# Improving Voice Recognition Prompts for Users in Various Application Environments

IPRO 316
Advisor: Matthew Bauer

Fall 2010

# Table of Contents


1. Abstract . . . . . 3
2. Background . . . . . 3
3. Objectives . . . . . 4
4. Methodology . . . . . 4
    4.1. Work Breakdown Structure . . . . . 4
        4.1.1. Phase One . . . . . 4
        4.1.2. Phase Two . . . . . 5
        4.1.3. Phase Three . . . . . 5
    4.2. Changes Made to Work Breakdown Structure . . . . . 5
    4.3. Experiment Methodology . . . . . 5
5. Project Budget . . . . . 5
6. Team Structure and Assignments . . . . . 6
    6.1. Phase One . . . . . 6
    6.2. Phase Two . . . . . 6
    6.3. Phase Three . . . . . 7
7. Team Members' Background and Expectations . . . . . 7
    7.1. Team Members' Background . . . . . 7
    7.2. Team Members' Expectations . . . . . 9
8. IPRO 316 Code of Ethics . . . . . 10
9. Results . . . . . 11
    9.1. Expected Results . . . . . 11
    9.2. Observed Results . . . . . 11
10. Obstacles . . . . . .14
11. Acknowledgments . . . . . 15
12. References . . . . . 15

# 1. Abstract

Users of speech recognition technology often hyperarticulate (i.e., exaggerate) their speech in response to recognition failures and subsequent requests to repeat (e.g., "I'm sorry, I didn't understand, please repeat the input."). Hyperarticulation usually leads to further recognition failure. The goal of the current project is to develop a protocol for testing different talker characteristics of voice prompts in speech recognizers with an aim towards minimizing hyperarticulated speech from users. This IPRO is equally suited to students interested in the more technical aspects of acoustic phonetics and voice recognition as well as the cognitive aspects of predicting user behavior in technology-mediated environments.

# 2. Background

This IPRO continued the basic work of IPRO 343 Fall 2008 and Spring 2009, and IPRO 316 Spring 2010 in examining acoustic and cognitive factors that contribute to understanding speech for public and commercial purposes.

Hyperarticulated speech is exaggerated or more extremely produced speech (Lindblom 1990). Speakers will hyperarticulate their speech to overcome noisy work environments (Tufts and Frank 2003), to address children (Kuhl 1997), to address hard-of-hearing listeners (Picheny, Durlach, and Braida 1985), to address pets (Burnham, Kitamura, and Vollmer-Conna 2002), to accent words (Cho 2005), to convey fussiness (Eckert 2005), to indicate salient points within a sentence (Cho 2005), and to express frustration, sadness, excitement and other emotions (Lee et al 2005, Litman and Forbes-Riley 2006, Ververidis and Kotropoulos 2006).

Hyperarticulation involves enhancement of the acoustic signal and modification of the normal movement of the vocal organs. In particular, hyperarticulated speech is louder and higher pitched. Speech segments are longer, and the acoustic vowel space is larger. Jaw displacement from rest position is more extreme, and tongue body movement is more exaggerated, such that articulations requiring the tongue body to be high and front in the vocal tract are sometimes higher and more forward in the mouth (Lindblom and Moon 1994, De Jong 1995, Johnson et al. 1993, Smiljanic and Bradlow 2005).

Several studies have shown that when speech recognizers fail to identify a string of speech and then ask users to repeat the input, users will hyperarticulate their responses (Oviatt, MacEachern and Levow 1998, Swerts, Litman, and Hirschberg 2000, Goldberg, Ostendorf, and Kirchoff 2003, Hirchberg, Litman, and Swerts 2001). Interestingly, as a result of such hyperarticulation, once users are issued such failure-to-understand prompts, recognition rates fall significantly as hyperarticulation increasingly distorts the speech string (Swerts, Litman, and Hirschberg 2000). Thus, an ability to correctly predict how exactly speakers will hyperarticulate speech in failure-to-understand situations is a present challenge for speech researchers (Oviatt, MacEachern and Levow 1998).

One factor related to hyperarticulation in failure-to-understand responses is user emotion. A significant body of literature has shown how emotions of speakers affect their speech (Williams and Stevens 1972, Goldberg, Ostendorf, and Kirchoff 2003, Linnankoski et al 2005, Nordstarnd et al 2004, Lee et al 2005, Litman and Forbes-Riley 2006, and see Ververidis & Kotropoulos 2006 for a bibliography of several dozen other papers). In human-computer interactions, hyperarticulation from frustration is frequently exhibited but can be minimized if

the wording of the error message is apologetic, rather than direct (e.g. “I’m sorry, I didn’t understand. Please say the sentence again,” vs. “Say the sentence again.”) (Goldberg, Ostendorf, Kirchoff 2003.) Another factor related to hyperarticulation in failure-to-understand responses is the user’s desire to be intelligible. Lindblom and Moon (1994) observed that speakers instructed to “speak clearly” will usually hyperarticulate their speech, even if doing so undermines intelligibility of speech.

At issue is whether other talker characteristics of the voice prompt, such as its speaking rate, pitch, intonation, and its own degree of hyperarticulation, influence users’ speech in predictable ways and can further minimize recognition failure.

# 3. Objectives

The goal of the IPRO is to develop a protocol for testing different talker characteristics of voice prompts in speech recognizers with an aim towards minimizing hyperarticulated speech from users and improving recognition success rates.

I. The IPRO team will learn about the acoustic properties of normal and hyperarticulated speech in order to better understand the problem and potential solutions.
II. IPRO subteams will identify relevant factors in the quality of voice prompts to be tested during the experiments.
III. The IPRO team will devise and conduct experiments to test the effect of varying the properties of the voice prompt's speech.
IV. The IPRO team will summarize recommendations for improving voice prompts in voice recognition systems so as to reduce the amount of hyperarticulated speech from users.

# 4. Methodology

## 4.1 Work Breakdown Structure

### 4.1.1 Phase One

Table 1: Description and deadlines of tasks

| Task | Description | Deadline |
|---|---|---|
| Learn Acoustic Foundations of Speech | The team will learn the fundamentals of acoustics and how this affects the way speech is interpreted by humans and computers. | 9/9/10 |
| Project Plan | Revise and Submit the project plan. | 9/12/10 |
| Budget Proposal | Revise and Submit the proposed budget. | 9/12/10 |
| Ethics Training | Complete web training on research ethics. | 9/28/10 |
| Evaluate Existing Voice Prompts | A team will collect recordings of existing voice prompts for further analysis. | 9/16/10 |
| Devise Solutions | The team will devise solutions and experiments to test those solutions. | 9/30/10 |
| Midterm Presentation | A team will compile the data acquired and give a presentation on the current state of the project. | 10/14/10 |

4.1.2 Phase Two

Table 2: Description and deadlines of tasks

| Task | Description | Deadline |
|---|---|---|
| Recruitment | A team will recruit IIT students to be our test subjects. | 10/14/10 |
| Design Stimuli | A team will devise the stimuli necessary for the experiments. | 10/21/10 |
| Design Measurement Tools | A team will design tools needed to gather data during the experiments. | 10/21/10 |
| Administer the Experiments | The team will administer the experiments on test subjects and compile the results. | 11/10/10 |
| Plan of Analysis | A team will construct a plan to analyze data obtained from the experiments. | 11/16/10 |

4.1.3 Phase Three

Table 3: Description and deadlines of tasks

| Task | Description | Deadline |
|---|---|---|
| Analyze Results | The team will analyze the results of the experiments. | 11/25/10 |
| Final Report | A team will write up the final report, including the analysis of the results and further recommendations. | 12/2/10 |
| Final Presentation | A team will present the findings from the IPRO. | 12/3/10 |

## 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

Participants were placed in a soundproof booth for high audio fidelity. A microphone and headphones were placed on their head and they were given a script to read. After they were set up, the volume levels were checked. In order to simulate voice recognition software, the subjects during recording could not see or interact with the testers. While the participants were listening to the stimuli and responding from the script, the testers were recording the subject using a solid-state recorder. At the same time, the testers were listening to the subject's voice and playing the stimuli to most simulate voice recognition software. The experiment took approximately 3-5 minutes to complete for each participant. Participants were all college-aged students of IIT or Shimer College. Among them, many were non-native speakers of English. Once they were done and were taken out of the booth, they were given their incentives of pizza and a raffle ticket.

# 5. Project Budget

| Expenses | Days | Price Per Day | Total |
|---|---|---|---|
| Pizza | 4 | $75.00 | $300.00 |
| Raffle | - | - | $200.00 |

| IPRO Day Expenses | - | Price | Total |
|---|---|---|---|
| Exhibit Materials | - | $90.00 | $90.00 |
| Other Expenses | Amount | Price Per Unit | Total |
| Audio Equipment | - | $20.00 | $20.00 |
| TOTAL EXPENSES | | | $610.00 |

# 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:

**IPRO 316 Team Leader**: Naomi Peterson
**Final Report Leader**: Nithin Winston
**Ethics Training Leader**: Shashank Gopal
**Experiment Organizer**: Andrew Bossemeyer
**Minute Taker**: Alexander Webster
**Agenda/Time Keeper**: Robert Millonzi

## 6.1 Phase One

Table 4: Description of assignments in Phase One

| Group | Members | Description |
|---|---|---|
| Learn Acoustic Foundation of Speech | All | We will learn some IPA and the acoustic properties of speech in order to determine how best to improve voice prompts in recognition systems. |
| Project Plan | Ruth Morrison | Ruth will write the project plan. |
| Ethics Training | All | We will become certified to administer necessary experiments. |
| Evaluate Existing Voice Prompts | Alexander Webster, Vincent Echavarria | This group will collect recordings of existing voice prompts and evaluate their merits. |
| Devise Solutions | All | We will come up with possible solutions to the problems with existing voice prompts. |
| Midterm Presentation | Nithin Winston, Andrew Bossemeyer, Gabriel Klansky | This group will create the slides for and give the Midterm Presentation. |

## 6.2 Phase Two

Table 5: Description of assignments in Phase Two

| Group | Members | Description |
|---|---|---|
| Recruitment | Robert Millonzi, Andrew Bossemeyer, Shashank | This group will recruit IIT students to participate in the |

| | | |
|---|---|---|
| | Gopal | experiments. |
| Design Stimuli | Ruth Morrison, Nithin Winston, Gabriel Klansky | This group will decide on voice quality variables to test during the experiments. |
| Design Measurement Tools | Alexander Webster, Andrew Bossemeyer | This group will design measurement tools used in the experiments. |
| Administer the Experiments | All | We will administer the experiments and record the data collected. |
| Plan of Analysis | Alexander Webster, Andrew Bossemeyer | This group will plan how to analyze the data gathered during the experiments. |

### 6.3 Phase Three

Table 6: Description of assignments in Phase Three

| Group | Members | Description |
|---|---|---|
| Analyze Results | All | We will analyze the data collected in the experiments. |
| Final Report | Nithin Winston | This group will write up the final report containing the findings from the experiments and our recommendations. |
| Final Presentation | Andrew Bossemeyer, Shashank Gopal, Naomi Peterson | This group will give the final presentation. |
| IPRO Booth | All | We will present the findings to all interested at IPRO day. |

## 7. Team Members' Background and Expectations

### 7.1 Team Members' Background

Table 7: Team Members' Background

| Name | Major | Year | Teams | Skills | Interests |
|---|---|---|---|---|---|
| Alexander Webster | Electrical Engineering/ Computer Engineering | 3rd | Minute Taker, Learn Acoustic Foundations of Speech, Ethics Training, Evaluate Existing Voice Prompts, Devise Solutions, Design Measurement Tools, Administer the Experiments, Plan of Analysis, Analyze Results, IPRO Booth | Java, C, Open Office, Breadboarding, MS Paint, Circuit Design, Fourier Analysis | Music, Games, Computers, Gadgeteering |
| Nithin | Biomedical | 4th | Learn Acoustic | MS Paint, | Books, |

| | | | | | |
|---|---|---|---|---|---|
| Winston | Engineering | | Foundations of Speech, Ethics Training, Devise Solutions, Design Stimuli, Administer the Experiments, Analyze Results, Final Report, IPRO Booth | MATLAB, MS Office, AutoCAD, Organizational Skills | Television, Music |
| Vincent Echavarria | Computer Science | 3rd | Learn Acoustic Foundations of Speech, Ethics Training, Evaluate Existing Voice Prompts, Devise Solutions, Administer the Experiments, Analyze Results, IPRO Booth | Java, C++, C, MS Office, OpenOffice, LaTeX | Reading, Games, Computers, Movies |
| Robert Millonzi | Architecture | 5th | Agenda/Time Keeper, Learn Acoustic Foundations of Speech, Ethics Training, Devise Solutions, Recruitment Administer the Experiments, Analyze Results, Final Presentation, IPRO Booth | Photoshop, Illustrator, In Design, and other design software | Architecture, Music, and various other arts |
| Andrew Bossemeyer | Architecture | 5th | Experiment Organizer, Learn Acoustic Foundations of Speech, Ethics Training, Devise Solutions, Midterm Presentation, Recruitment, Design Measurement Tools, Administer the Experiment, Plan of Analysis, Analyze Results, Final Presentation, IPRO Booth | Graphic Design, Leader ship | Baseball, Volleyball, Photography, Sketching |
| Ruth Morrison | Computer Information Systems | 5$^{th}$ | Learn Acoustic Foundations of Speech, Project Plan, Devise Solutions, Design Stimuli, Administer the Experiment, Analyze Results, IPRO Booth | C/C++, Java, Word Processors and LaTeX, Familiarity with IPA and Linguistics | Language, Computers, Programming, Reading |
| Shashank Gopal | Computer Science and Computer Engineering | 4th | Ethics Training Leader, Learn Acoustic Foundations of Speech, Ethics Training, Devise Solutions, Recruitment, Administer the Experiments, Analyze Results, IPRO Booth | Communication, Elective Teamwork, Organization | Music, Reading, Coding |
| Gabriel | Humanities | 4th | Learn Acoustic Foundations of Speech, | Writing, Presenting, | Semiotics, Photography, |

| | | | | | |
|---|---|---|---|---|---|
| Klansky | | | Ethics Training, Devise Solutions, Midterm Presentation, Design Stimuli, Administer the Experiments, Analyze Results, IPRO Booth | Photography, Linguistics background | Communicatio n, Philosophy |
| Naomi Peterson | Computer Science | 4th | Project Leader, Learn Acoustic Foundations of Speech, Ethics Training, Devise Solutions, Administer the Experiments, Analyze Results, Final Presentation, IPRO Booth | Java, MS Office, Leadership, Communication | Speech accents, Music, Computers, Reading |

## 7.2 Team Members' Expectations

Table 8: Team Members' Expectations

| Name | Short Term Goals | Long Term Goals |
|---|---|---|
| Alexander Webster | To create working systems that suit the needs of the experiments and, hence, further research into voice-recognition technology. | To gain valuable experience working with a development team towards furthering a research end. |
| Nithin Winston | I would like to partake in research that will benefit and promote the field of voice-recognition technology. | I would like to have more experience working with a team on a research project. |
| Vincent Echavarria | I want to help improve voice recognition prompts. | I would like to learn more details about voice recognition technology because it looks to be a major part of everyday life in the future. |
| Robert Millonzi | I want to see this group provide meaningful research into the development of voice-recognition software. | To work in a team scenario with various disciplinary backgrounds to achieve a common goal. |
| Andrew Bossemeyer | Develop a command prompt that decreases hyper-articulated responses | To obtain more experience being a team player and be effective in a team environment. |
| Ruth Morrison | I'd like to learn more about the auditory properties of speech, and how other people react to them. | I hope to gain experience with working as part of a team and conducting experiments in order to further research. |
| Shashank Gopal | I would like to learn to use Praat. I would like to understand linguistics. I would like to use ultrasound to understand tongue movement. | I would like to help improve voice recognition prompts. |
| Gabriel Klansky | I hope to run an experiment and analyze the results. I also hope to learn how to analyze speech. | My long term goals are to learn how to be a team player and work in a group effectively. In tandem with that, I hope to learn to subdue my aggressiveness for others. |
| Naomi Peterson | I would like to understand people better, specifically what causes their spoken response to audio directions to change and what changes are caused. | I hope to gain valuable experience in learning new things quickly in a team environment so I can jump into helping with problem-solving almost immediately. |

# 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 made 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 require 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

We expected that by the end of the semester, the IPRO team will have established which talker characteristics of voice prompts elicit the most successfully recognized speech, and will be able to make recommendations leading to more successful voice recognition systems.

## 9.2 Observed Results

A 2x3 Analysis of Variance for independent samples was performed. Factors include speaking rate of a prompt during a simulated voice recognition exercise (slow, medium, and fast) and its intonation contour (normal and skewing monotone). The dependent variable is a measure of success in voice recognition before versus after an error prompt during the exercise. Specifically, spoken responses before and after the error prompt, were submitted to IITSphinx voice recognition software. The software examines audio files and assigns it an interpretation. Then, a comparison of IITSphinx's interpretation was made to the actual string of words, and each response was assigned a score based on the following formula:

$$\frac{\text{Number of words correct (regardless of position) - \# of swaps - \# of additions}}{\text{Length of correct string}}$$

This calculation was done automatically in IITSphinx. Differences in recognition of responses before and after an error prompt, for each participant, were then calculated as follows:

[% correct after error] – [% correct before error]

Results revealed a significant effect for intonation $F(1,73) = 4.7$, $P < 0.05$, but no effect for speaking rate, $F(2,73) = 0.39$, $P = 0.68$, and no interaction $F(2,73) = 0.91$, $P = 0.41$. Results are described in the tables below.

Table 9. Results for intonation

| | **Mean** | **SD** |
|---|---|---|
| Normal | -3.25% | 22.43% |
| Monotone | 7.90% | 22.84% |

Table 10. Results for speaking rate

| | **Mean** | **SD** |
|---|---|---|
| Slow | 1.85% | 21.35% |
| Medium | 4.93% | 28.43% |
| Fast | -0.67% | 18.93% |

Overall, results show that recognition rates of responses following an error prompt improve significantly when the intonation contour of the prompt is natural-like, but nearly monotone. Results for speaking rate of the prompt suggest a medium rate leads to slightly

improved recognition of response compared responses to slower or faster spoken prompts, but differences did not reach significance.

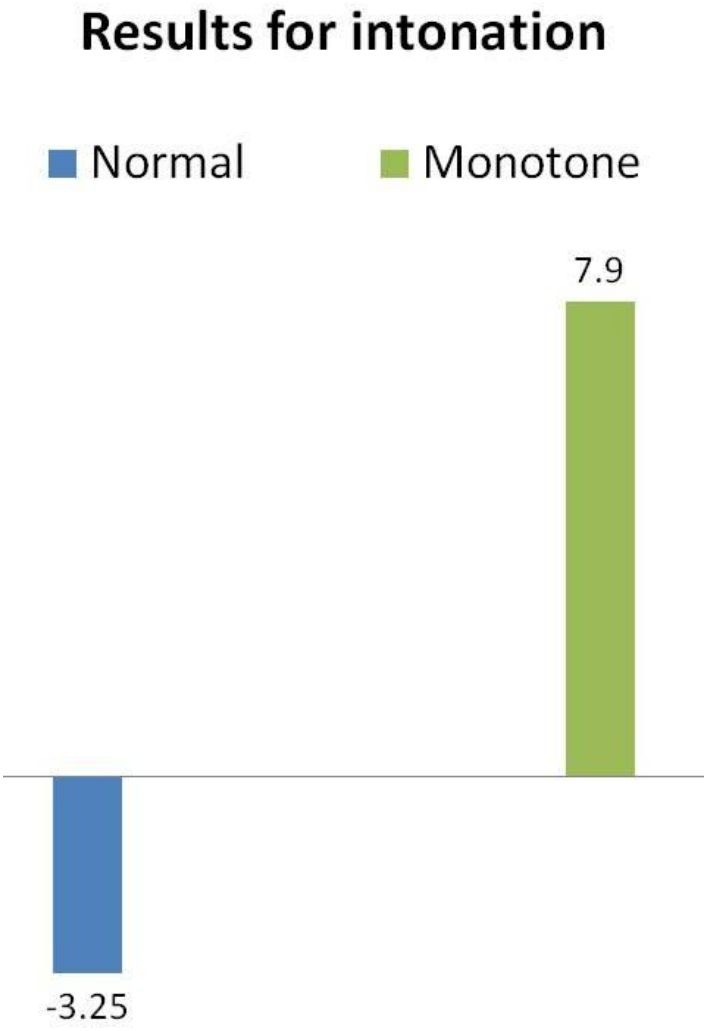


Figure 1: The results due to the variation of intonation

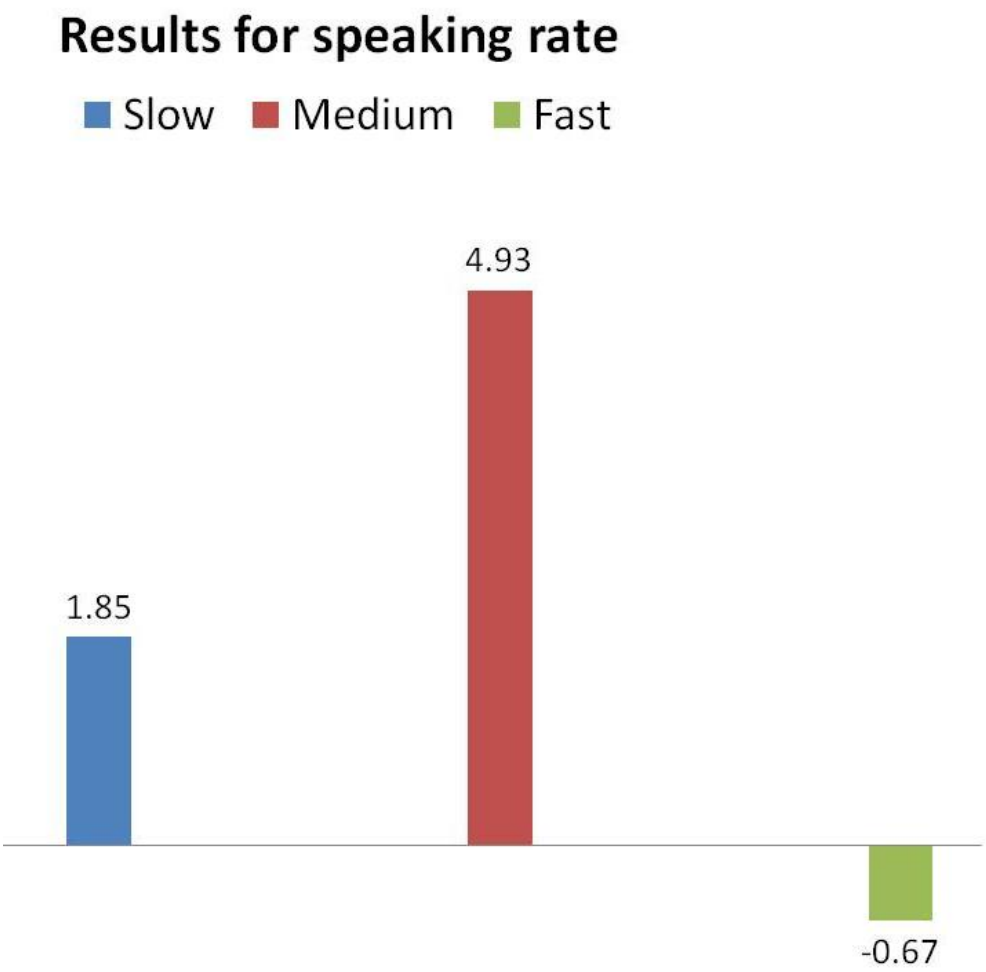

Figure 2: The results due to the variation of the rate of speech

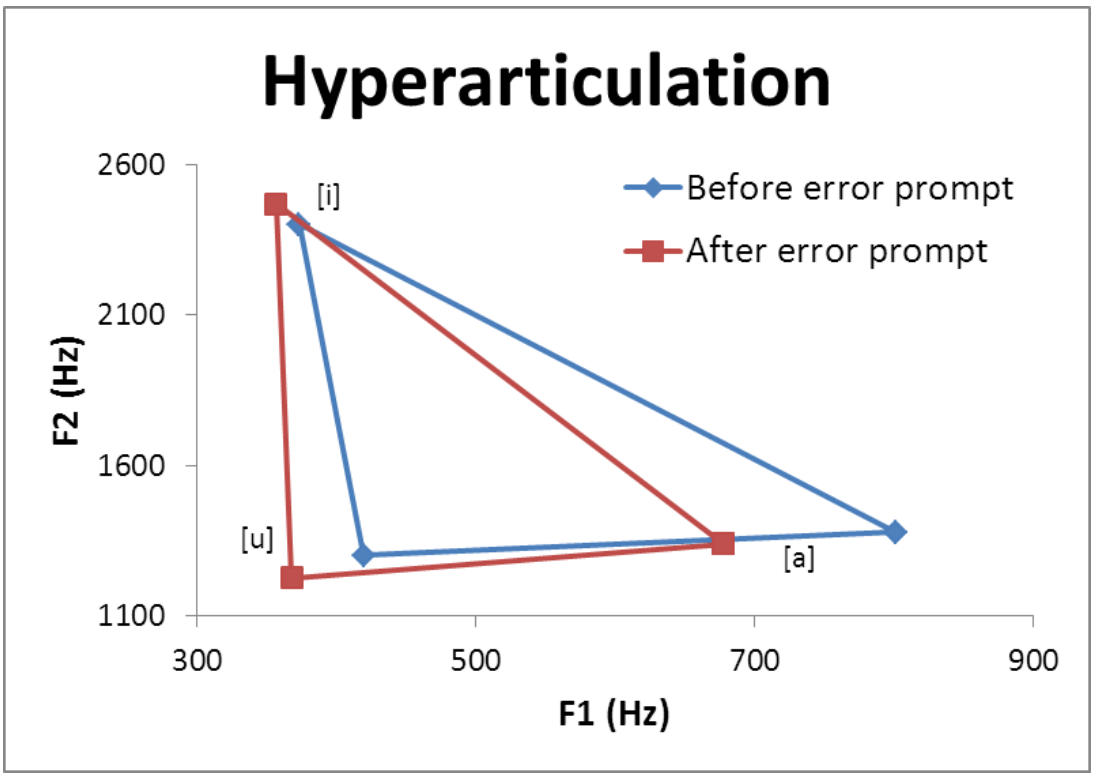

Figure 3: Hyperarticulation analysis

Figure 3 shows the hyperarticulation triangles based on the average first and second formant frequencies of three standard vowels spoken before and after an error prompt in a voice recognition system. The three vowels are [a] as in economies, [i] as in three, [u] as in two. It shows that the [a] sound was less clearly articulated while the [i] and [u] sounds were hyperarticulated slightly when responding to the error prompt. This can be seen because less articulated vowel sounds are closer to the center of the chart, while hyperarticulated vowels are closer to the edges or axes of the charts.

As pointed out in the description, the [i] and [u] vowel sounds were hyperarticulated while the [a] vowel sound was diminished in the response to the error prompt. This is most likely a result of the [i] and [u] sounds being in numbers, which occur much more frequently in the English language than the word "economies"—the word the [a] sound was pulled from. Since the numbers occur more often, the test subjects were probably more certain of the "correct" pronunciation, and emphasized the vowels accordingly. Conversely, since "economies" is a less familiar word to most, and because it was the final word in the test statement, it is likely that the test subjects glazed over their pronunciation of "economies" producing a less clear [a] sound after the error prompt.

## 10. Obstacles

Few obstacles were encountered by this IPRO team over the semester. At the beginning of the semester, the team entered the IPRO not knowing much about how words are pronounced and how the ear translates sounds into recognizable speech. We spent the first three weeks of class just learning how each syllable sounds and how it is pronounced. We spent a brief period of

time also learning how sounds are pronounced in other languages. Another challenge encountered was that the team had to learn how to use the Praat software in order to properly manipulate various sound files.

## 11. Acknowledgements

The IPRO 316 team would like to thank the test subjects who were willing to spend their time participating in our study. The team's instructor, Prof. Matthew Bauer, also deserves a special mention for imparting his expertise and giving advice on the team's project. Furthermore, the team also wishes to thank the Miller Pizza Company for providing the pizzas for the study's subjects and also acknowledge the security guards of the building for giving access to the building and keeping it secure.

## 12. References

1. Barra-Chicote, R., Yamagishi, J., King, S., Montero, J. M., Macias-Guarasa, J. "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech." Speech Communication 52 (2010): 394-404.

2. Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C. "Automatic speech recognition and speech variability: A review." Speech Communication 49 (2007): 763-786.

3. Bradlow, A. R., Torretta, G. M., Pisoni, D. B. "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics." Speech Communication 20 (1996): 255-272.

4. Jones, C., Berry, L., Stevens, C. "Synthesized speech intelligibility and persuasion: Speech rate and non-native listeners." Computer Speech and Language 21 (2007): 641-651.

5. Palaz, H., Bicil, Y., Kanak, A., Dogan, M. U. "New Turkish intelligibility test for assessing speech communication systems." Speech Communication 47 (2005): 411-423.

6. Skantze, G. "Exploring human error recovery strategies: Implications for spoken dialogue systems." Speech Communication 45 (2005): 325-341.

7. Stevens, C., Lees, N., Vonwiller, J., Burnham, D. "On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference." Computer Speech and Language 19 (2005): 129-146.

8. Van Wijngaarden, S. J., Verhave, J. A. "Prediction of speech intelligibility for public address systems in traffic tunnels" Applied Acoustics 67 (2006): 306-323.