.

There are two methods for generating the speech waveforms in a text-to-speech system, concatenative synthesis and formant synthesis.

4.6.1 Concatenative Synthesis

Concatenative Synthesis is based on the joining, or concatenation, of prerecorded speech. There are three main types of concatenative synthesis in use in modern text-to-speech applications.

Unit selection synthesis involves the use of a sizable database of recorded speech. Utterances are recorded and then segmented into their individual phones, diphones, half-phones, syllables, words, phrases, and sentences. This process of dividing the utterances is done through the use of a speech recognizer[19]. An index of these segments is created based on the segmentation type and the acoustic parameters; pitch, duration, position in syllable, and neighboring phones. In text-to-speech conversion, the system attempts to find the best utterance in this database by determining the best chain of candidate units. This process can be handled by use of a weighted decision tree[20-21].

Unit selection generally provides incredibly good results due to the low amount of digital signal processing, manipulation of the audio output, done to the synthesized speech. A drawback to such a system however is the size of the database, which can contain upwards of hours of prerecorded speech.

4.6.2. Diphone Synthesis

Diphone synthesis makes use of a minimal speech database that contains all the diphones that occur in a given language. Only one example of each diphone is contained inside the database, resulting in it's minimal approach. When text-to-speech conversion occurs, these diphones are combined and then digital signal processing is performed on the resulting concatenation. Because of the large amount of post-processing done, diphone synthesis suffers from poor naturalness and audio glitching[22].

4.6.3 Domain-Specific Synthesis

Domain-specific synthesis concatenates prerecorded words and phrases to create utterances. It's general use is in applications where the given amount of text the system will be expected to synthesis will be limited to a particular domain, such as train scheduling, automated phone service systems, or weather reports[23]. The limited domain approach of domain-specific synthesis means it can not be used for general-purpose systems.

4.6.4 Formant Synthesis

Formant synthesis differs from concatenative synthesis as no human speech samples are used. Synthesized speech output is instead created using additive

synthesis and an acoustic model[34]. Additive synthesis is a sound synthesis technique that creates timbre by adding sine waves together[32]. In formant synthesis, audio parameters such as the fundamental frequency, voicing, and noise levels are varied to create a waveform simulating speech. This is also known as rules-based synthesis.

An unfortunate drawback to Formant synthesis is that it generates artificial, robotic-sounding speech, in exchange the speech it makes is highly intelligible at a variety of speeds and avoids acoustic glitching that may occur in concatenative systems. Another advantage is that since Formant synthesis systems do not keep a database of speech, they are small and easily portable, working well in highly limited systems.

Formative speech was mainly used in many highly specialized, limited devices and toys of the day, including the Speak-and-Spell, and early arcade games such as Gauntlet, Star Wars, and Astro Blaster[36]. These systems had very artificial sounding voices, but also had limited data storage and a need to vocalize information to the user that must be understood.

4.7 The Flite System

Olympus uses the Flite synthesis library for its use of text-to-speech. Flite's core library consists of a core architecture of fundamental objects[37]. These are referred to as the CST, or C Speech Tools.

For actual synthesis the library has three parts to form a complete synthesizer. The language model provides a phoneset, tokenization rules, text analysis, prosodic structures, and other language components that can be shared between voices. The lexicon is a pronunciation model that contains letter to sound rules for words that are not included in the system's preexisting vocabulary. The lexicon is dependent upon both domain and unit inventory of the given language. The voice itself consists of the unit inventory, and speaker-specific prosody models. A voice depends on the primitives provided by the language model.

The language and lexicon models are language dependent and can be shared across multiple voices of the same language. Voices in Flite are converted from preexisting FestVox voices.

Lexicons are used in Flite to give pronunciations for words, though it is incomplete, mandating the use of a system to derive pronunciations for words not in the lexicon. These are done in the form of a simple sorted list of characters and phones. Phones are encoded with single bytes and letters left as ASCII. For dealing with homophones, the lexicon distinguishes them through the use of an extra character to denote a part of speech. Each word in the lexicon is converted to a list of letters and put into a sorted table, which is indexed into a list of phones.

Letter-to-sound rules are built using CART, classification and regression tree, techniques. These generated trees are minimized.

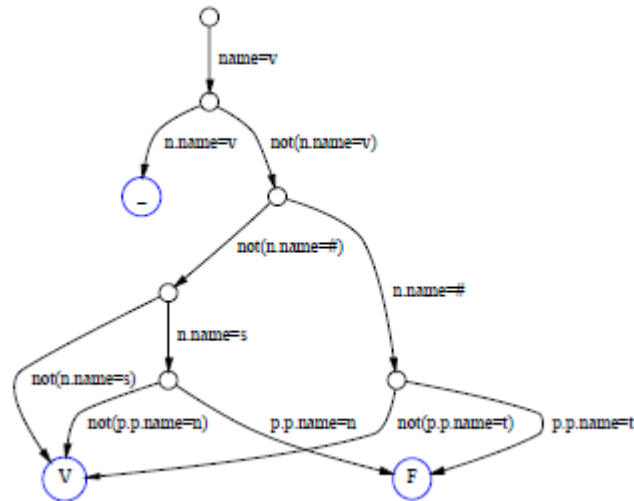


Figure 6: A sample representation of the minimized CART tree for FLite[37].

Following this is the unit database, and is the second largest structure in the Flite system. The unit database is a collection of the speech units that are to be concatenated during speech synthesis. The Flite system uses a residual excited pitch synchronous LPC method to modify the speech segment's pitch and duration[38].

Flite was originally chosen for both, the Aphatar project and for the Olympus system due to it's low weight, and fast speed. The decreased system requirements and run-time load both help with the portability and speed of the project.

4.8 Microsoft Speech API

Microsoft SAPI is the final system used by the avatar for this project. It offers a number of advantages over Flite. It contains a much larger amount of speech units in comparison to Flite, offering more realistic speech. It also allows users to create their own speech synthesizer voice sets, with different accents and genders. The largest advantage that Microsoft SAPI offers over Flite however, is the visime system.

MS SAPI was programmed with built in support for aiding with lip-syncing through the use of the visime system. In the MS SAPI system, the given text string to be synthesized is broken up into the individual phones, and then given a visime. The visime represents the sound being made, and has a beginning and an end point which can be read in real-time using the OnVisime event.

MS SAPI supports two sets of visimes. There is a set of 13 visimes called the Disney 13; these are used in hand-animation and were created by the company Disney. Due to the nature of hand animation it is possible to reuse a large number of visimes and still remain unnoticeable to the viewer. There is also a set of 22 visimes which are for use in a more realistic scenario. For this project, due to the computer-generated nature of the avatar, and the desire for smooth lip-syncing, we have chosen to use the set of 22 visimes.

Table 2: The list of MS SAPI visimes and corresponding sound as per the MS SAPI API.

<u>Visime Number</u>	<u>Sound Simulated</u>
SVP_0 = 0	silence
SVP_1 = 1	ae, ax, ah
SVP_2 = 2	aa
SVP_3 = 3	ao
SVP_4 = 4	ey, eh, uh
SVP_5 = 5	er
SVP_6 = 6	y, iy, ih, ix
SVP_7 = 7	w, uw
SVP_8 = 8	ow
SVP_9 = 9	aw
SVP_10 = 10	oy
SVP_11 = 11	ay
SVP_12 = 12	h
SVP_13 = 13	r
SVP_14 = 14	l
SVP_15 = 15	s, z
SVP_16 = 16	sh, ch, jh, zh
SVP_17 = 17	th, dh
SVP_18 = 18	f, v
SVP_19 = 19	d, t, n
SVP_20 = 20	k, g, ng
SVP_21 = 21	p, b, m

CHAPTER 5

METHODOLOGY

5.1 Virtual Clinician

The virtual clinician consists of a simple program that integrates into the Olympus framework. The program consists of an Avatar Control Panel and an Avatar Display. This Avatar Control panel allows quick access to the speech synthesis API, provided by Microsoft SAPI, and allows direct control over avatar movements, costuming, and scripts.

The face used in the virtual clinician program consists of four pieces. The base of the face, which is the entire face sans the eyes and mouth to serve as a background. This base also contains the art assets for the costuming that the avatar can wear. There are at the moment four distinct costumes for the avatar, a doctor's outfit, a business suit, a t-shirt, and a waitress outfit. Each of these outfits corresponds to the various scenarios.

The other pieces are the left eye, right eye, and mouth pieces. The left eye and right eye consists of a series of frames of animation, ending in a perceived emotion. The emotions the avatar is capable of displaying with the eyes are neutrality, disappointment, happiness, surprise, and curiosity. The mouth piece consists of a series of frames of animated speech, ending in either a smile, a frown, or a neutral position.

The Avatar Display Panel itself is transparent, and fits over a series of

backgrounds that can be changed suit the scenario. So if the scenario is that of a restaurant, the avatar will wear the waitress costume, and be placed over a background of a restaurant. This is to help immerse the patient into the scenario.

The Avatar Control Panel is a user-friendly GUI designed for ease in controlling the avatar. It consists of a empty text box the the “wizard” can type text into, and then a vocalize button. When the vocalize button is selected, the speech in the text box is passed to a Microsoft SAPI, and is spoken by the computer system, on top of this, a thread is created that animates the mouth movements of the avatar in time with the speech synthesis created by Microsoft SAPI.

Also included is a large “Quik-Speak” box. Pre-written scripts can be placed into a text file, and read into the Avatar Control Panel. When this happens, the Quik-Speak box is rendered, containing each line of the script. The wizard can select a specific line, and instantly have it spoken by the avatar. Since the scripts are written ahead of time, it is possible to get a majority of the potential responses the avatar will need to reply with, such as “What would you like to order?”, “Where would you like to sit?”, or “Here is your check”. The purpose of this is to minimize the amount of ad-libbing that must be done by the wizard, and minimize the time spent between the client speaking to the avatar, and the avatar's response.

5.2 *Aphatar 2.0*

Since this study began, the avatar has undergone a drastic change to make improvements and add compatibility to the Olympus API. The Aphatar 2.0 system has been designed to run single window, and was made with portability in mind. The backgrounds are now built into the program window itself, and the avatar costumes and expressions are no longer separated into pieces.

The changes between the Virtual Clinician, and that of Aphatar 2.0 are as follows: Aphatar 2.0 makes use of MS SAPI's visime system, visimes are a series of shapes and mouth movements that are used to animate lip-syncing. The new avatar system consists of 23 frames of mouth movements, each syncing up to a specific sound or syllable. When speech is entered to be synthesized, it is passed to MS SAPI, and MS SAPI separates it into individual segments, and passes these back to the Aphatar 2.0 program as they are spoken. The Aphatar 2.0 then selects the correct frame of animation from the list of 23 frames, and displays it. The results of this back and forth is realistic real-time lip-syncing done by the avatar.

The Aphatar 2.0 also has a networking integration mode. An option on the GUI allows the user to open up a network connection. This connection will search the user-given port and IP and connect to it. If successful, it will wait for incoming string data and pass that data to MS SAPI for animation and speech synthesis. This networked functionality allows it to integrate directly with the Olympus system. When the Olympus system receives speech input, and is ready to pass a response back as output, the formed sentence string is sent over the network to any listening

application. If the connection is open and successful, Olympus will pass that formed sentence string to the Aphatar 2.0 program.

5.3 Olympus Integration and API

The core of this project is the Olympus/RavenClaw architecture[14, 30]. Olympus is an integration of a number of components that are necessary for the building and deployment of an advanced spoken dialog system, including speech decoding, text-to-speech, parsing, dialog management, and spoken language generation. This system will be modified to have the capability to coordinate and control the behavior of the virtual clinician, such as trigger speech motions when speech is synthesized and changing the avatar's emotional state. This capability is provided by the domain reasoner.

The scripts and role playing scenarios that will be used in this test are used to create corresponding tasks in the dialog system. Here, “task” is defined as a series of scripted exchanges that will take place between the client and the Aphatar system. These scripts are written in advance to target a specific utterance or deficiency that the client is to practice, and rules for how to interpret the client's speech and create a proper response.

The “Autonomous Virtual Clinician” study and early testing implementations focus on determining if the client can correctly and fluently reproduce the targeted utterance or deficiency. Heuristics will identify incorrect word choices, excessive pausing, transposition, and other deficiencies that are a result of aphasia and speech disorders. In a situation where the system identifies an

issue or cannot understand the client due to the above issues, the system will ask the client to repeat what they have said.

Olympus allows integration into any application through the use of a network. Olympus will send the receiving input data, and the formed output data through the network to specified ports. A majority of the Olympus system (Sphinx, Kalliope, TTY Server, etc.) connect to each other through standardized ports using localhost. The Aphatar 2.0 system can also integrate itself by connecting to Olympus' Rosetta or Kalliope systems and receiving the output from these systems, which are the finished proposed sentence strings in response to the speech input.

CHAPTER 6

EXPERIMENTATIONS

6.1 Pilot Study

The pilot study contains data that has been collected from two participants (48 and 58 yrs old) with aphasia. Both participants were at least one year post-onset and displayed only a moderate level of aphasia affecting speech and language production but not comprehension. Each participant reported to the study twice a week, for two weeks for a session lasting one hour. Four scripts were used in each session, two based on common errands and two based on social interactions. Each script was practiced twice per session. Participant one practiced with the human clinician and participant two practiced with the virtual clinician. In the second week, four new scripts were practiced, and participants were paired with the other clinician. The outcome measure was the proportion of total narrative words used in sentences when conversing with a clinician.

No significant differences emerged from this data, but there was a trend in which the participant who had practiced with the human clinician before the virtual clinician produced a greater percentage of narrative words when paired with the virtual clinician (65% to 68%, or 3% difference). The participant who started with the virtual clinician and ended with the human clinician produced more narrative words with the human clinician (82% to 90%, or a 8% difference). Future studies will control for order of virtual and human clinician interactions with the client.

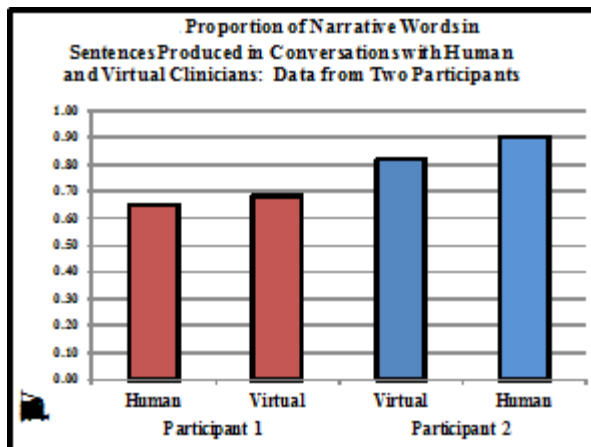


Figure 7: The results of the pilot study, showing the percentage of narrative words produced in conversations of both participants, and the clinician with whom they conversed.

This data is consistent with other studies that suggest people will interact with virtual humans without detriment to the language performance. This lends credence to the hypotheses that training individuals with aphasia with a virtual clinician will provoke similar responses and verbal output as one would get when conversing with a human clinician.

6.2 Wizard of Oz Study

The Wizard of Oz study is the first study done using this system. The purpose of this study is to identify the range of conversation variations, and key words that come up in these conversations. Testing is performed on participants selected from the Aphasia Rehabilitation Research Laboratory at Temple University. Individuals were chosen who suffer from chronic aphasia, at a

minimum of six months post stroke. Client evaluation on language, short-term memory, and cognition were done at Dr. Nadine Martin's laboratory prior to enrollment. Participants chosen are limited to those with primarily production difficulties in word retrieval or sentence productions. Potential participants were administered the Western Aphasia Battery-Revised[31] test to determine the type of aphasia and to profile the participant's language and communication abilities.

Participants are sat in front of a large screen in which the virtual clinician is projected. The clinician's dress and background are fitted the given script's scenario that is used for the test, a scenario that involves being seated for a diner will involve a diner background scene, and the avatar in a waitress uniform. The participant is recorded using the Kinect system as to gather future data on the participant's reaction and emotional state, as well as record the audio of the session.

The human clinician, now known as the Wizard, is seated in a separated room and given access to the Aphatar Control Panel system. The Aphatar Control Panel consists of a series of controls to change the emotional state of the avatar, and a text input box. To facilitate the speed and flow of a conversation, a nearby Quik-Box is implemented to allow the Wizard to select any of the script's lines and instantly have the avatar synthesis them into speech. Furthermore, the inputted text can contain a command-key to change the avatar's emotional state on a per-sentence basis. The Aphatar Control Panel takes the inputted or selected text string, and then passes it to Microsoft's Speech API, which synthesis the given text string into spoken speech which the participant can hear. When the spoken speech is

outputting, the Avatar plays a generic talking animation until the speech is finished output. The participant then responds to the avatar, and the Wizard chooses the correct response from the list, or in situations which have not been accounted for, types in their own ad-lib response.

The main focus for this part of the project was user friendliness. As the Wizard will be operating the avatar and the speech output from behind the scenes, inputting the text to be synthesized must be able to be done as quickly and easily as possible. As one of the goals for this test is to find and prepare the automated scripts for potential variations, the ability to ad-lib must also be added.

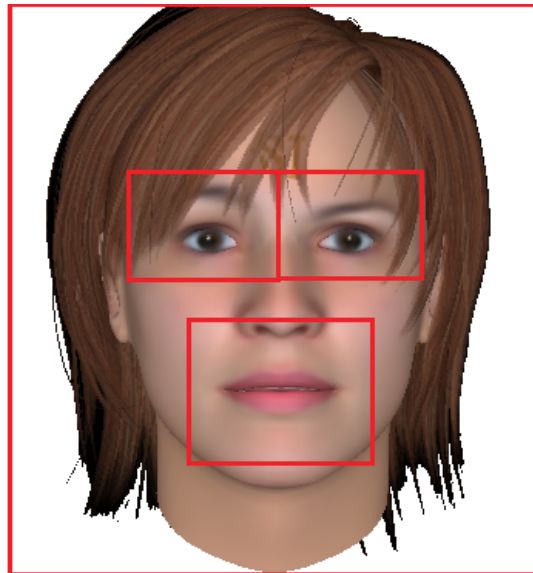


Figure 8 - The face used for the Wizard of Oz project. The borders represent the different pieces of the avatar that can be changed to mimic emotion. This one is displaying 'interest'.

The Avatar created for this project is a white, female avatar. It was built

using the FaceGen Software Suite. The avatar is positioned into a number of preset emotional and animation base states, such as a opening and closing mouth, or moving eyebrows and eyes and these are saved into frames. The frames are further broken apart into four individual pieces; the basic background head and hair, the mouth, and the left and right eye/eyebrow. This allows the avatar to mix and match states to simulate what would be a recognized facial emotion. For references to how to model these emotions, my own face was used. The mix and matching of states allows easy implementation of some emotional states such as confusion, where the avatar will raise one eyebrow, while lowering the other in thought, and paves the way for easy creation of future and variable emotional states.

Please Indicate whether you agree or disagree:	Not sure (0)	Agree (1)	Disagree (2)
The computer therapist seemed like a real person			
I knew other people were in the room when I was talking to the computer therapist			
Talking with the computer therapist was difficult			
I was comfortable talking with the computer therapist			
I could not pay attention to the computer therapist			
I felt more comfortable talking with the real therapist			
It was less easy to talk with the computer therapist than the real therapist			

Figure 9: A sample of the adapted questionnaire used to gauge client's reception of the virtual clinician.

6.2.1 Data Gathering Methodology

Data for the study is analyzed using a technique developed by Linda Nicholas and Robert Brookshire, the Correct Information Unit[33]. The Correct Information Unit analysis is a standardized rule-based scoring system to evaluate the informativeness and efficiency of connected speech. The system shows

significant differences between those who suffer from a form of speech impairment through brain damage, such as aphasics, and those of normal healthy speakers.

The method of this system is as follows. Transcripts of the Wizard of Oz study are gathered, and two scorers score the transcripts independently of each other to ensure reliability.

Clients are tested on the validation and understanding of the main concepts of the scenario. The main concepts of a scenario varies by scenario; in one scenario in which the avatar acts as a travel agent and the client books a trip to Las Vegas, the main concepts would be the idea of going on a trip to Las Vegas, understanding that they are talking to a travel agent, and their own reasons for the trip and what they wish to do there. The client's understanding of the trip is assigned one of the following five scores.

AC: accurate and complete

AI : accurate and incomplete

IN : inaccurate and complete

II : inaccurate and incomplete

Clients are then run through a number of sessions of varying scenarios. These scenarios and their combined scores gives us the means, standard deviations, and ranges for the understanding of the main concept of the scenarios.

Other things which are discussed and measured during the testing is the number of narrative words, words that have meaning to the scenario and convey

information, and utterances and filler words.

It is our belief going into this experiment that with the training and practice with the avatar, the CIU scores will improve with the sessions, and the ratio of narrative words to utterances will improve.

6.2.2 Results

To gather the results, patients were subjected to a series of therapy sessions. The pre-treatment consisted of a two scenarios run, one with the human clinician, and the other with the virtual clinician. Following this, testing was done in sessions spaced across a few days. A sessions was done where both scenarios are done with the human, both are done with the virtual clinician, and two done with the human and virtual clinician handling a single scenario. After these sessions, a final post-treatment sessions was held, following the same structure as the pre-treatment session.

As of this writing, full data analysis is presently being preformed on all the collected data from the clients that we have run using the above methodology. Though analysis is not finished, the pre-treatment and post-treatment phases of the one member of the study has been scored. It is as follows.

Table 3: Results of the first Wizard of Oz study.

	Pre-Treatment CIU Score	Post-Treatment CIU Score
Patient EH4	--	--
Human Clinician	0.77	0.78
Virtual Clinician	0.82	0.82
Patient CN39		
Human Clinician	0.58	0.7
Virtual Clinician	0.8	0.89

The CIU is scored based on the number of narrative, context-words said by the patient in comparison to the number of filler, out-of-context, or utterances made. Results between the human and virtual clinician varied but were close, depending on the individual patient. This stresses the need for the avatar experience to be tailored to the patient for the best results.

It was observed that the avatar elicited more initiation from the patient during the conversations. One abnormality that arose was in Patient CN39 pre-treatment, the patient produced less words, mostly single word responses such as “Yes.” or “No.”, but were considered informative for the conversation. This resulted in a large number disparity when scoring the CIU score.

Overall, the results show that therapy with the avatar is a valid addition to normal therapy, and can produce positive results in users.

6.3 Future Study – Autonomous Virtual Clinician Study

This is a future study that will be done using the information gathered from the Wizard of Oz study. We will once more be gathering participants suffering from aphasia using the same criteria for the Wizard of Oz study, and the information gathering protocols will remain the same. However, all variations and situations that occur in the Wizard of Oz study that deviate from the script will need to be programmed and handled automatically by the Virtual Clinician, rather than a human driver.

Modification and creation of new and better scripts will be mandated based on the common error modalities and variations in conversational flow discovered and observed in the Wizard of Oz test. It is my hope that despite the lack of a human driver, the Spoken Dialog System will be able to maneuver these errors and variations gracefully, and not effect the participants reaction to the virtual clinician negatively. A potential issue that may arise due to the virtual clinician's inability to understand or gracefully adapt to conversations on the fly may cause the participant's immersion to be broken. In such a case a participant may become uncomfortable with the computer interaction and in worse case scenarios, be unwilling or unable to further participate with it

CHAPTER 7

CONCLUSION AND FUTURE WORK

The pilot data hints toward a positive outlook, as both participants showed improvement and tested positively when dealing with the virtual clinician. It is our hope that future results that will come from running more participants through the “Wizard of Oz” test will confirm these results and remove any testing order related biases. Following this acquisition of data, tests can begin using the Olympus/RavenClaw system to remove the driver and make the virtual clinician fully automated.

The Olympus/Ravenclaw ASR system must be expanded to deal with speech and language errors that are present in aphasia. These include disordering of words, word substitutions, pauses and grammatical errors. Such modifications to an existing speech recognition system could prove useful in later advances in speech recognition and accuracy rates involving those with speech disorders or heavy accents as both cases will need to be taken into account.

Script revisions will focus on increasing their fit to the Olympus/RavenClaw ASR system. Such revisions will allow the conversation to flow more smoothly when handled by an automated process and can help create more natural sounding sentence structure and responses that come from the virtual clinician.

The avatar itself will be improved upon as well. Avatars for the virtual clinician must be made in a variety of races and both genders. The avatar can also be

dressed to fit in with the scenario to ease the client into a conversation and provide a sense of immersion during the scene. For example, when order food during a restaurant scene, the avatar will be wearing the outfit of a waiter. Providing the illusion and immersion in this form can remove any unease a client feels with the virtual clinician.

Future work will be carried out with the Olympus/RavenClaw system to give signals for emotion or expressions that the virtual clinician should perform. For example, the avatar should show a confused expression if it did not understand the client, or smile if it did. This must also be done to help synchronize the speech to the virtual clinician's mouth movements so that the effect of the virtual clinician talking is not jarring to the client or looks wholly unnatural.

It is our plan to eventually turn this virtual clinician into a software suite that can be given to a client to use at home, and thereby enable continued therapy after in-clinic treatment has ended. This will provide another means of support to help individuals with aphasia lead productive lives that include communicative interactions with others. We expect significant outcomes of these experiments will be: 1) maximizing use of residual language skills in everyday functional communication situations, 2) efficacy of practice, 3) cost effectiveness and ongoing care beyond in-clinic treatment, and 4) developing a greater database to address questions about factors that influence successful use of residual language skills in functional communication situations. This includes cognitive, neurological and psychosocial factors, severity of impairment, social support systems and other

potential variables. Developing this technology will allow for greater consistency of measurement of this stage of language treatment and will provide an important tool for investigations into many potential factors that influence successful communication with aphasia

REFERENCES

- [1]. Holland, A., Fromm, D.S., DeRuyter, F., & Stein, M. (1996), Treatment efficacy: Aphasia. *Journal of Speech and Hearing Research*, 39, S27-S36.
- [2]. Crinion J.T., Leff, AP. (2007) Recovery and treatment of aphasia after stroke: functional imaging studies. *Current Opinion in Neurology*, 20, 667– 673.
- [3]. Meinzer M, Flaisch T, Breitenstein C, Wienbruch C, Elbert T, Rockstroh B. (2008) Functional rerecruitment of dysfunctional brain areas predicts language recovery in chronic aphasia. *Neuroimage*, 39, 2038-46.
- [4]. Fridriksson, J. (2010). Preservation and modulation of specific left hemisphere regions is vital for treated recovery of anomia in stroke. *The Journal of Neuroscience*, 30 (35), 11558-11564.
- [5]. Aten, J. L., Cligiuri, M. P. & Holland, A. L. (1982) The efficacy of functional communication therapy for chronic aphasic patients. *Journal of Speech and Hearing Disorders*, 47, 93-96.
- [6]. Ahlsen E. (2005) Argumentation with restricted linguistic ability: performing a role play with aphasia or in a second language. *Clinical Linguistics and Phonetics*, 19, 433-451.
- [7]. Cherney, L. R., Halper, A.S., Holland, A.L, and Colke, R. (2008). Computerized script training for aphasia: Preliminary results. *American Journal of Speech-Language Pathology*, 17, 19-34.
- [8]. Holland, A., Frattali, C., & Fromm, D. (1998). *Communicative abilities in daily living*. Austin TX: Pro-Ed.

- [9]. Garcia, L.J., Rebolledo, M. Metthe, L., and Lefebvre, R. (2007) The potential of virtual reality to assess functional communication in aphasia. *Topics in Language Disorders*, 27, 272-288.
- [10]. Microsoft. (2011) Xbox 360 + Kinect. [Online]. <http://www.xbox.com/en-US/kinect>
- [11]. Stephanie Crawford. (2011, November) How Microsoft Kinect Works. [Online]. <http://electronics.howstuffworks.com/microsoft-kinect2.htm>
- [12]. Microsoft. (2011, August) Xbox Support - Kinect Sensor Components. [Online]. <http://support.xbox.com/en-US/kinect/more-topics/kinect-sensorcomponents>
- [13]. Zalevskym Shpunt, Maizels, and Garcia (2005) Method and System for Object Reconstruction
- [14]. Bohu, Raux, Harris, Eskenazi, and Rudnick (2007) Olympus: an open-source framework for conversational spoken language interface research
- [15]. Huang, X., Alleva, F., Hon, H.-W., Hwang, M.-Y., Lee, K.-F. and Rosenfeld, R., 1992. The SPHINX-II Speech Recognition System: an overview
- [16]. Carnegie-Mellon University (2012) CMUSphinx : Basic concepts of speech [Online]. <http://cmusphinx.sourceforge.net/wiki/tutorialconcepts>
- [17]. Allen, Jonathan, Hunnicutt, M. Sharon, Klatt, Dennis (1987). *From Text to Speech: The MITalk system*. Cambridge University Press. ISBN 0-521-30641-8
- [18]. Van Santen, Jan P. H.; Sproat, Richard W.; Olive, Joseph P.; Hirschberg, Julia (1997). *Progress in Speech Synthesis*. Springer. ISBN 0-387-94701-9.
- [19]. Alan W. Black, Perfect synthesis for all of the people all of the time. IEEE TTS Workshop 2002.

- [20]. John Kominek and [Alan W. Black](#). (2003). CMU ARCTIC databases for speech synthesis. CMU-LTI-03-177. Language Technologies Institute, School of Computer Science, Carnegie Mellon University.
- [21]. Julia Zhang. *Language Generation and Speech Synthesis in Dialogues for Language Learning*, masters thesis, Section 5.6 on page 54.
- [22]. Charpentier, F.; Stella, M., "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.* , vol.11, no., pp.2015,2018, Apr 1986 doi: 10.1109/ICASSP.1986.1168657
- [23]. L.F. Lamel, J.L. Gauvain, B. Prouts, C. Bouhier, R. Boesch. Generation and Synthesis of Broadcast Messages, *Proceedings ESCA-NATO Workshop and Applications of Speech Technology*, September 1993.
- [24]. Cherney, L. R., Halper, A.S., Holland, A.L, and Colke, R. (2008). Computerized script training for aphasia: Preliminary results. *American Journal of Speech-Language Pathology*, 17, 19-34.
- [25]. Thompson, C.K., Choy, J.J., Holland A., Cole R. (2010). Sentactics®: Computer-Automated Treatment of Underlying Forms. *Aphasiology*. 24, 1242-1266S.
- [26]. Morrison R. (2009). Empathy from Avatars: Propositions for Improving Trust Development in Pseudo-Social Relationships with Avatars. *European Journal of Social Sciences*, 12, 298-309.
- [27]. Cerulo, K. A. (2009). Nonhumans in social interaction, *Annual Review of Sociology*, 35, 531-552.

- [28]. Turkle S. (2003). Technology and human vulnerability. A conversation with MIT's Sherry Turkle. *Harvard Business Review*. 81: 43-50.
- [29]. Turkle, S, Taggart W, and Kidd CD. (2006). A sociable robot to encourage social interaction among the elderly. Proceedings of the 2006 *IEEE International Conference on Robotics and Automation*, Orlando, Florida, May 2006.
- [30]. Bohus, D. and Rudnicky, A. The RavenClaw dialog management framework: architecture and systems.. *Computer Speech and Language*, 2008, 23(3), 332-361.
- [31]. Saffran, E.M., Berndt, R.S. & Schwartz, M.F. (1989). The quantitative analysis of agrammatic production: Procedure and data. *Brain and Language*, 37: 440-479.
- [32]. Julius O. Smith III (2011) [Online] "[Additive Synthesis \(Early Sinusoidal Modeling\)](https://ccrma.stanford.edu/~jos/sasp/Additive_Synthesis_Early_Sinusoidal.html)"
https://ccrma.stanford.edu/~jos/sasp/Additive_Synthesis_Early_Sinusoidal.html
- [33]. Laura E. Nicholas and Robert H. Bookshire (1993). A system for Quantifying the Informativeness and Efficiency of the Connected Speech of Adults With Aphasia. *Journal of Speech and Hearing Research*, 36 :338-350.
- [34]. Dartmouth College (2011) "Music and Computers" [Online].
<http://music.columbia.edu/cmc/MusicAndComputers/>
- [35]. John Holmes and Wendy Holmes (2001). *Speech Synthesis and Recognition* (2nd ed.). CRC. ISBN 0-7484-0856-8.
- [36]. Michael Adcock "Arcade Emulation How-To" (1997) [Online]
<http://textfiles.meulie.net/games/ARCADE/aehowto.txt>
- [37]. Alan Black and Kevin Lenzo. (2001) Flite: a small fast run-time synthesis engine.

Carnegie-Mellon University

[38]. Hunt, M., D., Z., and R., C. Issues in high quality LPC analysis and synthesis. In *Eurospeech89* (Paris, France, 1989), vol. 2, pp. 348–351.

[39]. Wenjun Zeng, Zehngyou Zhang (2012) Microsoft Kinect Sensor and Its Effects

[40]. Cai, Gallup, Zhang, and Zhang (2010) 3D Deformable Face Tracking with a Commodity Depth Camera.

[41]. K. Khoshelham (2011) Accuracy Analysis of Kinect Depth Data

[42]. J Shotton, Fitzgibbon, Cook, Sharp, et al. (2011) Real-Time Human Pose Recognition in Parts from Single Depth Images

[43]. Antonio R. Damasio, "[Aphasia](#)", "N engl J Med," Feb. 20, 1992

[44]. Carenotes, "[General Information: Aphasia](#)", *Truven Health Analytics Inc.*, 2012

[45]. van de Sandt-Koenderman WM (February 2011). "Aphasia rehabilitation and the role of computer technology: can we keep up with modern times?". *Int J Speech Lang Pathol* **13** (1): 21-7. doi:10.3109/17549507.2010.502973. PMID 21329407.

[46]. Cherney, L. R., Halper, A.S., Holland, A.L, and Colke, R. (2008).

Computerized script training for aphasia: Preliminary results. *American Journal of Speech-Language Pathology*, 17, 19-34.