Improving Ability to Verify Audio CAPTCHAs to Serve Visually-Impaired

IPRO 316 Final Report
Instructor: Matt Bauer

Spring 2010

Contents

List of Figures

List of Tables

1 Abstract

CAPTCHAs (Completely Automated Public Turing Test to Tell Computers and Humans Apart) are used to prevent automated access to sensitive information online. In its usual format, users are presented with distorted text and asked to enter the displayed text in an answer box. If successful, humans, but not computers, will be able to interpret the distorted text. Another format of CAPTCHAs asks users to identify audio information (usually a string of digits or phrase of words) that has been distorted or placed against a background of noise ("white" noise, reversed speech, etc.). Users type the words they hear into an answer box. The audio format is intended to be accessible to blind and low-vision users who cannot use the visually-based format. Unfortunately, audio CAPTCHAs are difficult for humans to use (Bigham and Cavendar 2009) but relatively easy for computers to solve (Tam et al. 2008), which is exactly the opposite outcome desired. To take two extreme examples, in one recent study (Sauer et al. 2008), users were able to solve only 46 percent of audio CAPTCHAs, while in another study (Burztein and Bethard 2009), a computer program was able to break 75 percent of audio CAPTCHAs. At issue is whether audio CAPTCHAs can be designed so that users can easily solve them but computers cannot. Yan and Ahmad (2008) propose testing different kinds of background noise to determine which is the most effective at blocking computers but admitting humans. To this end, Tam et al. (2008) suggest using other human voices as background noise (to thwart computers) but familiar phrases as the string to decode (to aid listeners). The current project will focus on selecting from a set of potential solutions to test. These solutions include:

1. Using two concurrent undistorted, high-quality streams of speech, both of which can serve as the target signal or background noise.
2. Manipulating talker characteristics, such as speaking rate and intonation.
3. Using audio puzzles, such as a math problem, sound question, or an instruction.
4. Use portions of music and ask the user to input what lyrics they have heard.
5. Changing the interface to something that would be easier for low-vision users to use but perhaps harder for computers.

2 Background

2.1 Previous Semester's Work

This IPRO continues the basic work of IPRO 316 F08 and S09 [relabeled as IPRO 316 for Spring 2010] in examining acoustic and cognitive factors that contribute to understanding speech for public and commercial purposes. In fall 2008, the project focused on auditory factors that may improve accuracy of taking customer orders in a simulated fast food drive-thru environment. In spring 2009, the project identified and proposed benchmarks for speaking rate and pitch of synthetic speech designed for public announcements. The goal of the current project is to improve user ability to solve audio CAPTCHAs (Completely Automated Public Turing Test to Tell Computers and Humans Apart) while preventing computers from doing so.

2.2 Current Semester's Work

The basic problem the current IPRO team will address is how to make an audio CAPTCHA highly intelligible to human users while impossible to solve by computers. Much of the current work aimed at improving audio CAPTCHAs ignores characteristics of the signal for users to interpret, such as speaking rate (syllables per second), fidelity of the signal, degree of hyperarticulation ("enunciation" of speech sounds), etc. The current project will add to existing work by testing these and other factors known to affect a speech signal's intelligibility. In terms of contemporary significance, audio CAPTCHAs are currently used mostly by people with blindness or low vision, but Soupionis et al. (2009) note that audio CAPTCHAs may be useful in Voice over IP contexts to combat spam over Internet telephony (e.g. bots making calls for political or product advertisement). Therefore, the quality of audio CAPTCHAs has widespread significance for the general population.

2.3 Ethical Issues

The Institution Review Board (IRB) reviews research proposals that involve human participants. We submitted our application to the IRB which was approved. Under last semesters application we were required to indicate the purpose and scope of the experiment, the type of testing that would be involved and importantly, indicate the ethical implications of the study. The ethical issues which were indicated were:

1. There may be a slight discomfort from wearing headphones for an extended period of time, as well as possible discomfort from sitting in a chair for an extended period of time.
2. There was also a possibility that participants could experience increased stress from being asked to make decisions quickly.

In addition, this semester every member of IPRO 316 completed the National Institutes of Health (NIH) Ethics Training Course with a combined total of over twenty hours of training.

3 Objectives

This IPRO's goal is to study and recommend one or more various potential means of decreasing computers ability to break audio CAPTCHAs while at the same time improving human ability to break them. To that purpose:

I. The IPRO team will study speech and speech recognition in terms of its acoustical properties in order to better understand the problem and potential solutions.

II. IPRO sub teams will explore possible solutions and their pros and cons in order to select which options should be pursued.

III. The IPRO team shall devise and conduct an experiment to determine how effective the chosen solutions are at improving audio CAPTCHAs.

IV. The IPRO team shall devise a recommendation as to how to improve audio CAPTCHAs based on an analysis of the data obtained in the experiment.

4 Methodology

4.1 Work Breakdown Structure

4.1.1 Phase One

Table 1: Work Breakdown Structure - Phase One

Task	**Description**	**Deadline**
Learn Acoustic Foundations of Speech	The team will learn the fundamentals of acoustics and how this transposes into speech.	01/28/10
Project Plan/IRB Form	Revise and submit the project plan and the IRB form.	02/05/10
Budget Proposal	Revise and submit the proposed budget.	02/05/10
Ethics Training	Complete the web training on research ethics.	02/10/10
Evaluate Existing CAPTCHAs	A team will evaluate existing CAPATCHA systems for further analysis.	02/17/10
Devise Solutions	A team will devise possible solutions and applicable experiments to test those solutions.	02/17/10
Midterm Presentation	A team will compile the data acquired and present.	02/23/10

4.1.2 Phase Two

Table 2: Work Breakdown Structure - Phase Two

Task	**Description**	**Deadline**
Recruitment	A team will recruit persons to be our test subjects.	04/01/10
Design Stimuli	A team will devise the stimuli necessary for the experiment.	04/13/10
Design Measurement Tools	A team will formulate accurate tools to measure results achieved from experiment.	04/13/10
Administer the Experiment	The team will administer the experiment and compile results.	04/13/10
Plan of Analysis	A team will construct a plan to analyze data obtained from experiment.	04/13/10

4.1.3 Phase Three

Table 3: Work Breakdown Structure - Phase Three

Task	**Description**	**Deadline**
Analyze Results	A team will analyze results from experiment.	04/15/10
Final Report	A team will craft the final report including the analysis of results and further recommendations.	04/20/10
Final Presentation	The team will present the findings from the IPRO.	04/23/10

4.2 Changes Made to Work Breakdown Structure

In almost all cases, no changes were necessary to the work breakdown structure. All goals were completed on time.

4.3 Experiment Methodology

Experiment methodology here.

4.4 Sphinx

Sphinx is a collection of packages used to recognize speech from streams of audio. It was originally developed by Carnegie Mellon University and is an open source (modifications to the code are possible/permitted) project. While the package is quite good, numerous modifications were necessary in order to use this package for testing on stimuli developed.

First, a large language model, or corpus, needed to be procured. The language model serves as the statistical model that the speech recognizer uses to determine what is being said in an audio sample. The basic idea behind this statistical model is that there is a certain probability that a certain word will be next to another word (a so called "2-gram"); or that three words will be found next to each other ("3-gram), and so forth. Google's role on the Internet as a search engine means they index almost every site on the Internet. As a result of this, they have an enormous amount of data on how language is used in the public stream. Thus, they were able to assemble a 1 trillion word token corpus which we were able to purchase and use.

Sphinx is able to use this language model as a kind of road map while processing audio streams. Like a human, Sphinx breaks audio streams up into segments (i.e., syllables) and collects them for processing. These segments are distinguishable from one another due to the unique combination of sound waves that have critical turning points called formants. Based on these formants, Sphinx is able to determine what vowel or consonant sound it's dealing with, and compares that to the language model which allows it to statistically pick the most likely string. However, the original approach taken by Sphinx that had to be changed was that it used what's called a "first-fit" algorithm. Basically, as soon as it finds a result that could work, it uses that. This approach makes sense when dealing with a real-time voice recognizer, but when the most accurate results are desired, this approach is not the best. Thus, the second thing we had to do was to program Sphinx to instead work over the entire language model and find the absolutely best result it can.

Lastly, Sphinx was designed for real-time voice recognition and not for recognizing large samples of audio stimuli. Therefore, we had to program an interface to Sphinx that would allow us to test batches of audio files.

5 Project Budget

Table 4: Project Projected and Actual Expenses

	Spring 2010 Projected Expenses			**Spring 2010 Actual Expenses**		
Experimental Expenses	**Days**	**Price Per Day**	**Total**	**Days**	**Price Per Day**	**Total**
Participant Incentive/Support - Pizza	4	$125.00	$500.00	4	$125.00	$500.00
IPRO Day Expenses	**Amount**	**Price Per Unit**	**Total**	**Amount**	**Price Per Unit**	**Total**
Team Polo Shirts	8	$24.25	$194.00	-	-	-
Exhibit Materials	-	$90.00	$90.00	-	-	-
Other Expenses	**Amount**	**Price Per Unit**	**Total**	**Amount**	**Price Per Unit**	**Total**
Equipment - Microphone	1	$150.00	$150.00	-	-	-
Equipment - Software	1	$150.00	$150.00	1	$150.00	$150.00
TOTAL EXPENSES			$1084.00			$650.00

6 Team Structure and Assignments

To better facilitate the completion of the project's objectives, the team has been divided into groups and roles have been assigned as follows:

The Groups are as follows:

IPRO 316 Team Leader: Maxwell Kaim

Final Report Leader: Gabriel Klansky
Ethics Training Leader: Michael Fabian
Experiment Organizer: Daniel Kipp
Minute Taker: Adam Ciarkowski
Agenda/Time Keeper: Sean Wallace

6.1 Phase One

Table 5: Team Structure and Assignments - Phase One

Group	**Members**	**Description**
Learn Acoustic Foundations of Speech	All	We will learn the Phonetic alphabet and the acoustical properties of speech to improve the tools to be used in the task at hand.
Project Plan	Sean Wallace, Daniel Kipp	This group is in charge of constructing this document (the project plan).
Ethics Training & IRB	All	These people will become certified to administer an experiment of the nature required and handle the submission of the budget and the Institutional Review Board form (IRB).
Evaluate Existing CAPTCHAs	Michael Fabian, Erick Schneider, Maxwell Kaim	This group will collect data on existing CAPTCHAs relating to their format and evaluate what works and what should be changed.
Devise Solutions	All	The IPRO group will come up with potential solutions to the problem that can be explored by way of brainstorming.
Midterm Presentation	Gabriel Klansky, Michael Fabian, Adam Ciarkowski	This group shall handle the midterm presentation in terms of both fabrication and the actual presenting.

6.2 Phase Two

Table 6: Team Structure and Assignments - Phase Two

Group	**Members**	**Description**
Recruitment	Gabriel Klansky	The IPRO group will run experiments to determine the effectiveness of a chosen solution.
Design Stimuli	Erick Schneider, Maxwell Kaim, Daniel Kipp	These people will be in charge of designing the stimuli to be used in the experiments.
Design Measurement Tools	Sean Wallace, Adam Ciarkowski	This group will design the measuring tools for the collection of experimental data.
Administer the Experiment	Erick Schneider, Adam Ciarkowski	These people will be in charge of any administrative aspects of the experimental process.
Plan of Analysis	Sean Wallace, Adam Ciarkowski	This group will plan how to analyze the data gathered in the experimental process.

6.3 Phase Three

Table 7: Team Structure and Assignments - Phase Three

Group	Members	Description
Analyze Results	Sean Wallace, Daniel Kipp	These people will perform the analysis of the experimental data.
Final Report	All	The IPRO group will build a report of our findings and recommendations.
Final Presentation	Gabriel Klansky, Michael Fabian, Adam Ciarkowski, Sean Wallace	This group will present our findings in the final presentation.
Oral Presentation	Erick Schneider, Daniel Kipp	This group will perform the IPRO oral presentation.
IPRO Booth	All	The team will present findings to all interested at IPRO day.

7 Team Members' Background and Expectations

7.1 Team Members' Background

Table 8: Team Members' Background

Name	Major	Year	Team(s)	Skills	Interests
Gabriel Klansky	Humanities	3rd	Final Report Leader, Learn Acoustic Foundations of Speech, Ethics Training, Devise Solutions, Midterm Presentation, Recruitment, Final Presentation, Final Report, IPRO Booth	MS Office, Praat, semi-fluent in French	Music, Art, Photography, Writing, Food
Erick Schneider	Computer Science	3rd	Learn Acoustic Foundations of Speech, Ethics Training, Devise Solutions, Evaluate Existing CAPTCHAs, Design Stimuli, Administer the Experiment, Oral Presentation, Final Report, IPRO Booth	MS Office, Java, Visual Basic, Scala, GIMP	Computers, Gaming, Books
Maxwell Kaim	Computer Science / Psychology	4th	Team Leader, Learn Acoustic Foundations of Speech, Ethics Training, Devise Solutions, Evaluate Existing CAPTCHAs, Design Stimuli, Final Report, IPRO Booth	Java, C, C++, Python, Open Office, Open Canvas, Photoshop, Psych Testing Experience	Language / Semantics, Natural Language Processing, AI Programing, Drawing, Creative Writing

Continued on next page...

. . . Continued from previous page

Name	Major	Year	Team(s)	Skills	Interests
Adam Cia-rkowski	Computer Science	3rd	Minute Taker, Learn Acoustic Foundations of Speech, Ethics Training, Devise Solutions, Midterm Presentation, Design Measurement Tools, Administer the Experiment, Plan of Analysis, Final Presentation, Final Report, IPRO Booth	MS Office, Java, C, Sound Mixing	Music, Games, Computers, Bike-riding
Michael Fabian	Computer Science	3rd	Ethics Training Leader, Learn Acoustic Foundations of Speech, Ethics Training, Devise Solutions, Evaluate Existing CAPTCHAs, Midterm Presentation, Final Presentation, Final Report, IPRO Booth	C, C++, Perl, Java, HTML, MS Office, Visual Basic, SQL	Technology, Movies, Cognitive Science
Sean Wallace	Computer Science	3rd	Agenda/Time Keeper, Learn Acoustic Foundations of Speech, Ethics Training, Devise Solutions, Project Plan, Design Measurement Tools, Plan of Analysis, Analyze Results, Final Presentation, Final Report, IPRO Booth	C, Objective-C, Java, PHP, SQL, Adobe Programs, MS Office, JavaScript, HTML	Technology, Wakeboarding, Movies
Daniel Kipp	Mathematics	3rd	Experiment Organizer, Explore Multiple Voices and Music Lyrics, Learn acoustic foundations of speech, Devise Solutions - Design Experiment, Ethics Training, Generate Project Plan, Running Experiments, Designing Stimuli, Analyze Results, Final Report, Final Presentation Booth	OpenOffice, MS Office, Audacity, Java, C, Praat, Fourier Analysis, Probability & Statistics, Matlab, Derive, Python	Music, Reading, Math, Video Games, Puzzles, Climbing Things, Skiing, Board Games

7.2 Team Members' Expectations

Table 9: Team Members' Expectations

Name	Short Term Goals	Long Term Goals
Gabriel Klansky	I hope that I will be able to express the difficulties and intricacies of this project in the presentations and reports. Also I look forward to running an experiment and seeing through to completion.	I hope to further my knowledge and ability in phonetics and language processing. I hope that through the project I will improve my skills in writing in a academic setting.
Erick Schneider	To bring my abilities and experience to the group to facilitate the development of a new type of CAPTCHA that will be both easy for humans to recognize and understand, yet hard for an intelligent agent to reason out.	To do the best possible in order to place high in IPRO day and to lay out a foundation for the next group of individuals who will work on this project.

Continued on next page. . .

. . . *Continued from previous page*		
Name	**Short Term Goals**	**Long Term Goals**
Maxwell Kaim	Keep the team on track and together in terms of our deliverables, benchmarks, goals, et cetera. To keep my own contributions on time and their content up to group expectations.	To learn something if not useful than at least interesting about language and language analysis.
Adam Ciarkowski	Explore several different background sounds and being able to successful incorporate them with spoken words, Prepare presentations, Discuss various options for solutions.	Work together as a group to build a solid final presentation, Learn how the experiment process works when testing others, Further improve my knowledge about sound and speech.
Michael Fabian	To successfully help in recruiting participants for our experiments, work with the group as a whole to devise the next step in research in thus upcoming technology, and to help in providing a supportive environment through ideas of each team member may be freely expressed.	To provide valid experimental results for future research to build upon through strong teamwork and dedication to the problem at hand.
Sean Wallace	Better round my knowledge and experience with technology by utilizing previous skills to accomplish new things.	To learn more about language as it pertains to technology and hopefully help focus future study on this subject more precisely.
Daniel Kipp	Within my groups I hope to complete my tasks well and on time. I hope to learn things about speech and acoustics and I expect things to run reasonably smoothly.	I hope to have a successful project that is also interesting.

8 IPRO 316 Code of Ethics

Ethical considerations are the main priority for IPRO 316. With this in mind, IPRO 316 has an obligation to articulate its basic values, ethical principles, and ethical standards. The IPRO 316 sets forth these values, principles, and standards to guide members conduct. The Code is relevant to all student and faculty members, regardless of their professional functions, the settings in which they work, or the populations they serve.

All, personal and professional, conduct taken by IPRO 316 members shall adhere to state and legislative laws. Toleration of lawbreaking will not occur, regardless of any progress breaking or bending the rules will bring. Should any of the laws be broken, then consequences none other than arrest shall be made.

No member shall reveal facts, data, or information without prior consent of students participating in experiment or data conveyed to him or her by advising faculty members. Discussion of results and or the progress IPRO 316 makes through experimentation that involves revealing results of specific individuals with non-IPRO 316 members shall not occur as all data should be kept confidential.

All personal conduct taken by members of IPRO 316 that either directly or indirectly relates to coursework for the progress of IPRO 316 shall remain professional. At any time a member is publicly representing IPRO 316 they shall behave with the utmost professional manner. Any misconduct will reflect poorly against IPRO 316 and could compromise its continuation.

Any progress to be achieved by IPRO 316 shall be innovative and any challenges will be taken constructively. Actions taken that can influence the goals of IPRO 316 are to only be for improvement. Any detrimental effects could compromise its continuation.

The services provided by IPRO 316 members requires honesty, impartiality, fairness and equity. These services also must be dedicated to the betterment of public health, safety, and welfare of the group and community. If it is found and proven that a member of IPRO 316 has said or was responsible for acting against any of these qualities, it is up to the advisor to determine his or her future with IPRO 316.

IPRO 316 members adhere to abilities of utmost honesty and integrity in all relations. At no time shall any data or analysis be revealed that contain sensitive information without being discussed with all members and advisor. Severity of the consequences can only be determined by the type and seriousness of the released information.

Student members of IPRO 316 shall not attempt to obtain recognition or attempt to increase their status within the group by untruthfully criticizing or creating deception among other members. Rewards of completing a task shall be given to all members involved, not disregarding any member so as to take full credit. If partial credit is found and not directed towards a specific individual because the leader evidentially chose not to disclose this fact shall face consequences determined by the advisor of IPRO 316.

9 Results

9.1 Expected Results

At the beginning of the semester, we anticipated that our results would allow us to provide a recommendation as to how to improve audio CAPTCHAs by increasing computer failure rate and human success rate. Also, we expected our findings to illustrate why our recommendation makes sense as opposed to other potential solutions as well as approximate the magnitude of its advantage.

9.2 Observed Results

Over the course of our experimentation, we found that most of our participants handled the test well with only a few exceptions with people who could not comprehend the audio. The majority of the participants in the study scored higher than the mean of the total scores (see histogram in figure 1). The mean scores of all the human tests show that no matter the transformation, the ability for humans to recognize the sentences was over 70% (see figure 3). However, with Sphinx's trials, we found that while it scored over 60% on the untransformed files, it was unable to understand the 25ms transformation and performed poorly (less than 20%) for Dutch, echo, and fast transformations (see figure 2). Because we were able to find participants from all over the world, we were able to also determine how well native English speakers did compared to non-native English speakers for each of the transformations. Based on this data, we were able to conclude that on average native speakers of English scored 20.23% higher than non-native speakers of English (see figure 4).

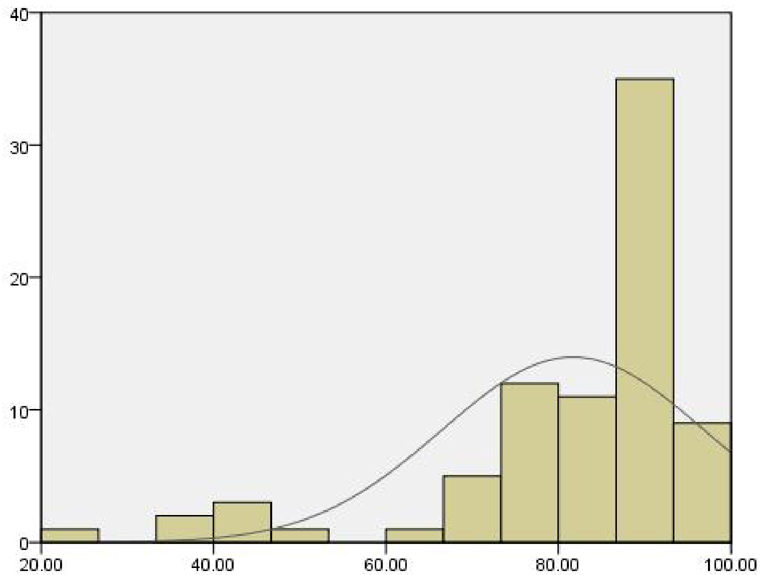

Figure 1: Histogram of overall average scores achieved by experiment participants.

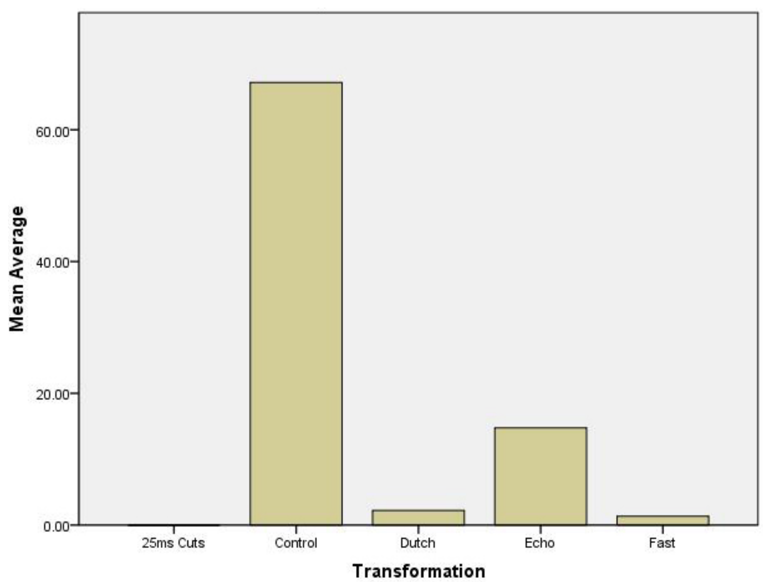

Figure 2: Mean scores achieved by Sphinx broken down by transformation type.

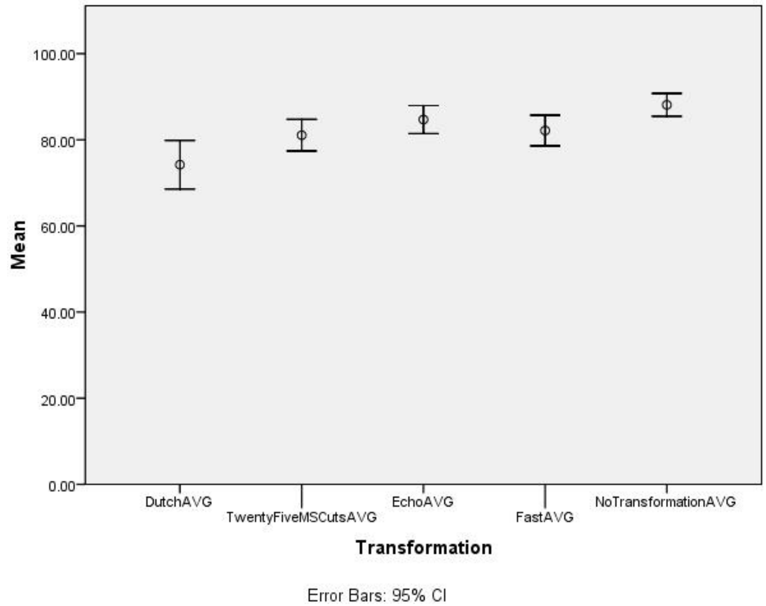

Figure 3: Mean scores achieved by experiment participants broken down by transformation type.

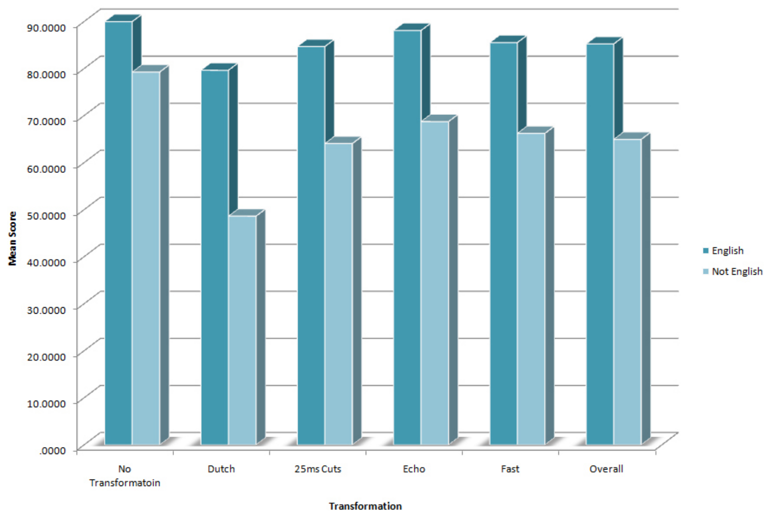

Figure 4: Means achieved by native English and non-native speakers of English.

10 Obstacles

There were few obstacles encountered during our IPRO, but those that did show up were easily dealt with. Here is a brief description of the obstacles and how we resolved them.

1. At the beginning of the semester, we entered the IPRO not knowing much about how words are pronounced and how the ear translates sounds in to recognizable speech. We spent the first two weeks of classes just learning how each syllable sounds and how to it is pronounced. We spent a brief period of time also learning how sounds are pronounced in other languages.

2. Right before our testing phase, we had to come up with different transformations that could be done on a sound file to make it harder for a computer to understand what's been said. Using data from previous years, other research done in the field, and what we learned in the first two weeks, we were able to come up with ideas, then implement them for use in our experiment.

3. As we finished creating the different transformations and making sound files for testing, we realized we had to find a program that would allow us to test to see how well a computer could understand our models. We spent an entire class period looking up speech-to-text programs that were available to us before settling with Sphinx, an open-source speech-to-text program. See section 4.4 for more information about Sphinx.

4. Once we had Sphinx, we realized that the corpus of words that had originally come with the program would not be enough for what we needed to test. We were forced to search for another corpus of words that we could implement in Sphinx to allow us to test our sound files against it. Eventually we discovered that Google had a trillion word corpus and we were able to purchase it and use it in our program. See section 4.4 for more information.

5. Making the necessary modifications to Sphinx to allow us to batch test audio files. While the methods described in section 4.4 might sound trivial, they were nothing but. In fact, it took nearly the entire semester to make the necessary modifications to allow us to test our audio stimuli as we needed.

11 Recommendations

To further improve the quality of audio CAPTCHAs, in order to both help the user understand speech better and speech recognizers to understand speech less, the team recommends that:

1. Both the 25ms cuts and the faster speech transformations should be looked into further to possibly improve the success rate of these transformations.

2. The Dutch and echo transformations shouldn't be looked into further. The speech recognizer had the highest recognition, among the transformations, with the echo transformation. The Dutch transformation was difficult for bilingual users to understand.

3. An improved speech recognizer to perform better against our stimuli, this way we can actually see how these CAPTCHAs would perform in the real world in which complex speech recognizers are being used to crack current CAPTCHAs.

4. New and unique stimuli should also be considered as there should be a large variety of stimuli in current audio CAPTCHAs to avoid speech recognizers targeting a certain transformation or set of transformations. An example of this could be the spoken puzzles concept discussed earlier which does not rely on transformations but instead a question requiring an answer.

These recommendations are worth looking into as they can eventually result in improved audio CAPTCHAs. Improving a speech recognizer to target certain transformations and potentially eliminate them can give us an idea on how well our stimuli and future stimuli will perform in a human vs. computer-type of results. New ideas for stimuli will add to a variety of transformations, and the more complex these transformations are (while still allowing humans to recognize the spoken words fairly well), the better for the future of audio CAPTCHAs.

12 Acknowledgments

We would like to acknowledge Google for the most comprehensive language model available and Carnegie Mellon University for their continued research in the field of speech recognition and making Sphinx publicly available.

13 References

Bigham, Jeffry P., and Anna C. Cavender. “Evaluating Exists Audio CAPTCHAs and an Interface Optimized for Non-Visual Use.” Print.

Bursztein, Elie, and Steven Bethard. “Decaptcha: Breaking 75% of EBay Audio CAPTCHAs.” Print.

Tam, Jennifer, Jari Simsa, Sean Hyde, and Luis Von Ahn. “Breaking Audio CAPTCHAs.” Print.

Tam, Jennifer, Jiri Simsa, David Huggins-Daines, Juis Von Ahn, and Manuel Blum. “Improving Audio CAPTCHAs.” Print.

Yan, Jeff, and Ahmad Salah El Ahmad. “Usability of CAPTCHAs Or Usability Issues in CAPTCHA Design.” Print.