Spring 5-5-2010

Effect of Visual Input on Vowel Production in English Speakers

Amanda C. Richardson
Macalester College, richardson.amandac@gmail.com

Effect of Visual Input on Vowel Production in English Speakers

Amanda Richardson
Advised by Christina Esposito & Eric Wiertelak

Macalester College
2010

Abstract

This study analyzes whether there should be a visual component to a model of speech perception and production by comparing the jaw opening, advancement, and rounding of American English and non-English vowels in the presence and absence of a visual stimulus. Surprisingly, jaw opening did not change production, but the presence of the visual stimulus was found to be a significant factor in participants' vowel advancement for non-English vowels. This may be explained by lip rounding, but requires further research in order to develop a full understanding of the impact of visual input on vowel production to be used in teaching and learning languages.

Introduction

Role of Vision in Speech Perception

From neural signals that coordinate our movement to conversations held with family members and peers, the ability to receive and project information is essential to successful functioning in a world of communication. Language is universally learned, but it is difficult to understand the nuances of how this learning takes place. There are aspects, in addition to listening new words, which give us new information about how sounds are formed in spoken language (i.e. the way that the mouth moves). The word *bait*, for example, is visually distinguishable from the word *Kate* because producing a [b] requires that the lips come into contact with one another. The placement of the lips during the first sound in *Kate*, [k], does not require that contact. This inherent property of our speech mechanism would suggest that one element that may play a role in our ability to learn language is the visual input that we receive as we listen to the production of sounds.

The goal of this study is to see if participants are better able to produce vowel sounds when they are provided with a visual stimulus along with the auditory production of the sound. The most well-known studies in this field look at perception of consonants, while this study investigates how the perception then goes on to affect the production of speech sounds. It is also important to keep in mind that most of the following studies focus on consonant sounds because of their salient visual features. Because less work has been done on vowel perception, this study draws upon theories that originated from studies focusing on consonants and explores if and how they might apply to vowels.

McGurk and MacDonald (1976) performed a study that examined the way that visual input affects human perception of speech sounds. They presented visual clips of a

person pronouncing one sound simultaneously with audio clips of a different, but similar, sound and asked participants to report what they had heard. For example, the participant may have heard the syllable /ka/, but watched a visual of a person saying /ba/. The researchers found that a significant number of participants made an error in their reports, saying that they had heard /ba/ when they had been given /ka/ as the auditory input. This showed that visual information could influence what the participant reported hearing.

This phenomenon became known as the McGurk effect and inspired a wave of research that examined different variables in the presentation of the stimuli. One study manipulated the time and speed at which the visual stimuli were presented (Munhall, Gribble, Sacco, & Ward 1996). The experiments examined the effect of delaying the audio and playing both the visual and audio inputs at different speeds. The McGurk effect was stronger when both types of stimuli were played at the same speed and time, but it was still present in the unmatched trials, which suggested that while participants were sensitive to the timing difference, the concordance of the two types of information was not needed to produce the McGurk effect.

Recently, a study extended the research to include somatosensation (Gick & Derrick, 2009). Participants that listened to the syllable /ta/ were more likely to think that they were hearing /pa/ if they felt a puff of air against their skin as the word was played. Gick & Derrick reasoned that this is because /p/ is aspirated at the beginning of a word, meaning that there is a burst of air released when it is pronounced. The puff of air simulates the aspiration and affects the perception, just as the visual input changed the perception in McGurk and MacDonald's original study.

Another study measured cortical activity during the perception and production of syllables (Skipper, van Wassenhove, Nusbaum, & Small 2007). This study showed activity in the frontal cortex during the perception of the sound, indicating that as people watch someone speak, they are also processing the movements that would be required to produce the sounds. While producing the sound, the cortical mappings for each syllable /ta/, /pa/, and /ka/, was distinct, but when perceiving the sound, there was much more overlap. This is contributing evidence to the McGurk effect and some insight into how the brain is processing stimuli during perception and production of sound.

The idea of observational learning is one that has received a lot of attention. A type of neurons, called mirror neurons, were discovered in primates (di Pellegrino 1992). di Pellegrino was mapping sensory-motor areas in rhesus monkeys and noticed activity not only when they were performing the act of picking up a peanut, but also when the monkeys saw people perform that same action. Rizzolatti & Arbib (1998) proposed that mirror neuron systems were responsible for evolution of language from gestural communication to modern speech. Rizzolatti & Craighero (2004) point out that modern speech is seen as arbitrary because the phono-articulatory actions that we use to make words are unrelated to the meaning. Mirror neurons in the auditory and audio-visual system may allow for the imitation of phono-articulatory movements independent of semantic meaning (Rizzolatti & Craighero 2004).

Kuniko Yasu Nielsen (2004) did a study in which she tested fifteen English consonants to determine the intelligibility of individual speech sounds. Audio and video clips were presented simultaneously, while the quality of the audio clip varied to the point of being generally incomprehensible. Participants were given a forced choice task

to report what they had heard. Nielsen found that the presence of the visual cues improved intelligibility, especially in the case of inter-dental sounds, when the tongue is placed between the teeth, and labio-dental sounds, when there is contact between the lips and teeth. This finding was consistent with the phonetics of the English language, considering that those sounds are made with the teeth and lips, parts of the mouth that are visible to the listener. Taken together, the results of these studies suggest that there is something at work during speech perception and production, beyond the auditory input that we receive.

Language Models

With the knowledge that there is a connection between auditory perception and other sensations, we want to understand the underlying brain structures that are responsible for connecting these sensory modalities. We link them by creating language models that shows the flow of information from stimulus energy in our environment to comprehension and/or output. One of the first researchers to look into models of speech perception and production was Alvin Liberman. He developed *motor theory*, in which he discussed phonetic *gestures*. He defined gestures as the way that the vocal tract constricts in order to produce consonant sounds and considered gestures invariant aspects of what the listener observes (Liberman et al. 1967). This was later revised to say that the listener perceives the *intended* gesture, as Liberman conceded that actual gestures certainly do vary (Liberman 1985). Liberman suggested that the way in which we process these intended gestures in both their perception and production is an innate module.

The question then became what this innate module looked like. Norman Geschwind (1970) addressed this question by expanding upon the research done by Paul Broca and Karl Wernicke in the mid-1800's. In a post-mortem autopsy of an aphasic patient who had difficulty creating meaningful strings of words, Broca discovered a lesion anterior to the motor strip, above the lateral fissure in the left hemisphere (Broca 1861). This is now commonly referred to as Broca's area. Due to the disjointed speech of people with damage to this area, this part of the brain is associated with speech production and fluidity of speech. According to Garrett (2009), in 1870, Wernicke found an area in the left temporal lobe that, when damaged, interfered with the ability to comprehend speech, linking this area, often called Wernicke's area, to speech comprehension.

Geschwind (1979) combined the discoveries of Broca and Wernicke as well as modern ideas of the brain and developed the Wernicke-Geschwind Model of Language. The model specified the neural pathways for language perception and production. If someone were asked a question, for example, information would first be processed by the primary auditory cortex, followed by Wernicke's area. Once the question had been comprehended, speech could be produced by relaying a signal to Broca's area, and eventually the motor cortex, where the movement for sound production would be generated. A written question, on the other hand, would be processed by the occipital lobe and then translated by the angular gyrus, located in the parietal lobe, between the occipital lobe and Wernicke's area. The information would continue along this pathway to Wernicke's area and again follow to Broca's area and the motor cortex (Garrett, 2009). The Wernicke-Geschwind Model involves the occipital lobe insofar as vision is

necessary for reading text, but spoken word is always referred to as being processed by the primary auditory cortex.

Frank H. Guenther is currently working on perfecting a computerized speech model, called DIVA, (Directions Into Velocities of Articulators) (Guenther 1994). This model is a neural network that, unlike many speech synthesis programs, has aspects that correspond to specific brain structures. The system learns a sound when a speech sound map cell is activated, triggering the motor commands that direct the system to attempt production of the sound. The system's production of the sound causes two subsystems to begin to operate, a feedforward and feedback, as pictured below. At first, the auditory feedback control will be the primary system for learning the sound, but with each attempt, the feedforward control is updated and will soon be the sole command for sound production (Guenther 2006).

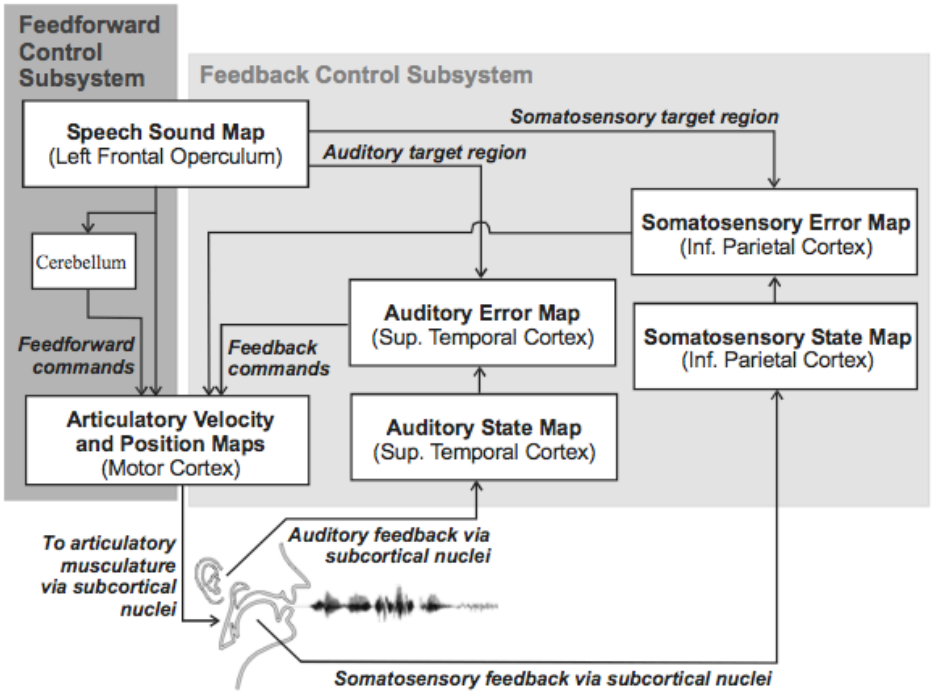

*Figure 1***.** A simplified version of the DIVA model of speech acquisition and production. adapted from "Cortical interactions underlying the production of speech sounds," by F. H. Guenther, 2006. *Journal of Communication Disorders*, **39**, 350-365. Reprinted with permission.

Guenther includes the motor cortex in both the feedforward and the feedback control loops. It is responsible for creating the sound from the given input. The temporal cortex is involved in the auditory feedback loop. Once the system has produced the sound it is able to take the resulting sound and compare it to an auditory state map, note any errors in the production and adjust motor movements based on those errors. The parietal cortex is responsible for somatosensory feedback in his model. The system stores the sensations that are present during production and uses them to inform future productions. We notice that the occipital lobe does not play a role in Guenther's model of spoken language. Additionally, he points out that there are multiple motor movements that can

produce the same speech sound, including speaking with a clenched jaw (Guenther, 1995).

While we are able to understand a person who speaks with a clenched jaw, there are still differences in the quality of the sounds that result from moving the mouth in a different manner. The way that we shape our vocal tract affects the resulting speech sound. As air is pushed from the lungs up through the larynx, placement of the uvula directs air flow out of the oral and/or nasal cavity. Other mobile structures in mouth, like the tongue and lips adjust the shape of the mouth to produce different qualities of sound. As the air moves through a narrower or wider space, the sound varies. Even the length of the vocal tract affects ability to produce vowel sounds, as is seen in a primate's inability to make some human vowel sounds due to the difference in the shape of their vocal tract (Liberman et al. 1969).

So, how do we decide what shape to create with our mouths when producing spoken language? When we are unfamiliar with a sound, auditory input may not suffice. Our ability to produce a novel sound may be influenced by visual cues we receive from the speaker. I would argue that recognizing the formation of certain sounds may use the same pathway as that of object recognition in the brain. This pathway is known as the ventral stream, or the "what pathway," as it is how we determine "what" we are seeing in our visual field. The ventral stream begins at the primary visual striate cortex, V1, which feeds information forward to the prestriate cortex, V2, creating a visual coherency from the information. From there, information travels to V4, the extrastriate cortex, which plays a role in attention modulation, finding salient features in visual information. The next structure in the ventral stream is the medial temporal lobe, located on the inner side

of the temporal lobes, where processing of auditory stimulus takes place. The medial temporal lobes also include the hippocampus, which is important in creating memory (Goodale & Milner 1992). These ties between the ventral stream and audition and memory may allow those with visual access to speech information to better perceive and produce sounds.

This purpose of this study was to investigate the role of visual input on speech production by presenting nonsense syllables to participants in two groups. The members of the control group received only auditory input of the sound and were asked to reproduce it to the best of their ability. The members of the experimental group watched a video of the speaker producing the syllables, receiving both visual and auditory input. The acoustic properties of all productions were then measured and compared to the original stimulus, or *target production* values to determine how the production of the participants differed from that of the target. Analyses were done to answer the questions:

(1) How well were people with visual cues able to match the target production of vowel sounds in comparison with those who did not receive visual cues?

(2) Was there a correlation between confidence ratings and whether or not a vowel was English?

(3) Was there a difference in ability of either the visual or non-visual group in reproducing English versus non-English vowels?

(4) Were there specific vowels that were more difficult for participants to reproduce?

Methods

Participants

Eighteen students took part in this study. All spoke American English as their first language and had normal hearing and normal or corrected-to-normal vision. None of them had formal training in the International Phonetic Alphabet (IPA), which means that they had not been taught how to recognize and produce English and non-English sounds in the study. Formal training would have made them familiar with non-English sounds, losing the novelty effect, which would affect the results. The ages of the participants ranged from eighteen to twenty-three years old. Ten identified as male and eight identified as female. All participants were recruited from the current or former Macalester College student body, keeping age, and sex in consideration. Linguistics majors were excluded from the study and participants were asked to provide information on what languages they had studied to account for familiarity with non-English sounds.

Stimuli

The stimuli consisted of two sets of sixteen nonsense words, which were monosyllabic and followed the pattern CVC (consonant, vowel, consonant). One set contained six American English vowels (/ɪ/, /eɪ/, /ɛ/, /æ/, /oʊ/, /ɑ/) and two central vowels, (/ɨ/ and /ʉ/), which are considered part of the American English dialect in this experiment. The vowels were placed between two alveolar consonants, namely [t] and [s], and [z] and [d], chosen for minimal movement of the jaw and lips and for uniform place of articulation in a fashion that would result in all syllables being nonsensical in American English. Alveolar sounds are those that are produced by placing the tongue on

or near the alveolar ridge, which is located behind the front teeth of the top jaw. The other eight stimuli consisted of eight non-English vowels (/y/, /ɯ/, /ʏ/, /ø/, /œ/, /ɶ/, /ɣ/, /ɒ/) between the same two sets of consonants. The non-English vowels were chosen as the rounded or unrounded counterparts to the eight English vowels. The difference in roundness is predicted to be a salient visual cue and so is sure to be included in the stimuli. In the first set, both consonants were voiceless ([t],[s]). The initial consonant was an oral stop ([t]), that is, a consonant produced by the constriction and release of airflow in the mouth, while the final consonant was a fricative ([s]), for which there is only partial constriction of airflow. The second set used voiced consonants, where the fricative ([z]) was word initial and the final consonant was an unreleased stop ([d̚]), meaning that the mouth is in position to produce the sound, but the air is not released. An unpaired t-test run comparing the words in the [t_s] formation versus the [z_ d̚] formation did not show significance, assuring us that the consonants did not affect the production of the vowel. The words were chosen in pairs in which the height and advancement of the vowel was the same, but the rounding was different, with a few minimal exceptions. The American English /æ/ is slightly more closed than the non-English /ɶ/, the English vowels /oʊ/ and /eɪ/ are diphthongs, and centralized, high vowels (/ɨ/and /ʉ/) were used. Below is a table of all thirty-two stimuli used in the experiment.

Set 1		**Set 2**	
American English Vowels	*Vowels not in Am. English*	*American English Vowels*	*Vowels not in Am. English*
tɨs	tys	zɨd̄	zyd̄
tʉs	tɯs	zʉd̄	zɯd̄
tɪs	tʏs	zɪd̄	zʏd̄
teɪs	tøs	zeɪd̄	zød̄
tɛs	tœs	zɛd̄	zœd̄
tæs	tɶs	zæd̄	zɶd̄
toʊs	tɤ̞s	zoʊd̄	zɤ̞d̄
tɑs	tɒs	zɑd̄	zɒd̄

Table 1. Stimuli used for the experiment were in two sets of sixteen words, all monosyllabic, nonsensical words, eight using American English vowels, and eight using vowels that are not included in the American English dialect.

The stimuli were recorded in the Linguistics Laboratory at Macalester College. The speaker, a male with no glasses and minimal facial hair, pictured below, spoke American English natively and had training in the IPA. He sat in front of a blue background approximately one meter in front of the video camera (Canon 2R70MC Digital Video Recorder). After reviewing each of the stimuli he was asked to pronounce each one three times upon cue from the experimenter. Both a visual and auditory recording was made of the pronunciations.

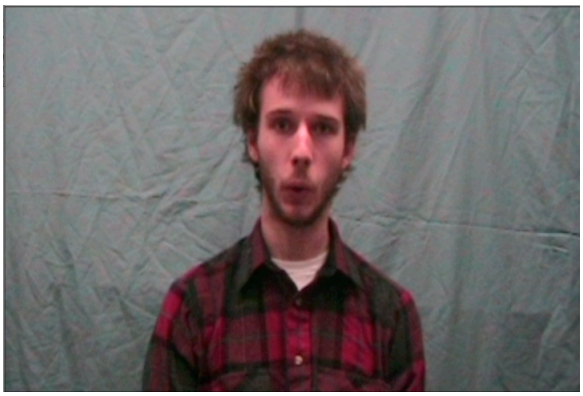

Figure 2. Sample image of the visual stimulus producing the vowel, /y/.

The recordings were then captured onto a computer using *Final Cut Studio*. Each word was edited to a one second clip of a single pronunciation (the best take of the three as determined by the experimenter). All video and sound manipulation was done at the Humanities Resource Center at Macalester College. The visual stimuli were presented on a 13.3-inch Mac Book and the audio at full volume.

Procedure

All of the participants were run in individual sessions, seated in front of a blue screen in the Macalester Linguistics Laboratory, one meter in front of a camera in the same orientation used for recording the speaker.

At the start of the procedure, participants were given four practice trials, two with American English vowels and two with vowels not present in the American English dialect. They were told that not all vowel sounds in the stimuli would be those of American English and asked to reproduce the sound with as much accuracy as possible. The experimenter instructed the participant to turn toward the screen before each stimulus was administrated.

Participants were randomly assigned to one of two groups. Ten participants received a visual stimulus paired with the auditory stimulus. The computer was set up half a meter to the right of the participant, but not in view of the video camera. Participants were asked to turn their heads to watch and listen to the speaker once on the screen and then orient their head toward the video camera and reproduce the word three times. They then verbally rated how confident they were in their production on a scale from 1 to 5, where 1 was no confidence in the production and 5 was complete confidence

in the production. The eight participants in the control group had the stimuli presented with the computer facing away from them and connected to speakers to assimilate the volume of the visual trials. After hearing the sound, they were asked to turn toward the camera and produce the sound three times. The thirty-two stimuli were presented in a randomized order that remained constant across trials.

Audio from the sessions was digitized into .wav files using *Roxio Easy VHS to DVD*. Analysis of the files was done using *Praat Version 5.0.20* (Boersa & Weenik 2009), software that is used to measure the phonetic properties of speech.

Two measurements were recorded, the first formant, F1, and the second formant, F2. Formants are the frequencies of sound waves that make up different vowel sounds and allow us to distinguish one sound from another. F1 is correlated with jaw opening, so the higher F1, the more open the jaw. F2 is correlated with advancement, so a higher F2 means that the tongue is further forward in the mouth. An [i], for example, would usually have a low F1 of approximately 280 Hz because the jaw is closed, but a high F2 value of approximately 2250 Hz because the tongue is forward. An [ɑ] on the other hand, is much more open, with an F1 value around 710 Hz and the tongue is father back, resulting in a lower F2 value than an [i] (close to 1100 Hz). Figure 3, below, provides a visual representation of how vowel formants are related to vowels in the IPA.

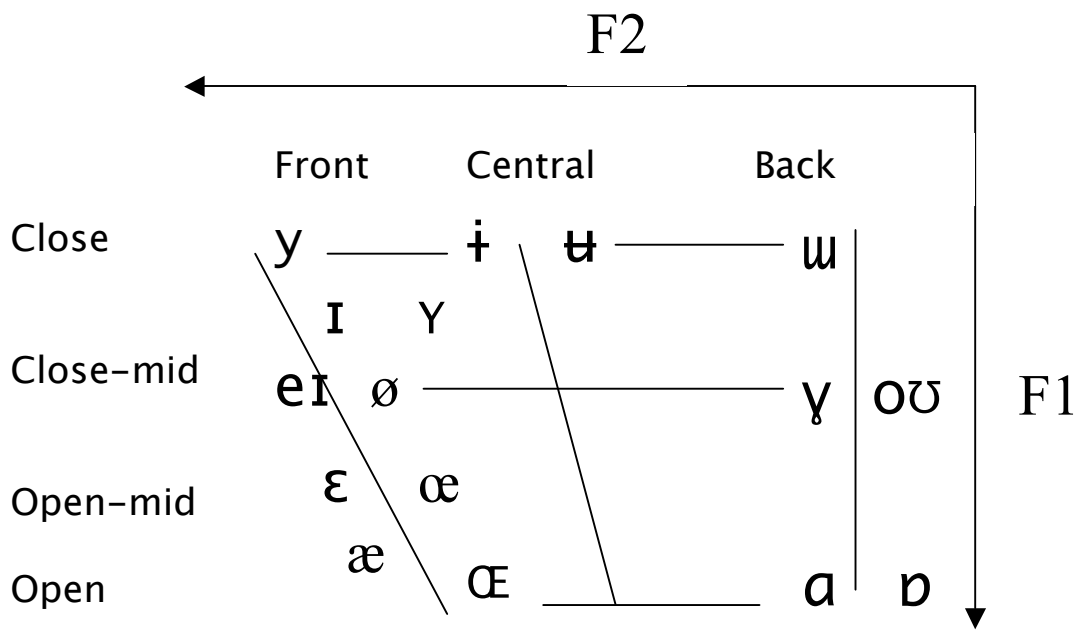

*Figure 3***.** Vowels used in the experiment and their relationship to formant values. The first formant, F1, should be higher for vowels like /ɑ/, /ɒ/, /ɶ/, and /æ/, which are more open vowels. Closed vowels (at the top of the chart) like /y/, /ɨ/,/ʉ/, and /ɯ/, will have lower F1 values. Front vowels, on the left side of the chart, will have high F2 values, while back vowels, on the right side of the chart, will have low F2 values. Reference the Appendix for further understanding of how this chart is constructed.

Average values of the three productions were used, although analyses were done to check for variance over multiple productions to control for somatosensory learning that may be occurring during repetition of new sounds (Guenther 2006). This same con

Statistical analysis was done in *Excel* and *R,* a programming language that provides tools for statistical modeling and graphics. The data used in *R* consisted of the stimulus number (1-32), listed twice to account for control and experimental conditions. The variables considered for each stimulus were *Visual*, a binary variable; either the participant received the visual stimulus or did not , *Eng*, also binary, either the vowel was English, or non-English, and *Conf*, a rating of confidence of production from 1-5, and the two vowel quality readings, *F1* and *F2*.

Results : Visual vs. Non-Visual

Looking at the figures below, there appear to be patterns in both F1 and F2. For F1 (Figure 4), participants in both of the visual and non-visual condition had significantly higher F1 values than the target value originally recorded with the speaker. F2 (Figure 5), however, shows a significantly higher value only for the non-visual group.

In order to first test the statistical significance of these differences in vowel formants, unpaired t-tests were performed. For F1, the difference between the target values and the control group was significant ($p=0.0076$), as well as the difference between the original and the experimental group ($p=0.026$). The target values and the control were also significantly different for F2 values ($p=0.0030$). The difference between the target and the experimental condition was not significant. A t-test run between the average confidence ratings of both groups did not show significance.

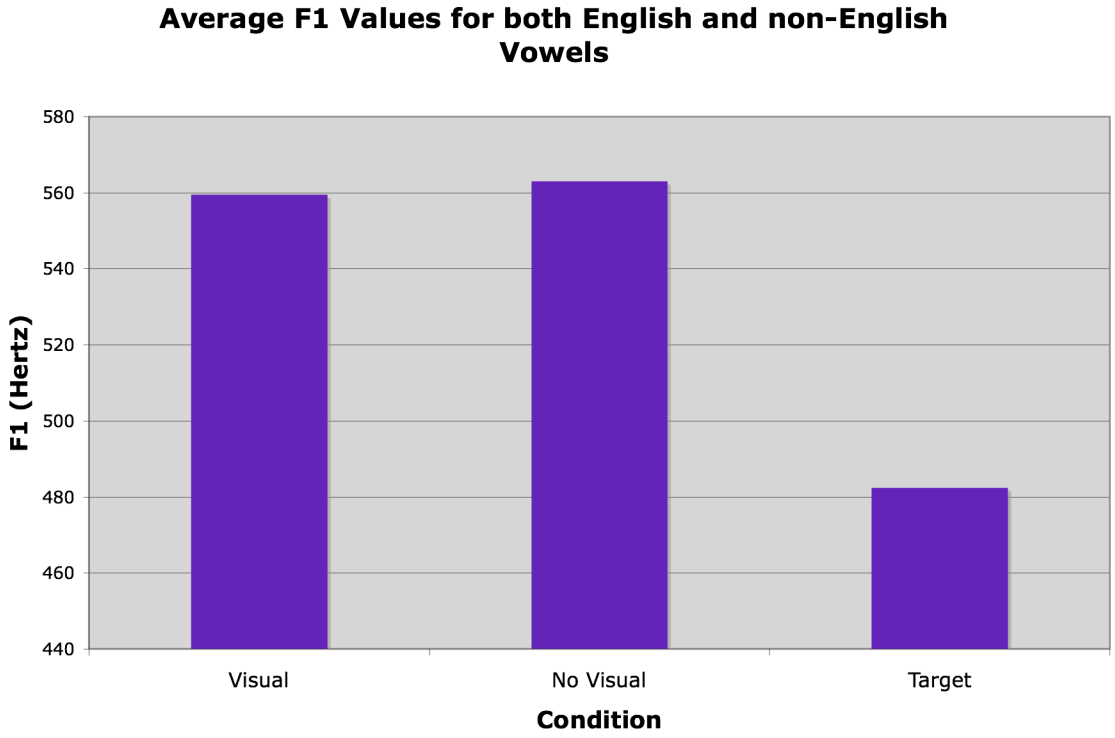

*Figure 4***.** Average F1 values plotted for both the experimental and the control group as well as the target values presented in the stimulus. Both the experimental (p=0.0076) and control (p=0.026) groups are significantly different from the stimulus, but not from each other. The mean experimental value was 559.59 Hz, the mean control value was 563.07 Hz, and the mean of the target stimulus was 482.40 Hz.

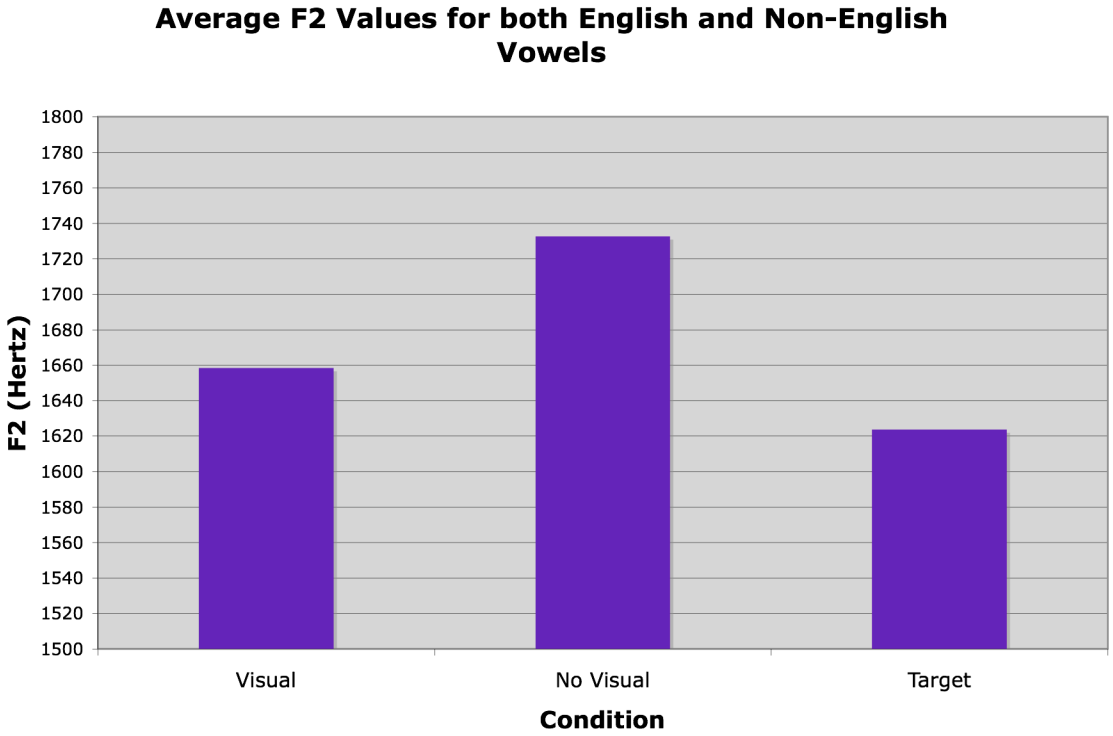

Figure 5. Average F2 values plotted for the visual and non-visual groups, as well as the value for the target stimulus. Both the visual and non-visual groups have higher F2 values when compared to the target values, but only the control is significantly different ($p=0.0030$), making the visual and non-visual also significantly different from one another ($p=0.014$). Mean values for the visual, non-visual and target values were 1658.50 Hz, 1732.67Hz, and 1638.66Hz, respectively.

A univariate model of F1 by Visual confirmed that whether or not the participant received the visual stimuli was not a good predictor of the F1 value produced by the participant. The univariate model of F2 by Visual confirmed that the presence of the visual stimuli was significant in predicting F2 values ($p=0.035$). Also significant in univariate models of F2 were Eng ($p=0.0081$) and Conf ($p=0.0104$). The estimated mean increase for an English vowel was 150.67 Hz for every increase in one confidence unit,

while the estimated mean increase for non-English vowels was 98.68 Hz. When both variables were present in the model, neither was significant. The correlation coefficient for Eng and Conf was 0.70.

Model for F2 Production (F2~Visual+Eng+Conf)				
Variable	Estimate	Std. Error	Test Stat	P-value
Intercept	1467.87	173.51	8.46	8.09e-12
Visual	-92.48	55.74	-1.48	**0.035**
Eng	93.88	78.58	1.20	0.24
Conf	52.37	51.54	1.02	0.31

*Table 2***.** The estimate, standard errors, test statistics, and p-values for all of the variables included in the model for F2. Significant p-values are in boldface font. Though non-significant, Eng and Conf were included, because of their significance in univariate models.

The results in Figure 6 suggest that participants were more confident in their productions when asked to reproduce an English sound, as opposed to a non-English sound. The average confidence rating for English vowels was significantly ($p=6.01*10^{-11}$) higher overall, 4.25 ± 0.59, in comparison to the average for foreign vowels, 3.16 ± 0.51.

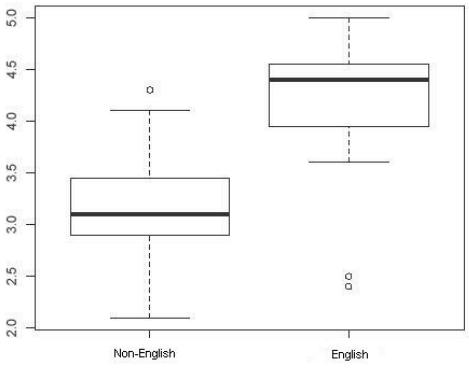

*Figure 6***.** Boxplot of confidence ratings for visual and non-visual groups. The dark line is the mean, with the edges of the box at the 25th and 75th percentiles and whiskers extending to 1.5 interquartile ranges from the mean. Circles represent outliers. Participants were significantly more confident about their production of English sounds than they were of non-English sounds.

Results: English vs. Non-English

In order to discuss the question of whether the English-status of the word affected production of the vowel, the data for English words was separated from non-English words for analysis. In comparing participants' formants with the target values, we saw the same results for non-English words as we saw for the overall data. Both conditions were higher than the target values for F1 (see Figure 7) and only the control was higher than the target for F2 (see Figure 8). In the English data, however, all of the F1 values were similar. The F2 values for both conditions appeared higher than the value for the target.

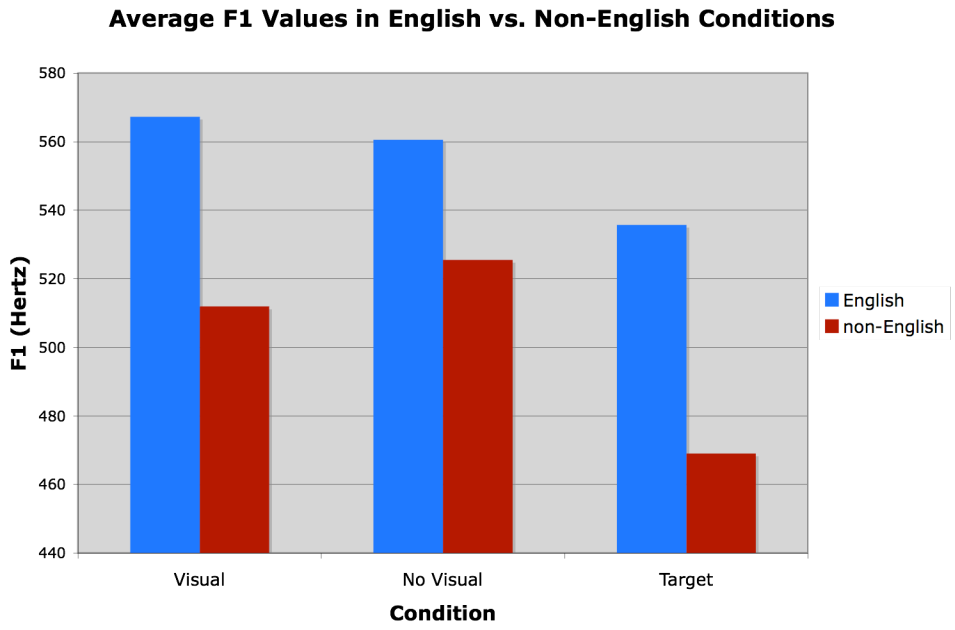

Figure 7. The F1 values in English stimuli are not statistically significant across any of the conditions. The relationship across conditions for the non-English stimuli mimic that of the combined data. The visual condition is significantly different from the target ($p=0.002$), as is the non-visual condition ($p=0.003$). The visual and non-visual conditions are not significantly different from each other.

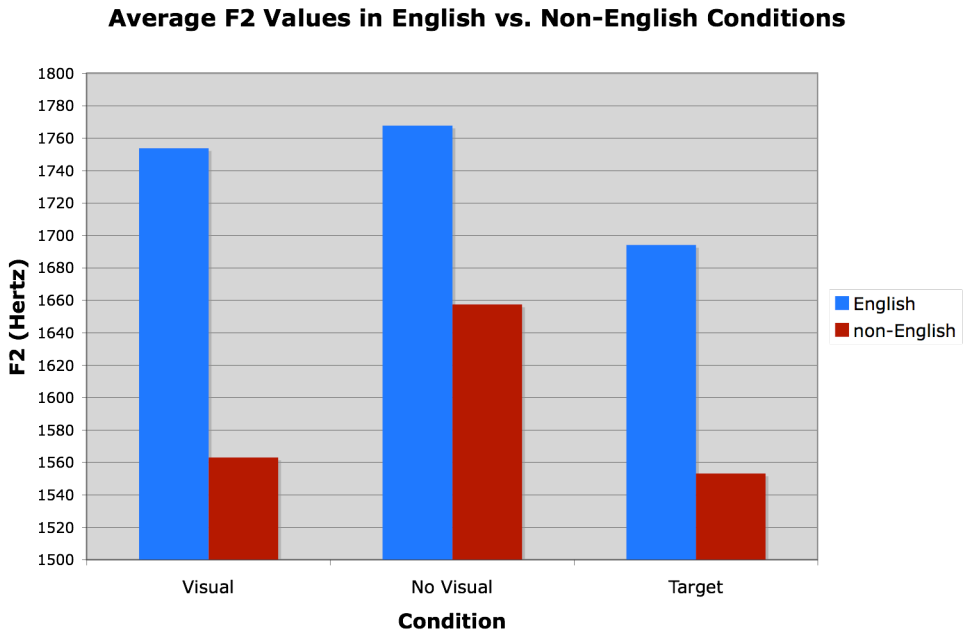

Figure 8. The F2 values in English stimuli are significantly higher than the target in both the visual (p=0.03) and non-visual(p=0.05) condition. Non-visual vowels again behave like the combined data, showing that the non-visual condition has significantly higher F2 values compared to the target and the visual condition.

To test for statistical significance, paired t-tests were run on the newly formatted data. For F1 of the non-English condition, the visual group was significantly higher than the stimulus (mean=449.03 Hz), with a mean of 531.92 Hz (p=0.0016) and the non-visual group was also significantly higher, with a mean of 545.53 Hz (p=0.0037). There was not a significant difference between the visual and non-visual groups. For the F2 value, the mean of the stimulus condition was 1553.17 Hz. The visual condition was not significantly higher than that value, with a mean value of 1543.05. The non-visual condition was significantly higher with a mean of 1677.446 Hz (p=0.033). The visual and non-visual groups were significantly different from each other (p=0.017).

None of the F1 values were significantly different from one another in the English condition, with averages of 567.27 Hz, 560.62 Hz, and 535.78 Hz for the experimental, control, and stimulus means, respectively. In the case of F2, the mean stimulus value was 1674.15 Hz. The visual mean was significantly higher than that at 1773.95 Hz (p=0.026). The non-visual mean was also significantly higher at 1787.89 Hz. The difference between the experimental and control groups was not statistically significant.

Individual vowels were also analyzed for accuracy. A vowel was said to differ from the target pronunciation if at least three people in the group (visual or non-visual) had significantly different values. There were no significant sounds that showed a difference in F1, however, across both groups, the most people differed from the target pronunciation in the vowels /ɯ/ and /œ/. F2 comparisons showed different pronunciations from the target in both groups for /ʉ/,/ɪ/,/ɛ/,/æ/, /ɑ/ and /y/. /ɨ/ was also significantly different for the visual group only.

Discussion

As previously mentioned, the first formant, F1, is correlated with jaw opening. The higher the F1 value, the more open the jaw of the speaker is. The second formant, F2 is correlated with the advancement of the vowel, that is, whether the tongue is placed at the front or toward the back of the mouth during production. The higher the F2 vowel, the farther forward the tongue is.

With this information, we look at the participants ability to reproduce the vowel sounds. In terms of F1, both the non-visual and visual groups were significantly different from the target with which they were presented. They were not significantly different

from each other, both having means higher than the target value. Higher F1 values indicate that participants actually tended to hyperarticulate their pronunciations by opening their mouths wider than the person who pronounced the sounds in the original speaker. A possible explanation for this unexpected result could be that in pronouncing unfamiliar sounds, participants were unsure of themselves and tried to overcompensate by making larger gestures, therefore hyperarticulating the sounds. This is supported by the fact that we do not see the same increased F1 when the data is split into English and non-English stimuli. There were no significant differences between the values in the English data. Since the English vowels were familiar, the participants were less inclined to hyperarticulate.

From the high correlation between English sounds and confidence ratings (see Figure 6), we first infer that the higher confidence ratings mean that participants were able to identify an English sound versus a non-English sound in order to reproduce the vowel. Secondly, we can see that when the stimulus was English, the participants were more comfortable and more confident, and were able to more accurately match the sound that they heard. In the case of non-English words, when the participants were less confident and less familiar, we saw the same effect as when we looked at the overall data, which showed more open jaws for both the visual and the non-visual groups. Therefore, because both groups performed in the same manner in this domain, the data show that the ability to replicate jaw opening has to do with familiarity and not to do with access to visual information.

The data on individual vowel production indicate that the confidence rating may even hinder the ability to accurately reproduce familiar vowel sounds in terms of

advancement. All of the sounds that were statistically different, with the exception of /y/, were English vowel sounds. Of course, each vowel differed from the vowel space of the speaker. When each production of each participant was compared to that of the speaker using t-tests, those that had the greatest number of statistically significant deviations from the target values were those whose vowel spaces most differed from the speaker in the stimulus audio (See Appendix). When they heard a sound that they use on a regular basis, they simply reverted to their own production, instead of focusing on replicating the sound that they heard and/or saw.

The overall data for F2 shows that the visual group was able to more accurately match the original F2. That is, they produced the sounds with a similar amount of fronting. Without the visual, the control group had higher F2 values, meaning that their tongue was farther forward in their mouths for these productions. Again, these results were replicated in the non-English condition, while in the English category, there was fronting in both groups. It appears that the presence of the visual stimulus helped participants to more accurately imitate advancement in non-English syllables, even though it is not a quality that can be observed in the video. The presence of fronting could be the result of the vowel being presented between two alveolar consonants, so that the vowel would be fronted for ease of articulation, but we see fronting even in the case of vowels that are already produced in the front of the mouth. In addition, this does not explain why fronting did not occur for people who had visual input, as advancement is a vowel quality that is not outwardly visible.

The explanation could be in an additional vowel quality, lip rounding. Lip rounding does affect F2 and is important as the most obvious visible quality. The fact that

/ɯ/, the unrounded counterpart of /u/, and /œ/, the rounded counterpart of /æ/, saw the greatest number of differences is further indication that rounding was a factor. Because /ɯ/ vowel is the unrounded version of the English sound /u/, people were hearing something similar to /u/, but they were not seeing rounded lips or, in the case of the non-visual group, they were hearing a difference and unable to reconfigure their mouths to replicate the sound. The vowel, /œ/ is the rounded version of /æ/, so the same sort of phenomenon could occur. The auditory input is recognizable as open and front, but people are unable to make the adjustment for the change in the lip rounding. We do not, however, see superior performance in the visual group, which we would expect since they were able to see the lip rounding, but we see just as many errors as in the non-visual group. Research into how lip rounding affects advancement of the vowel may help us to understand why the visual group was able to better match the stimulus in this aspect. This seems the most plausible explanation, as measurements of the participants' vowel spaces did not show any overall tendency to have more fronted vowels than the speaker in the stimulus. Measurement error is more likely when measuring the production of female participants due to voice quality, but the random assignment of participants into groups assured that there was not a skewed amount of either sex in either group. The number of multi-lingual people was also evenly distributed, so that their knowledge of foreign vowels did not skew results.

While it is not yet clear what caused the visual group's improved ability at reproducing F2, the fact that there is a significant aspect, albeit in an unexpected place, indicates that, with further research, the occipital lobe could play a part in the DIVA model of language processing. The model made sense in terms of our data, in that

familiar sounds triggered the feedforward loop, while the novel stimuli required a feedback loop of somatosensory and auditory information. It was not uncommon for a participant's vowel quality to change over the course of the three productions. To help build upon this model, future neurological research could be done on activity in the ventral stream during speech perception an production. Functional Magnetic Resonance Imaging (fMRI) would allow us to see if certain visual components are being recognized and identified as speech sounds are produced. Because we saw such a difference between the production or familiar and novel stimuli in this experiment, it would make sense to examine the level of activity in these same conditions. Is there a higher amount of visual processing when we are perceiving a novel sound?

Further linguistic studies would benefit from measuring lip rounding directly, while still considering vowel formants, as F2 is tied to lip rounding. It is also important to consider environmental factors that could be influencing the way that participants produce sounds. Being recorded in front of a camera may elicit emotional arousal, which could be accounted for by testing participants' Galvanic Skin Response. To the best of my knowledge, no research has been done on the relationship between nervousness and jaw opening, which could also be a further area of research. This study saw what could have been hyperarticulation in a potentially stressful situation, but it is possible that in an environment where they are not being alerted to their pronunciation, participants would actually reduce jaw opening. Additionally, an element that could provide insight into this study would be to run a group of listeners who are trained in perception and production of all sounds of the IPA. Even without the training, the participants in this experiment did extremely well in reproducing novel sounds, whether they saw the visual or not.

A consideration to keep in mind while interpreting these results is that it is harder to use visual cues for vowel discernment as opposed to consonants (Summerfield & McGrath 1984). The effect of access to visual information is much more prominent in consonants as seen in Nielsen's (2004) work. Vowels are often noticed as distinguishing one speaker from another, but once we are engaged with an individual speaker, the large differences in mouth movements are for the formation of consonants. It still stands to reason that the visual system is at work for comprehension and production of speech in all areas, but may be more crucial to the distinguishing of consonants.

Most helpful to teachers and language learners would be the exploration of certain qualities in learners that help them to benefit from the additional visual input. Is it helpful for children still in the critical age to have visual input when learning how to produce sounds? Is it helpful for second language learners? What are the benefits of visual input for people with different disabilities? These areas deserve investigation to aid language teachers in their ability to educate and language learners in their efforts to acquire language.

References

Boersma, P., & Weenink, D. (2009): **Praat: doing phonetics by computer (Version5.1.05)** [Computer program]. Retrieved September 15, 2009, from http://www.praat.org/

Broca, P.P. (1861). Perte de la parole; ramolissement chronique et destruction partielle du lobe antérieur gauche de cerveau. *Bulletins de la Société d'anthropologie de Paris*, *2*, 235-238.

di Pellegrino, G., et al., *Understanding motor events: a neurophysiological study.* Exp Brain Res, 1992. ***91****(1*):p. 176-80.

Erber, N. (1975). Audio-Visual Perception of Speech. *Journal of Speech and Hearing Disorders*, *40*, 481-492.

Garrett, B. 2009. Brain and Behavior: An Introduction to Biological Behavior, 2nd edition. Los Angeles: Sage Publishers.

Geschwind, N. (1970). The Organization of Language and the Brain. *Science*, *170(3961),* 940-944.

Geschwind, N. (1972). Language and the Brain. *Scientific American, 226(4),* 76–83.

Geschwind, N. (1979). Specialization of the Human Brain. *Scientific American, 241(3),* 180–199.

Green, K.P., & Kuhl, P.K. The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*, *45,* 34-42.

Goodale, M.A., & Milner, A.D. (1992). Separate pathways for perception and action. *Trends in Neuroscience 15,* 20–25.

Guenther, F.H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, *39,* 350-365.

Guenther, F.H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102,* 594-621.

Guenther, F.H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*, *72,* 43-53.

Gick, B., & Derrick, D. (2009) Aero-tactile integration in speech perception. *Nature, 462,* 502-504.

Ladefoged, P. (1975). A course in phonetics. Orlando: Harcourt Brace. 2nd ed 1982, 3rd ed. 1993, 4th ed. 2001, 5th ed. Boston: Thomson/Wadsworth 2006.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & M. Studdert-Kennedy. (1967). Perception of the speech code. *Psychological Review*, *74,* 431-461.

Liberman, A.M., Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1-36.

Liberman, P.H., Klatt, D.H., & Wilson, W.H. (1969) Vocal Tract Limitations on the Vowel Repetoires of Rhesus Monkey and other non-human Primates. *Science, 164,* 1185-1187.

McGurk, H., & MacDonald, J. (1976). Hearing Lips and seeing speech. *Nature*. *264***,** 746-748.

Munhall, K.G., Gribble, P., Sacco, L., & Ward, M. (1996) Temporal constraints on the McGurk effect. *Perception & Psychophysics.* ***58(3),*** 351-362.

Nielsen, K.Y. (2004). Segmental Differences in the Visual Contribution to Speech Intelligibility. *Proceedings for 8th International Conference on Spoken Language*

Processing, Jeju, Korea.
O'Neill, J.J., (1954). Contribution of the visual components of oral symbols to speech comprehension. *Journal of Speech and Hearing Disorders*, *19*, 429-439.
Skipper, J.I., van Wassenhove, V., Nusbaum, H.C., Small, S.L. (2007). Hearing Lips and Seeing Voices: How Cortical Areas Supporting Speech Production Mediate Audiovisual Speech Perception. *Cerebral Cortex*. *17(10),* 2387-2399.
Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *The Quarterly Journal of Experimental Psychology*, *36(1),* 51-74.
Raymaekers, R., J. R. Wiersema, et al. (2009). "EEG study of the mirror neuron system in children with high functioning autism." *Brain Res 1304*: 113-121.
Rizzolatti, G. and Arbib, M. A. (1998) Language within our grasp. *Trends in Neurosciences*, 21(5):188--194.
Rizzolatti, G., & Craighero, L. (2004). The Mirror-Neuron System. *Annual Review of Neuroscience, 27*, 169-192.
Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., & Jones, C.J., (1977). Effect of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, *20*, 130-145.

Appendix

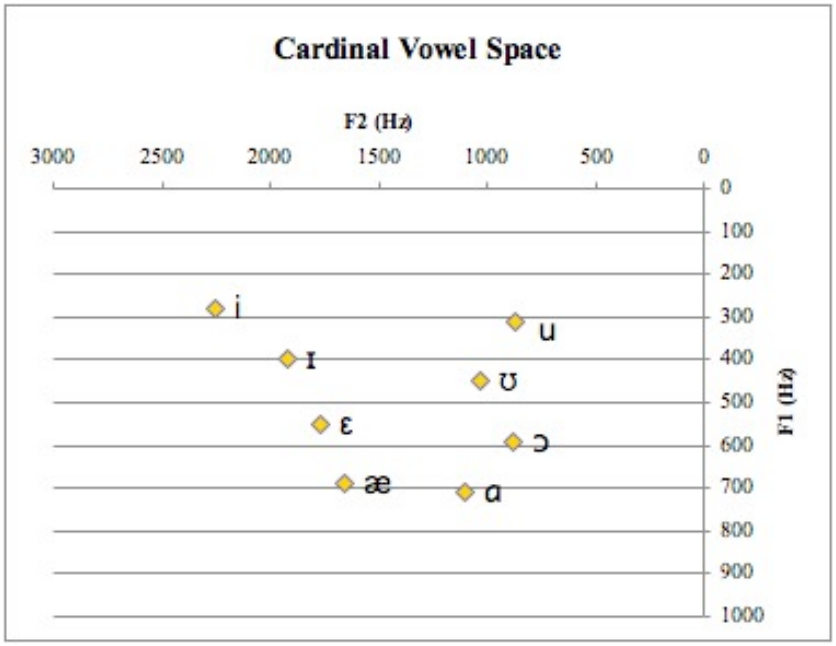

Figure A1: The typical vowel space for cardinal vowels. The first formant, F1 runs along the y-axis and the second formant, F2 runs along the x-axis.

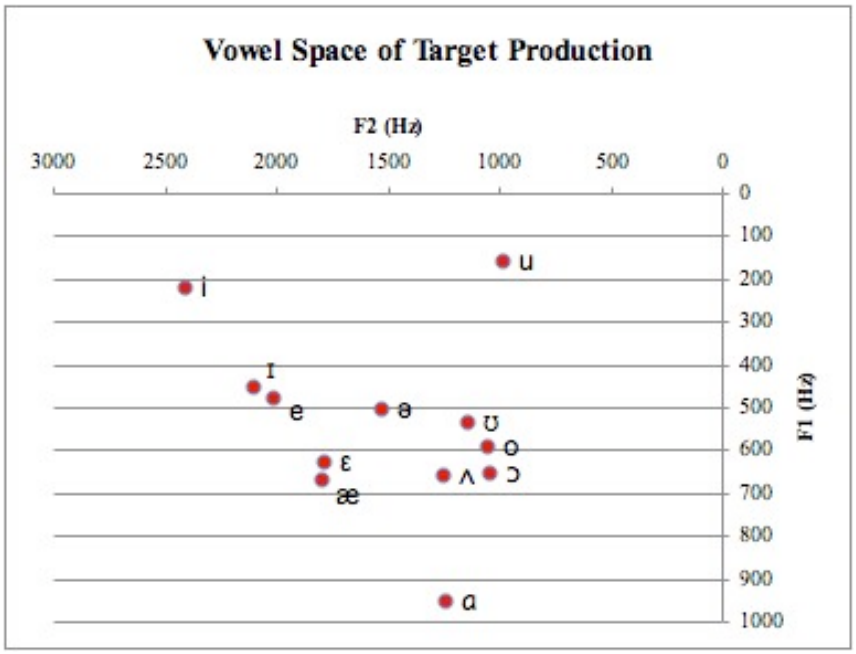

Figure A2: The English vowel space of the speaker in the production of target sounds.

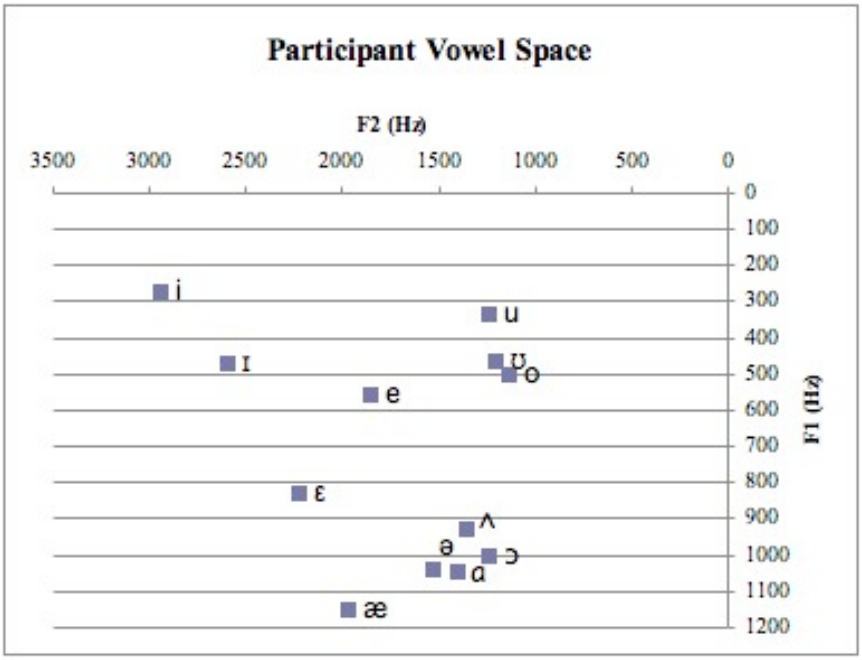

Figure A3: The vowel space of one of the participants. Similarities to *Figure A2*, like the placement of the /i/ and differences (i.e. the participant has a more centralized /e/ and a more open /u/) are what may be responsible for the difference that we see in individual vowels in the experiment.