Spring 2016

Feeling and Speaking: The Role of Sensory Feedback in Speech

Francesca R. Marino
Trinity College, Hartford Connecticut, francesca.marino@trincoll.edu

TRINITY COLLEGE

FEELING AND SPEAKING:
THE ROLE OF SENSORY FEEDBACK IN SPEECH

BY

Francesca Marino

A THESIS SUBMITTED TO
THE FACULTY OF THE NEUROSCIENCE PROGRAM
IN CANDIDACY FOR THE BACCALAUREATE DEGREE
WITH HONORS IN NEUROSCIENCE

NEUROSCIENCE PROGRAM

HARTFORD, CONNECTICUT
May 16, 2016

Feeling and Speaking:

The Role of Sensory Feedback in Speech

BY

Francesca Marino

Honors Thesis Committee

Approved:

____

Elizabeth Casserly, Thesis Advisor

____

Kent Dunlap, Thesis Committee

____

Sarah Raskin, Director, Neuroscience Program

Date: _______________________________________

TRINITY COLLEGE

Abstract

Sensory feedback allows talkers to accurately control speech production, and auditory information is the predominant form of speech feedback. When this sensory stream is degraded, talkers have been shown to rely more heavily on somatosensory information. Furthermore, perceptual speech abilities are greatest when both auditory and visual feedback are available. In this study, we experimentally degraded auditory feedback using a cochlear implant simulation and somatosensory feedback using Orajel. Additionally, we placed a mirror in front of the talkers to introduce visual feedback. Participants were prompted to speak under a baseline, feedback degraded, and visual condition; audiovisual speech recordings were taken for each treatment. These recordings were then used in a playback study to determine the intelligibility of speech. Acoustically, baseline speech was selected as "easier to understand" significantly more often than speech from either the feedback degraded or visual condition. Visually, speech from the visual condition was selected as "easier to understand" significantly less often than speech from the feedback degraded condition. Listener preference of baseline speech was significantly greater when both auditory and somatosensory feedback were degraded then when only auditory feedback was degraded (Casserly, in prep., 2015). These results suggest that feedback was successfully degraded and that the addition of visual feedback decreased speech intelligibility.

Introduction

I. Background

Sensory information plays an important role in many aspects of a person's life, particularly during perception, motor production, and other cognitive tasks. The implications of such processes participate in behaviors ranging from crossing the street to successfully conversing. These functions influence many daily operations; therefore, sensory processing requires a large cognitive demand (Powers, *et al*., 2012). Abnormal processing, specifically of acoustic information, can be a symptom of various neurological conditions, including central auditory nervous system tumors, epilepsy, and attention deficit hyperactivity disorder (Bamiou, *et al*., 2016). In some cases, these conditions generate learning and language disorders, which may cause developmental, social, and academic deficits (Kruger, *et al*., 2001). Aside from intrinsic biological origins, environmental factors directly impact sensory processing abilities. For example, during development, an environment with a high degree of noise significantly decreases the growth and differentiation of superior colliculus neurons in both the visual and auditory cortices (Xu, *et al*., 2014).

It is estimated that central auditory processing disorders affect 3-5% of the national population (Hear-It). Once an auditory processing disorder has developed, it is extremely difficult to improve the impaired communication and hearing abilities. The most common therapeutic method is auditory training, which attempts to improve communication capacity through listening practice sessions. Unfortunately, these programs often yield low proportions of success and compliance (Tye-Murray, *et al*., 2012). Since there are limited treatment options for these conditions, it is exceedingly important to study the neural mechanisms that regulate sensory processing and integration.

II. Sensory Feedback Alterations

Experimentally altering the sensory feedback of neurotypical adult speakers has a profound effect on speech behavior. Delaying auditory feedback significantly slows speech rates and causes talkers to produce more speech errors (Stuart & Kalinowsi, 2015), while changing the amplitude of auditory feedback causes talkers to compensate by altering speech loudness (Lane & Tranel, 1971). Acoustic feedback has been modified experimentally using cochlear implant simulations; this technology employs vocoding techniques to degrade feedback by mapping acoustic information onto a small series of frequency channels in real-time (Casserly, 2015). Such cochlear implant simulation studies have shown that degrading auditory feedback adversely affects the intelligibility of speech (Burkholder, *et* al., 2004; Casserly, *et al*., in prep.) and causes talkers to exhibit somatosensory compensation by collapsing vowel height (Casserly, 2015).

There are indications that manipulating somatosensory feedback has similar effects on speech behavior. The efficacy of somatosensory feedback has been degraded experimentally using bite blocks (Lane, *et al*., 2005), which modify articulator positions and cause talkers to exhibit compensatory behaviors during speech production (Houde & Nagarajan, 2011). Under these altered somatosensory conditions, removal of auditory feedback caused talkers to increase the fundamental frequency of vowels (Turgeon, *et al.,* 2015). These direct changes in production imply that sensory feedback has a critical function in speech behavior. Feedback likely allows talkers to self-monitor their speech production, accurately convey information, and interpret meanings of speech (Meekings, *et al*., 2015).

III. Neural Models

Although sensory feedback is clearly involved in speech fluency, it is not fully understood how this information is incorporated into speech behavior. The current neural models propose that feedback and feedforward control enhance the accuracy of speech production (Guenther & Vladusich, 2012; Houde & Nagarajan, 2011). Feedforward control is constructed from auditory and somatosensory information that accumulates over time; this is the system that initially enables talkers to produce the desired target sound. By contrast, feedback control develops an informative error-prediction loop to correct and regulate speech accuracy (Guenther & Vladusich, 2012). Speech errors can be processed two different ways, and each method yields distinct characteristics in the produced speech (the “H & H” model, Lindblom, 1990). Hypo-speech is output-oriented, as the increased neural energy expenditure allows the target words to become more clear and enunciated. In this case, the feedback system improves production accuracy by analyzing speech errors to alter speech behavior. On the contrary, hyper-speech is less regulated and reduces the cognitive effort for accurately producing speech. In this case, the feedback system does not incorporate speech errors to induce corrective actions; therefore, the speech system becomes the only regulatory mechanism to normalize speech accuracy (Lindblom, 1990). Depending on the environment and communication goals (e.g. level of speech accuracy or energy expenditure), talkers can vary between hypo- and hyper-speech.

There is an apparent connection between feedback and feedforward control; however, the specific features of this relationship are unknown. When learning a new behavior, people typically exhibit a gradual transition from feedback to feedforward control. To further understand this conversion, researchers trained participants to reach for an online visual target while controlling their right hand in a mirror-reversed visual task (Kasuga, *et al*., 2015). This method was considered to be learning a "new control strategy" because participants did not have any previous experience performing tasks of this nature. Subjects exhibited a strong feedback response during initial trial blocks; however, the magnitude of this response decreased over time. Conversely, the feedforward response was originally weak, but increased during subsequent trial blocks. These findings indicated a mechanistic reversal from feedback to feedforward control as the participants acquired more information through learning. Additional analysis of the visual-motor response onset, duration, and accuracy, also suggested that the participants' feedforward and feedback control systems developed separately (Kasuga, *et al*., 2015).

Once a speaker's feedback control circuit has developed, sustained damage or alteration to the auditory system can cause this circuit to degrade. Deafened adult speakers who undergo surgery to obtain cochlear implants exemplify this phenomenon. This procedure causes the original auditory feedback system to shift in frequency. After surgery, talkers learn to map the altered auditory feedback to the corresponding articulator positions, as well as form updated predictions about the analogous speech outputs (Lane, *et al*., 2007). Through this process, talkers update and reestablish their feedback control circuit.

It is evident that the brain has sophisticated mechanisms to regulate speech perception and production. Currently, there are two leading models to explain this phenomenon: the hierarchical state feedback control (HSFC) model (Hickok, 2014), developed from the state feedback control model (Houde & Nagarajan, 2011), and the directions into velocities of articulators (DIVA) model (Guenther & Vladusich, 2012). The principle distinction between these mechanisms is how each organizes the processing levels. In the state feedback control (SFC) model (Figure 1), a neural signal representing a speech sound travels from the motor cortex to the vocal tract. This initiates a change in the position of the talker's articulators - the larynx, tongue, pharynx, lips, etc. - and results in the final speech output. Over time, the speaker's brain collects perceptual data from speech outputs to form a prediction pathway for speech production. This pathway includes expectations for vocal tract position and the corresponding speech output, forming a method of improving speech accuracy using sensory feedback (Houde & Nagarajan, 2011). The HSFC model goes one step further, by organizing the levels of speech perception into a hierarchal system. Similar to the SFC model, a conceptual input causes an initial neural activation. This then projects to the high-level cortical loop, containing both sensory and motor regions, and finally reaches the low-level somatosensory-cerebellar-motor circuit (Hickok, 2014).

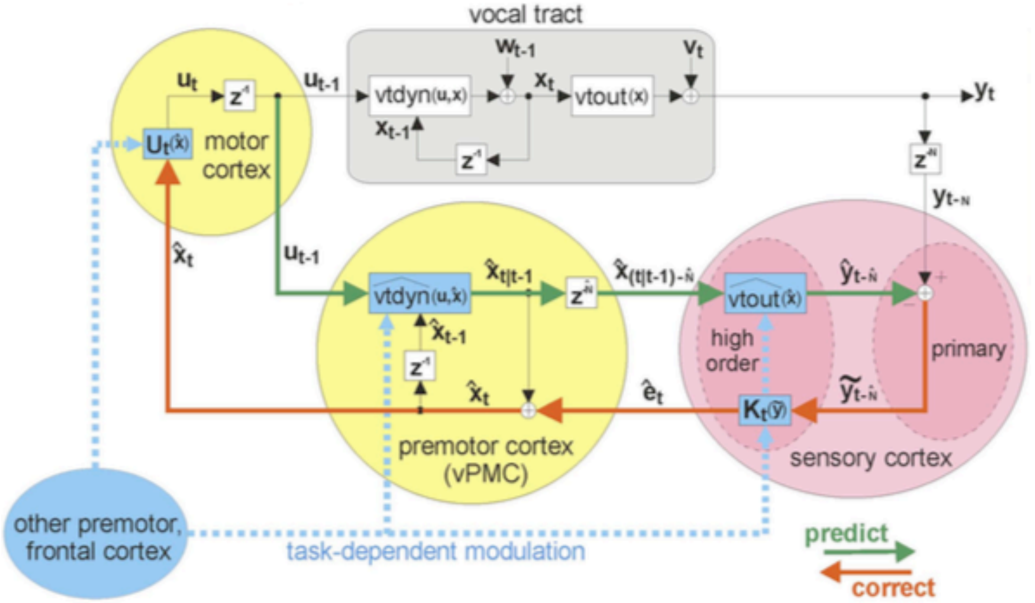

Figure 1: State Feedback Control model, showing the real-time speech production pathway (top) and the feedback-based error-prediction circuit (bottom) (Houde & Nagarajan, 2011).

Alternatively, in the DIVA model (Figure 2), speech outputs are produced after specific neurons in the talker's speech sound map are triggered (Guenther & Vladusich, 2012). The speech sound map is a collection of neurons in the left ventral premotor cortex and posterior Broca's area. This map is highly organized, such that each syllable is associated with a population of neurons in the cortex. Activation of the speech sound map initiates the projection of motor commands to the primary motor cortex. From here, two neural circuits regulate speech production: feedforward and feedback control. In the feedforward loop, the speech sound map projects directly to the cerebellum and primary motor cortex to initiate articulator movements. The feedback loop is more complex, as it contains distinct circuits for auditory and somatosensory information (Guenther & Vladusich, 2012).

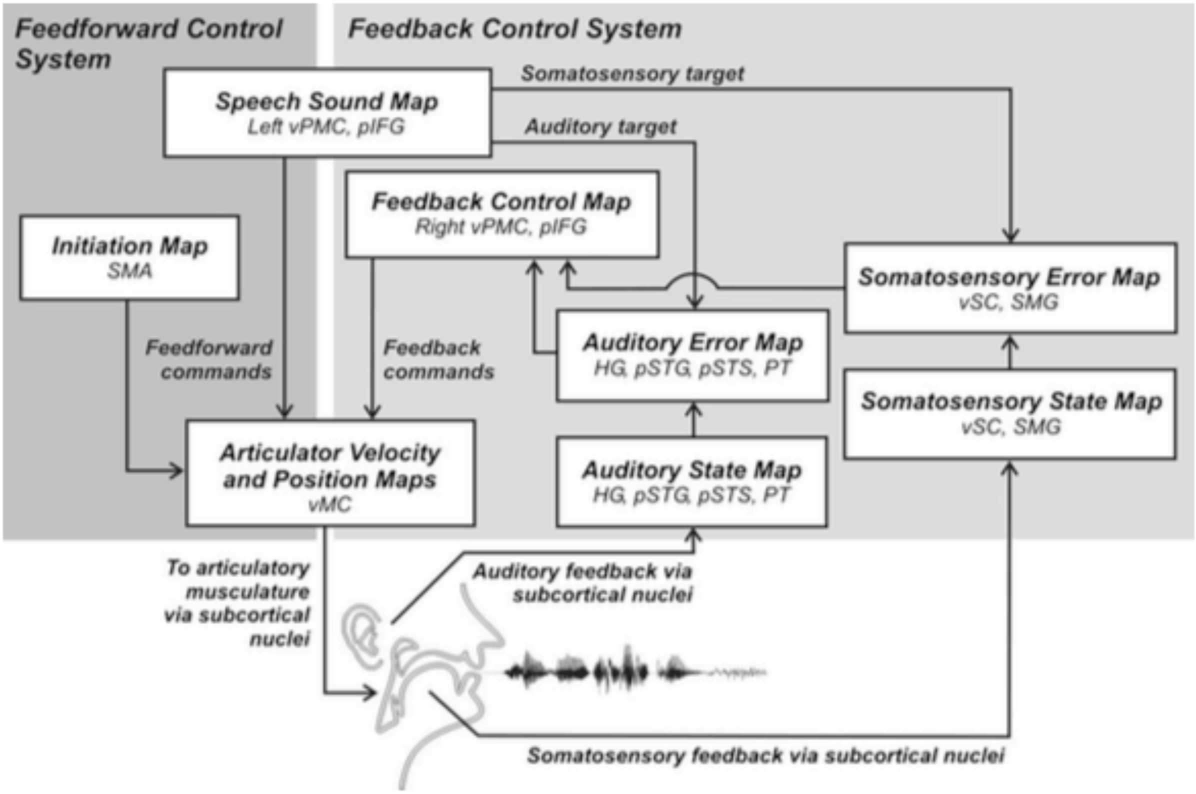

Figure 2: Directions Into Velocities of Articulators model, showing feedforward and feedback control circuits originating from the speech sound map (Guenther & Vladusich, 2012).

Although both the HSFC and DIVA models explain how feedback is incorporated into speech behavior, there are significant organizational differences between the accounts. The principal distinction between these two neural mechanisms is that the HSFC model appoints auditory feedback to syllable-level processing, while somatosensory feedback is assigned to phoneme-level processing. The HSFC model considers auditory feedback to be of a more advanced processing level than somatosensory feedback, whereas the DIVA model does not explicitly designate an organizational hierarchy. An interesting similarity is that both models depend on learned sensory information during the formation of speech production pathways (Guenther, 2014).

IV. Neural Regions for Speech

Many studies identifying the brain regions involved in speech production have employed classic fMRI analysis (Behroozmand, *et al*., 2015). During speech production, neural activity significantly increases in the following areas: temporal lobe, Heschl's gyrus, precentral gyrus, supplementary motor area, inferior frontal gyrus, insula, etc. (Behroozmand, *et al*., 2015). PET studies have also shown activation of neural regions associated with both auditory and visual processing during lexical perception tasks. Regions specific to vision include the striate, extrastriate, and occipital cortices, while auditory areas include the temporal and cingulate cortices (Petersen, *et al*., 1988). Given these findings, it is logical that the brain integrates different sensory streams to improve perceptual accuracy.

Spatial and temporal congruence are the predominant factors that affect the likelihood that sensory streams will be combined; this concept is referred to as the "multisensory binding process" (Powers, *et al*., 2012). Brain regions including the posterior superior temporal sulcus, inferior parietal lobe, insula, and superior colliculus are involved in the development of this process, as well as general multisensory integration (Powers, *et al*., 2012). Many brain regions are either dedicated to or involved in sensory processing; therefore, these multisensory integration functions must provide people with a significant perceptual advantage (Powers, *et al*., 2012).

V. Sensory Integration

Integrating sensory information conserves neural energy, increases perceptual accuracy, and improves response times (Altieri, *et al*., 2015). When participants were trained to associate a particular auditory tone with a visual cue, reaction times improved and energy expenditure decreased during the perceptual discrimination tasks (Altieri, *et al*., 2015). Furthermore, rhesus monkeys learned to integrate visual and somatosensory information to successfully produce a goal-directed movement. This behavior was significantly more accurate when both sensory streams were available for use (Dadarlat, *et al*., 2015).

It appears that the brain's ability to integrate sensory information develops over time. Children exhibit difficulty when determining whether stimuli are relevant to a given task, and, often, they incorrectly rely on the dominant sensory stream (Petrini, *et al*., 2015). For example, in visuospatial discrimination tasks containing non-informative visual information, children depend on visual information significantly more than the reliable spatial information. This phenomenon is likely due to previous experiences; during development, children learn to associate visual information with accurate processing of their visual environment. Functionally, this inability to combine sensory streams during childhood may allow individual sensory pathways, such as hearing or seeing, to develop (Petrini, *et al*., 2015).

VI. Visual Feedback

Speech perception studies indicate that incorporating visual information with auditory and somatosensory feedback is beneficial (Peele & Sommers, 2015). Listeners exhibited significant improvements in the ease and accuracy of speech perception through the use of this additional information (Peele & Sommers, 2015). When comparing sounds from an auditory perspective, it is difficult to distinguish syllables based on frequency content alone. When visual information is added, the improvements in speech perception are likely due to the disambiguation of words; this is done by clarifying both speech onset and rhythm. Intended speech becomes more obvious as listeners view the talker's articulator positions and decrease the potential lexical neighborhood. As listeners gain more information about the speech sound in question, the accuracy and ease of speech perception increases significantly (Peele & Sommers, 2015). That said, certain words can be visually recognized more easily than others; phoneme segments that contain the maximal amount of visual information are called visemes (Fisher, 1968). Examples of visemes include the articulation of the following segments: p, b, f, v, etc., as these sounds are produced using the lips (Ladefoged & Johnson, 2014).

The clarity of visual information produced during speech varies greatly between talkers (Lesner & Kricos, 1981). Lipreading studies indicated that listeners assessed talkers to have diverse levels of intelligibility and that words containing visemes were significantly rated as easier to understand (Lesner & Kricos, 1981). Some talkers had consistently above-average measures of intelligibility across the population of listeners, suggesting that certain people produce speech in a way that is easier to lipread than others (Lesner & Kricos, 1981). Despite individual differences, talkers were generally able to lipread themselves more easily than others (Tye-Murray, *et al*., 2014). It is evident that visual feedback has a significant purpose in speech perception and production; however, it remains unclear how this sensory stream is incorporated into the current neural models.

Current Investigation

As shown by the research summarized above, sensory feedback directly effects speech production. Absent or abnormal auditory feedback produces speech intelligibility deficits, as well as decreased speech accuracy and speed (Burkholder, *et al*., 2004; Casserly, *et al*., in prep; Stuart & Kalinowsi, 2015). Correspondingly, manipulating the usefulness of somatosensory feedback induces compensatory behaviors in talkers' articulator positions (Houde & Nagarajan, 2011). Due to these significant correlations, researchers have been working to determine the neural mechanisms linking speech production and perception. Although there are clear differences between the current theories, both the HSFC and DIVA models state that speech sound signals project to cortical areas including the primary motor cortex and primary somatosensory cortex (Houde & Nagarajan, 2011; Guenther & Vladusich, 2012). Both systems also include an error-prediction pathway, which is the principal function of the talker's feedback loop. Over time, the brain collects data from speech outputs and forms this pathway to improve speech accuracy (Houde & Nagarajan, 2011; Guenther & Vladusich, 2012).

Although researchers have a general idea of how the brain regulates speech behavior, many of the specific details remain unknown. For example, both the HSFC and DIVA models emphasize the importance of learned information when creating an accurate prediction pathway (Houde & Nagarajan, 2011; Guenther & Vladusich, 2012). It is yet to be determined how visual information is incorporated into these models, as this sensory stream does not have a strong learned component. Nonetheless, it is evident that visual information is assimilated into speech perception; talkers exposed to visual information, in addition to auditory and somatosensory feedback, exhibited improved intelligibility and accuracy in their speech production (Peele & Sommers, 2015). It is also undetermined how speech behavior is altered when visual information is the only reliable, non-degraded stream of sensory feedback.

Furthermore, it is currently unknown how talkers adjust to depend on the most reliable form of sensory information. In a speech production study, a group of normal-hearing adults experienced degraded auditory feedback by hearing their own speech through a cochlear implant simulation (Casserly, 2015). In response, the talkers significantly collapsed the height of the produced vowels, suggesting that they disengaged from the non-reliable auditory feedback stream (Casserly, 2015; Casserly, *et al.*, in prep.). It is evident that adults possess the ability to discriminate between beneficial and meaningless sensory information; however, it remains unclear how these unconscious decisions are made through the currently accepted neural mechanisms.

The primary aim of the present investigation is to further understand the neural mechanisms that regulate speech behavior. I attempted to determine whether talkers can incorporate visual information, or the maximally available sensory information, in control of speech behavior. Additionally, I sought to validate the experimental disruption of speech feedback and further understand how reduced somatosensory information alters speech production. Changes in speech intelligibility, due to degraded or maximally available sensory information, were assessed through a speech perception experiment. This method asked listeners to judge the intelligibility of speech, and these tasks have been used in the past to evaluate changes in speech intelligibility as a result of experimental sensory manipulations (Casserly, *et al*., in prep.; Holt, *et al*., 2011).

To determine whether speech intelligibility changed as sensory information became more or less reliable, talkers were recorded across three conditions: 1. baseline, 2. auditory and somatosensory feedback degraded, and 3. visual feedback added. Auditory feedback was degraded using a cochlear implant simulation, and somatosensory feedback was degraded using Orajel. Orajel is a topical numbing agent that temporarily decreases somatosensory feedback. Visual feedback was provided by placing a large mirror in front of the talkers. During each condition, participants were asked to produce speech in response to randomized stimulus words. The recordings were then used to determine changes in speech intelligibility – using auditory discrimination, visual recognition, and visual discrimination tasks – between the various conditions.

Design and Methods

Phase I: Data Collection

I. Subjects

Fifteen native English speakers were recruited on a volunteer basis from Trinity College in Hartford, Connecticut. The mean age of the participants was 19.53 years, and the study participant pool was 53% male and 47% female. None of the participants reported a history of hearing loss, and each participant underwent an audiometric screening with <5 dB hearing loss between 500 Hz and 8000 Hz at the time of the experiment.

II. Experimental Design

Each subject produced speech in response to stimulus words, which were presented on a thirteen-inch laptop screen in a sound booth. During the first condition, participants spoke without feedback perturbation to produce a baseline of normal speech. In the second and third conditions, participants wore a portable real-time vocoder (PRTV; see Section D) to degrade auditory feedback by simulating a cochlear implant in real time. Talkers also received one mL Orajel on the lips and tongue to reduce somatosensory feedback. Orajel was measured in a syringe, and participants self-administered the numbing agent by applying the gel to the articulators. Orajel remained undisturbed for one minute before the experiment began to ensure the full numbing effects were exerted. In the third condition, participants were also exposed to visual information, via a three-by-three-foot mirror placed directly across from the speaker in the sound booth.

III. Stimulus Materials

Subjects read aloud a set of 139 isolated English words, which were repeated across the three conditions. Words were selected from the pocket dictionary of English words (Nusbaum, *et al*., 1984) and first sorted by familiarity, such that all selected words were of the highest familiarity rating. This ensured that talkers were acquainted with the stimuli and would not "sound out" words during production. This list was then arranged by frequency; the high frequency words had a prevalence of 319-68,971 per million (Nusbaum, *et al*., 1984), the medium frequency words had an incidence of 97-150 per million, and the low frequency words had an incidence of 6-7 per million. The final word list contained 45 high-frequency words, 45 medium-frequency words, and 49 low-frequency words. High-frequency words require less feedback than low-frequency words; therefore, frequency balancing ensured that the stimuli necessitated sensory feedback.

Since visemes contain the maximal amount of visual information, the following phoneme segments were included in a proportion of the stimuli: [m, b, p, f, v, i, ʃ, r, w] (Fisher, 1968). The high-frequency category was comprised of three words beginning with [m, b, p, f, i, w], two words beginning with [ʃ], one word beginning with [v, r], eleven words containing [m, b, p, f, v, i, ʃ, r, w] elsewhere in the word (e.g. middle or end), and twelve filler words not containing any of the phonemes of interest. The medium-frequency category contained three words beginning with [m, b, p, f, r, w], two words beginning with [ʃ], one word beginning with [v, i], eleven words containing the phonemes of interest elsewhere in the word, and twelve filler words. The low-frequency category was composed of three words beginning with [m, b, p, f, i, r, w], two words beginning with [ʃ], one word beginning with [v], thirteen words containing the phonemes of interest elsewhere in the word, and twelve filler words.

IV. Simulation of Cochlear Implant Processing

Auditory feedback was degraded using a portable real-time vocoder (PRTV) that simulated a cochlear implant (Casserly, 2015). Subjects wore earbuds, occluders, and a lapel microphone, which was wrapped around the occluder headphones to position the microphone adjacent to the speaker's ear. The PRTV employed an eight-channel noise vocoder with a window of 252-7000 Hz. This software took the continuous acoustic input from the lapel microphone and applied a noise-vocoded cochlear implant simulation. This divided the natural acoustic signal provided by the speaker into the eight frequency-based channels. During this frequency shift, acoustic input below 252 Hz and above 7000 Hz was lost. This modified feedback was relayed to the talker, via headphones, within 10 ms. This degraded feedback contained less information regarding details of acoustic frequency than natural speech (Casserly, 2015).

V. Procedure

Participants first read and signed the informed consent statements for the experiment and use of Orajel. Each subject then completed a demographic survey and hearing test. During each condition, subjects were seated in a chair (0.845 m high) in a sound-attenuating recording booth, 0.624 m from a thirteen-inch laptop screen. The laptop screen was located 0.261 m from the front edge of the table, height 0.737 m. A video camera and condenser microphone were also placed across from the talker, 0.599 m from the front edge of the table (Figure 3). Audiovisual recordings were taken of the participant during each condition. The first condition consisted of subjects speaking under a normal, baseline condition. The PRTV and Orajel were introduced in the second condition and remained in place during the third condition. Also in the third condition, a three-by-three-foot mirror was placed across from the speaker behind the laptop screen. Approximately 99% of the mirror surface area was visible to the participant. The entire experiment lasted for ~40 minutes. The stimuli were presented in a random order that was consistent throughout each phase of the experiment.

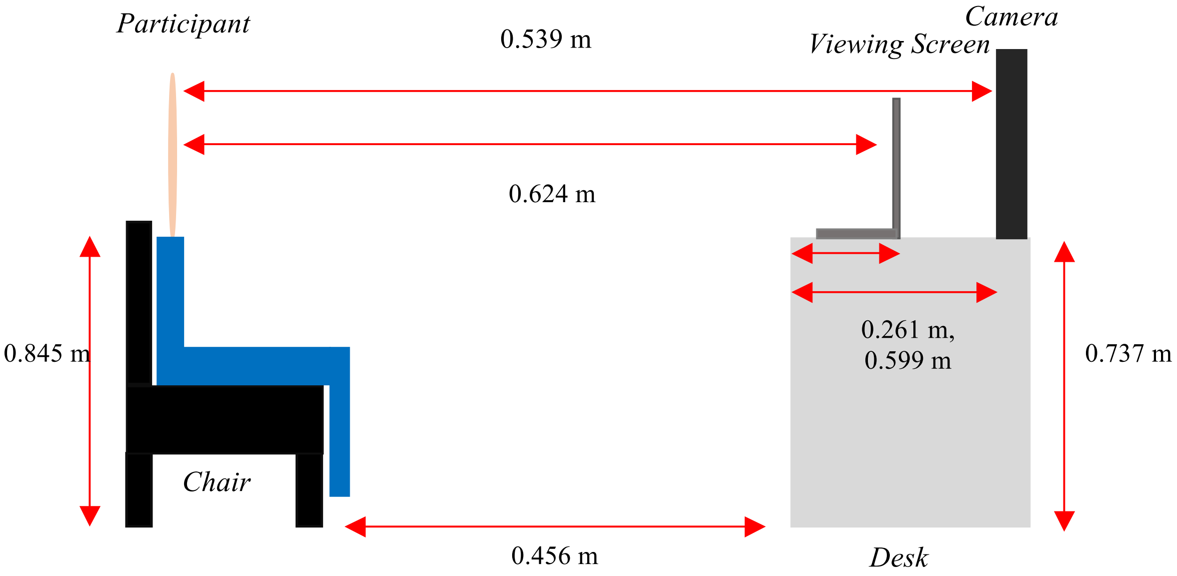

Figure 3: Visual representation of experimental setup in sound-attenuating booth for audiovisual recordings.

Phase II: Perceptual Response

I. Subjects

Thirty-eight native English speakers were recruited on a volunteer basis from Trinity College in Hartford, Connecticut. The mean age of the participants was 19.52 years, and the study participant pool was 32% male and 68% female. None of the participants reported a history of hearing loss, and each participant underwent an audiometric screening with <5 dB hearing loss between 500 Hz and 8000 Hz at the time of the experiment.

II. Experimental Design

Half of the subjects completed an auditory speech perception task (*n* = 19), and the other half participated in a visual speech perception task (*n* = 19). The auditory task focused on discrimination; participants heard two recordings and were asked to identify which was "easier to understand." For each comparison, both recordings were produced by the same talker. The visual task consisted of both discrimination and recognition. In the discrimination task, participants saw two recordings while knowing what the intended stimulus word was. They were then asked to determine which recording was "easier to understand" by lipreading. In the recognition task, participants saw one recording without knowing what the stimulus word was; they were then asked to transcribe what word they believed the talker was saying. In each task, participants responded to stimuli from all fifteen talkers, and the order of the stimulus words was randomized within each talker. In total, nineteen participants completed the auditory discrimination task, nineteen completed the visual discrimination task, and nineteen completed the visual recognition task.

III. Stimulus Materials

The stimulus words were chosen from the recordings collected in Phase I of the experiment. Nine unique words were selected for each of the fifteen talkers, resulting in a total of 135 words. For each talker, three words were high-frequency, three were medium-frequency, and three were low-frequency. For each frequency category, one word began with a phoneme containing maximal visual information, one word contained such a phonetic segment elsewhere in the word, and one word did not contain any of these letters. The words containing maximal visual information, also called visemes (Fisher, 1968), included the following letters: [m, b, p, f, v, i, ʃ, r, w]. Stimulus words were selected such that each viseme was equally represented amongst the talkers.

IV. Procedure

Participants first read and signed the informed consent statement for the experiment. Each subject then completed a demographic survey and hearing test. During each condition, subjects were seated in a small testing room containing a PC computer with a 19-inch monitor. Stimuli were presented and responses were recorded using Eprime software (Eprime 2.0, Psychology Software Tools, Inc.) For the auditory speech perception task, participants were asked to wear headphones and complete the auditory discrimination experiment. For the visual speech perception task, participants were asked to complete the visual discrimination and recognition experiments without wearing headphones. In each case, the entire experiment lasted for ~60 minutes.

Hereafter, baseline recordings will be referred to as condition 1, feedback degraded recordings as condition 2, and visual added recordings as condition 3. In the auditory discrimination task, the program was designed to portray each stimulus word three times. First as a condition 1v.2 comparison, second as a condition 1v.3 comparison, and third as a condition 2v.3 comparison. In each comparison, the listener was asked to identify which sound file was "easier to understand." In the visual discrimination and recognition tasks, the program was designed to portray the stimuli only as a condition 2v.3 comparison. In the discrimination comparison, the listener was asked to identify which video was "easier to understand" through lipreading. In the recognition comparison, the listener was asked to transcribe the stimulus word presented in the video. Since subjects were not wearing the PRTV in the baseline recordings, this condition was excluded from the visual speech tasks. I hypothesized that the visual presence of the PRTV could influence the speech selection of listeners in the visual speech perception task.

This bias was not present for the auditory version of the discrimination task because the recordings were sound files which did not contain any visual information.

V. Statistical Analysis

To determine whether speaker intelligibility changed across the feedback conditions, listener responses were compared across the three trial types. To test how often listeners chose speech from the highest feedback condition, one-sample t-tests were used to analyze the auditory and visual discrimination data. This analysis was also used to determine whether participants were choosing any conditions at rates significantly above chance. A paired-sample t-test was used to analyze the visual recognition data to determine whether accuracy differed between the two feedback-degraded conditions. In all 1v.2 and 1v.3 comparisons, speech from the baseline condition was considered to contain the highest level of feedback. In all 2v.3 comparisons, speech from the visual condition was considered to contain the highest level of feedback. Three of the visual recognition participants were excluded from statistical analysis because they exceeded the average lipreading performance (96-98% accuracy vs. M = 12.3%, SD = 10.09%).

If listeners selected baseline speech more often than chance in 1v.2 and 1v.3 comparisons, then this suggests that the experiment successfully degraded acoustic and somatosensory feedback. If listeners chose visual speech more often than feedback degraded speech in 2v.3 comparisons, then this indicates that speakers were able to incorporate visual speech feedback to improve intelligibility. If listeners selected feedback degraded speech more often than visual speech in 2v.3 comparisons, then this implies that the introduction of visual information decreased speech intelligibility. To determine whether somatosensory degradation influenced listener preference for baseline speech, an independent-samples t-test was used to compare the auditory discrimination 1v.2 data to data from an identical task performed with baseline speech and speech in which only auditory feedback was degraded (Casserly, *et al*., in prep.). If listeners chose baseline speech more often in the present investigation, then this suggests that the use of Orajel to experimentally degrade somatosensory feedback significantly increased preference for baseline speech.

Results

In the auditory discrimination task, analysis of speech selection patterns between the baseline condition and feedback degraded condition showed that listeners selected baseline speech as "easier to understand" 62.3% of the time (SD = 6.166%), which was significantly higher than chance (t = 8.467, p < 0.001) (see Figure 4). This same finding was observed in the analysis of speech selection patterns between the baseline condition and visual condition. Again, listeners selected speech from the baseline condition as "easier to understand" 62.3% of the time (SD = 5.547%), which was also significantly higher than chance (t = 9.404, p < 0.001) (see Figure 4). In this same task, analysis of speech selection patterns between the feedback degraded condition and visual condition showed that listeners selected feedback degraded speech as "easier to understand" 47.99% of the time (SD = 4.887%), which was not significantly higher than chance (t = -1.748, p = 0.099). This suggests a possible perceptual advantage for speech produced in the feedback degraded condition, as opposed to the introduction of visual feedback (see Figure 4).

In the visual recognition task, analysis of correct stimulus-word transcription responses between the feedback degraded and the visual conditions showed that listeners did not exhibit differences in lipreading accuracy between these conditions (M = 12.33%, SD = 10.095%; t = -0.209, p = 0.837; see Figure 5). However, in the visual discrimination task, analysis of speech selection patterns between the feedback degraded and visual conditions showed that listeners selected speech produced with added visual information as “easier to understand” 46.99% of the time (SD 4.929%), which was significantly lower than chance (t = -2.728, p = 0.013) (see Figure 6).

Finally, differences in baseline preference patterns were compared between the present study and the data collected in Casserly, *et al.*, (in prep), where normal speech and speech produced with only auditory feedback degradation were used in an identical discrimination task. This analysis was conducted to determine whether the additional degradation of somatosensory feedback in the present study affected listeners’ speech selection patterns – specifically, the degree to which they preferred baseline speech over feedback-degraded speech. In Casserly, *et al*., (in prep.), analysis of speech selection patterns between the baseline condition and auditory feedback degraded condition showed that listeners selected baseline speech as “easier to understand” 56.47% of the time (SD = 8.083%), which was significantly higher than chance (p = 0.0012). In the present auditory discrimination task, as described above, listeners selected baseline speech as “easier to understand” more often than speech with degraded auditory and somatosensory feedback. Baseline speech was selected 62.3% of the time (SD = 6.166%), which was also significantly higher than chance (p < 0.001). Independent-samples t-test analysis of these baseline speech selection patterns showed that listeners’ preference for baseline speech in the present investigation was significantly higher than that found in Casserly, *et al*., (in prep.) (p = 0.006; see Figure 7).

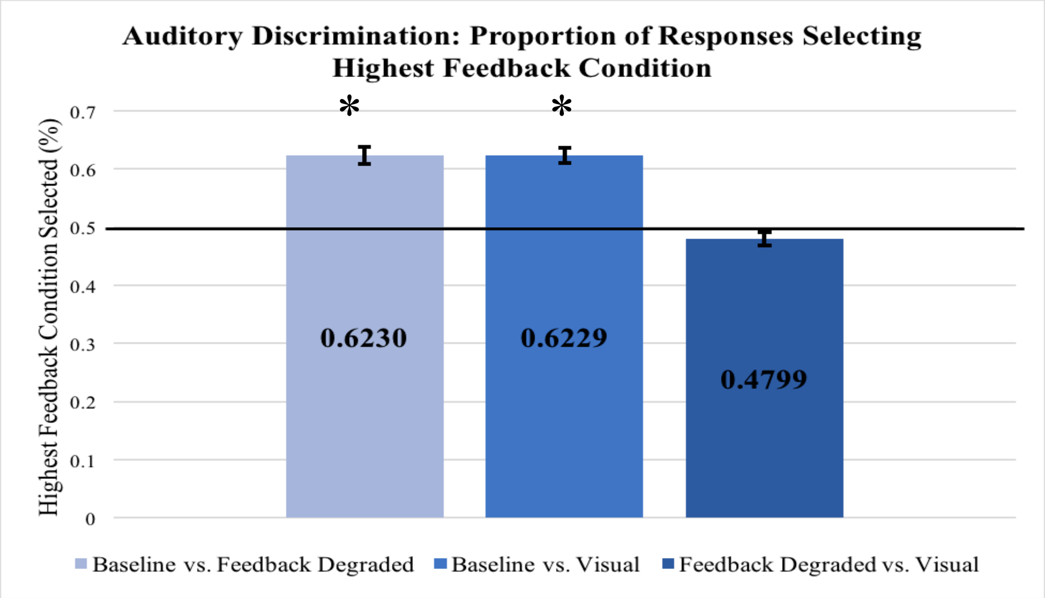

Figure 4: Proportion of time listeners selected the higher feedback condition as "easier to understand," acoustically, as opposed to visually. In both the baseline vs. feedback degraded analysis and baseline vs. visual-added analysis, baseline speech contained the most feedback. In the feedback degraded vs. visual-added analysis, visual feedback added speech contained the most feedback. Baseline speech was selected significantly more often than speech from either the feedback degraded or visual-added condition. This speech selection pattern was significantly higher than chance (horizontal line at 0.5).

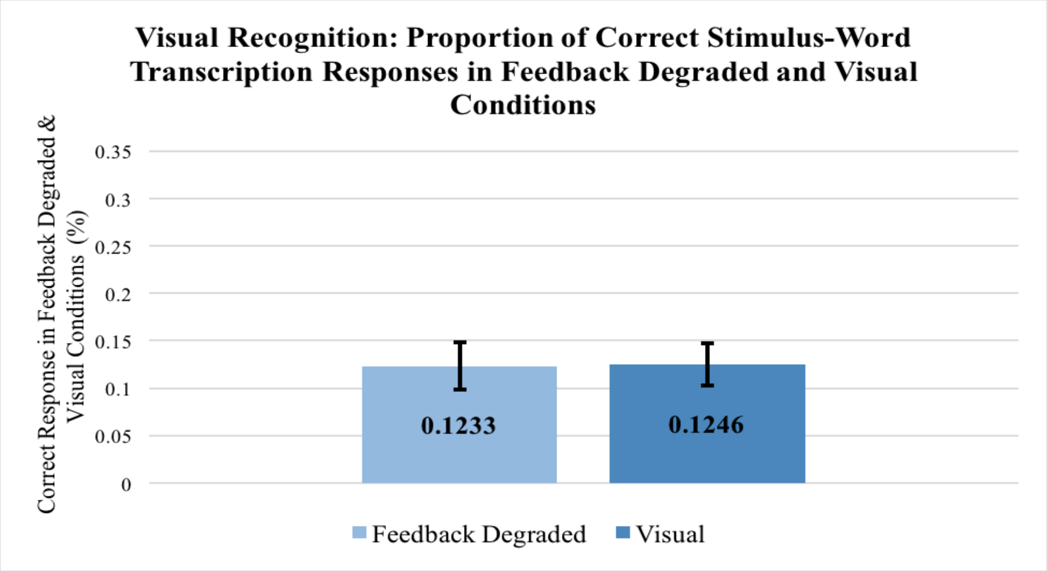

Figure 5: Proportion of correct stimulus-word lipreading transcription responses for the feedback degraded and visual-added speech conditions. Listeners did not exhibit differences in lipreading accuracy between the conditions.

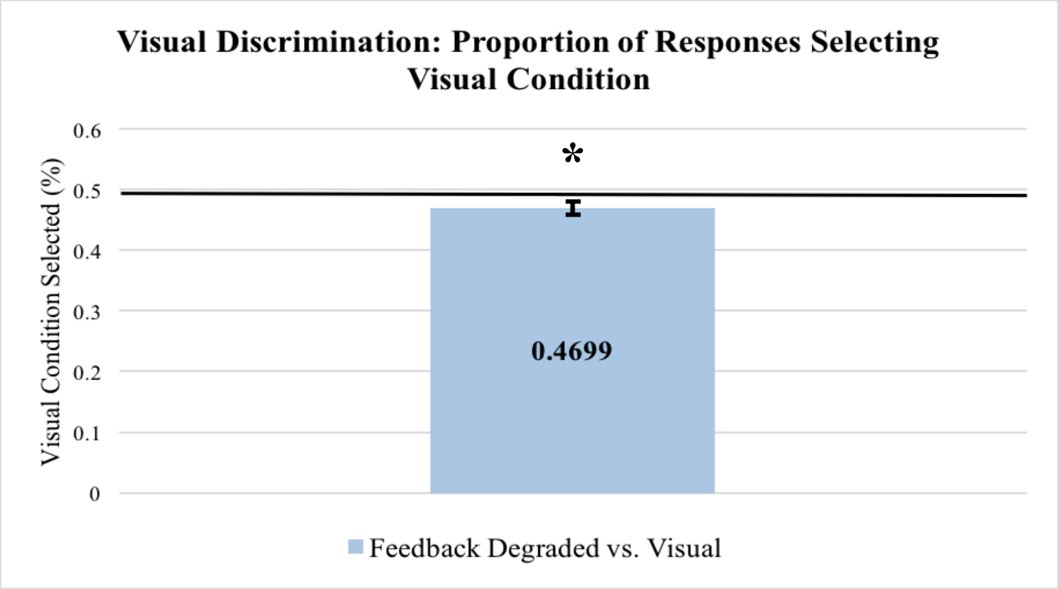

Figure 6: Proportion of the time listeners selected the higher feedback condition as "easier to understand," visually (via lipreading), as opposed to acoustically. In the feedback degraded vs. visual-added analysis, visual speech contained the most feedback and was selected significantly less frequently than chance, showing an advantage for feedback degraded speech.

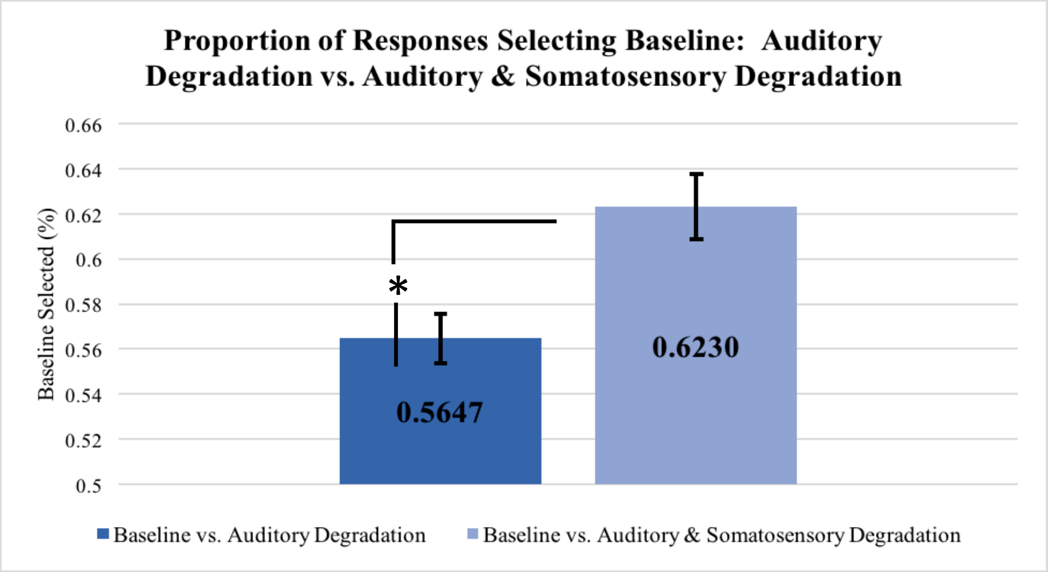

Figure 7: Proportion of time listeners selected the baseline condition as "easier to understand" in Casserly *et al*. (in prep.) and in the present study (baseline vs. feedback degraded condition only). The same auditory feedback degradation was used in both studies. In the present investigation, somatosensory feedback was also degraded using Orajel. Analysis of speech selection patterns showed that listeners in both studies selected baseline speech significantly more frequently than chance (horizontal line at 0.5). Selection rates differed significantly

between the two studies, with auditory and somatosensory degradation resulting in a greater intelligibility difference than the prior auditory-only degradation.

Discussion

I. Experimental Degradation of Feedback

In the present investigation, we experimentally modified speakers' sensory feedback and measured perceptual speech intelligibility. Auditory feedback was degraded using a cochlear implant simulation, which mapped natural spoken frequencies to eight frequency-based channels, in real-time (Casserly, 2015). Speakers self-applied Orajel to numb their articulators and degrade somatosensory feedback. We predicted that these experimental manipulations would be sufficient to disrupt speech feedback, and, therefore, impair production accuracy (e.g. speech intelligibility). This hypothesis was supported by the results of all three perceptual tasks.

In the auditory discrimination task, listeners selected speech from the baseline condition as "easier to understand" significantly more often than speech from the feedback degraded condition. In this study, the baseline condition provided listeners with more feedback than the auditory and somatosensory degraded condition. This pattern is consistent with previous studies that observed differences in perceptual speech intelligibility using listener judgments of normal speech and speech produced with feedback degraded through a cochlear implant simulation (Casserly, *et al*., in prep.;), as well as judgments of normal speech and speech produced under vocal tract alterations (Jones & Munhall, 2003).

In the present study, the baseline condition also provided listeners with more feedback than the visual condition. Visual feedback was introduced by placing a large, square mirror directly in front of the speaker while they were wearing the real-time cochlear implant simulator and had Orajel applied to their lips and tongue. Despite this added feedback stream, listeners still selected speech from the baseline condition as “easier to understand” for lipreading significantly more often than speech from the visual condition. This indicates that the addition of visual feedback did not provide listeners with enough information to fully compensate for the auditory and somatosensory degradation. This pattern is consistent with studies that examined speech perception with and without visual feedback (Schwartz, *et al*., 2004). Although intelligibility is greatest when both auditory and visual feedback are present, perceiving speech with only visual feedback is significantly more difficult than with only auditory feedback (Schwartz, *et al*., 2004). However, since listeners consistently selected speech from the baseline condition across all analyses, we conclude that our manipulations were effective in disrupting speech intelligibility via feedback degradation.

II. Somatosensory Feedback

We were also interested in determining the effects of somatosensory degradation on speaker intelligibility. We predicted that the modification of both auditory and somatosensory feedback would produce a greater decrease in speech intelligibility than the degradation of only auditory feedback. The comparison of auditory discrimination data between somatosensory and auditory degradation (current investigation) and only auditory degradation (Casserly, *et al*., in prep.) supported this hypothesis. In the present study, listeners selected speech from the baseline condition as “easier to understand” significantly more often than speech from the feedback degraded condition. Similar results were observed in Casserly, *et al*., (in prep.), where listeners selected speech from the baseline condition as “easier to understand” significantly more often than speech from the auditory degraded condition. However, the baseline preference rate was significantly higher in the present investigation than in Casserly, *et al*., (in prep.). These findings indicate that degrading both auditory and somatosensory feedback decreased speech intelligibility significantly more than degrading auditory feedback alone. In the present investigation, we did not alter auditory and somatosensory feedback separately. We chose to include only one feedback degraded condition because we were concerned than an extended experimental duration would cause the Orajel to wear off prematurely or allow the talkers to learn the stimulus words. From the difference in baseline preference rates across the two studies, however, we can conclude that we were successful in manipulating somatosensory feedback using Orajel. To our knowledge, this technique has never been previously used to degrade feedback.

This finding is consistent with theories that explain the neural mechanisms that regulate speech production (Houde & Nagarajan, 2011). The HSFC model implicates the auditory, motor, and premotor cortices as necessary in the formation of an accurate feedback circuit. Degrading the usefulness of somatosensory feedback prohibits the motor and premotor regions from contributing to the active error predictions made by the feedback loop. If acoustic feedback remains present, then the auditory cortex can provide information to the feedback circuit. In the case of total feedback degradation, neither auditory nor somatosensory feedback is available to the talker. The motor, premotor, and auditory brain regions would then be unable to contribute sensory information to the feedback loop, which would render the entire circuit uninformative. It is, therefore, not surprising that speech intelligibility significantly decreased when all forms of sensory feedback were diminished (Houde & Nagarajan, 2011).

III. Visual Feedback

This investigation also served to explore changes in speech intelligibility when visual information was the maximally available sensory stream. Although we predicted that the addition of visual feedback would improve talkers' intelligibility over the feedback degraded condition, the results of the visual discrimination and recognition tasks failed to support this hypothesis. Particularly, listeners selected speech from the feedback degraded condition as “easier to understand,” through lipreading, significantly more often than speech from the visual feedback condition during the visual discrimination task. Despite the fact that this preference was non-significant in the auditory version of this discrimination task, the data trended towards the same pattern. Listeners found talkers had lower intelligibility with added visual feedback than with completely degraded sensory feedback. Furthermore, there were no differences in listeners’ recognition accuracy between speech produced under the feedback degraded condition and visual condition. Together, the results suggest that the introduction of visual feedback made speakers slightly more difficult to lipread. The non-significant results in the visual recognition task caused us to consider speakers to be partially, but not completely, more challenging to lipread when visual feedback was added.

It is not surprising that the differences in visual speech intelligibility were only detected in the discrimination tasks, as these comparisons were more sensitive than the recognition task. Success in the recognition task can be likened to firing an action potential in that it is binary. Just as a neuron either receives enough electrical charge to fire or it does not, there is either enough perceptual information to allow for successful word recognition or there is not. In the recognition task, participants were asked to transcribe stimulus words through lipreading; therefore, the responses either hit or missed the threshold. Responses that were off by at least one phoneme segment were regarded as incorrect. In the discrimination task, however, participants selected the condition they found to be easier to understand; as a result, the responses were positioned on a gradient of intelligibility. In this task, some small discrepancy in the speaker’s production of the stimulus word could contribute to an unconscious preference in the listener’s perceptual ability.

In keeping with the analogy, this gradient can be equated to graded excitatory post-synaptic potentials that eventually summate to reach the threshold of excitation.

There are three potential explanations for the finding that introducing visual feedback adversely affected talkers' intelligibility. The first relates to the neural models of speech production, as this result is consistent with both the HSFC and DIVA models (Houde & Nagarajan, 2011; Guenther & Vladusich, 2012). These models emphasize the importance of learning when creating the feedforward and feedback loops. Typically, people are unfamiliar with watching themselves talk; therefore, it is possible that the lack of experience prevented the successful integration of visual information into the speakers' feedback circuits. In cases where visual information was found to improve speech intelligibility, auditory feedback was still available (Peele & Sommers, 2015). This suggests that while visual information is a beneficial addition, this sensory stream cannot act as the sole reliable form of feedback.

Aside from the potential difficulties associated with incorporating this sensory information into feedback circuits, it is also possible that introducing visual information created a cost in cognitive effort. This addition charged speakers with determining whether to ignore or employ the novel source of feedback. Before learning a behavior, there is an initial increase in neural energy expenditure, which decreases as the task becomes more familiar. In a visuospatial experiment, for example, participants exhibited a significant reduction in response time as they accumulated information about the task through learning (Kasuga, *et al*., 2015). Decreased response time and increased perceptual accuracy are commonly associated with decreased neural energy expenditure (Alteri, *et al*., 2015). It is likely that the talkers in the present investigation did not have sufficient learned information to successfully make use of the added visual feedback. Given these findings, it is possible that visual information could be incorporated into feedback circuits if people received adequate training or learned to employ this sensory stream.

The third alternative is that the mirror distracted the speakers. Deakin and Wakefield (2014) discussed the various complications experienced by two researchers during a Skype interview. The abnormal presence of visual feedback caused the researchers to lose focus, as immediate visual self-feedback is typically not present in everyday conversations. In the present investigation, it is possible that the mirror caused speakers to experience this distraction phenomenon. The decreased intelligibility, therefore, could be due to speakers becoming less focused during their production of the presented stimulus words.

Future Research

Additional studies are necessary to fully understand the effects of visual feedback on speech production. Specifically, future research could experimentally distinguish between the neural integration, cognitive effort, and distraction hypotheses that explain the observed decrease in speech intelligibility. In the Data Collection Phase, speakers were not explicitly told to look at the mirror while producing the stimulus words; therefore, it is possible that some speakers made use of the mirror, while others disregarded the added visual information. The instructions could be clarified or experimentally manipulated to test these hypotheses. Furthermore, additional studies are needed to determine whether visual information can be incorporated into feedback circuits when there is sufficient learned information. A future experiment could compare speech intelligibility between a population of talkers who have had increased exposure to visual feedback (e.g. through Skype or FaceTime) and a control group similar to the talkers in the present investigation.

Conclusion

Cochlear implant simulations and Orajel appear to be sufficient methods of degrading auditory and somatosensory feedback. Furthermore, auditory discrimination, visual recognition, and visual discrimination tasks seem to be viable techniques for detecting alterations in speech intelligibility. In this investigation, we observed that listeners selected speech from a baseline condition as easier to understand significantly more often than speech produced with degraded acoustic and somatosensory feedback, regardless of the presence of an alternative (visual) information source. The addition of visual feedback to the otherwise degraded speech condition did not improve speech intelligibility. In fact, in the visual discrimination task, the introduction of visual feedback significantly decreased talker's clarity. These results indicate that listeners could differentiate between normally produced and feedback manipulated speech. These findings also suggest that the addition of visual feedback was detrimental to the speakers' intelligibility. The conclusions of the present investigation are important for furthering the understanding of speech production mechanisms. In addition, these findings have potential clinical applications for generating novel therapies for deaf populations and individuals with central auditory processing disorders.

References

Altieri, N. *et al.* (2015). Learning to associate auditory and visual stimuli: Behavioral and neural mechanisms. *Brain Topogr*, 28: 479-493.

Bamiou, D. *et al*., (2016). Aetiology and clinical presentations of auditory processing disorders – A review. *Archive of Diseases in Childhood*, 85: 361-365.

Behroozmand, R. *et al*., (2015). Sensory-motor networks involved in speech production and motor control: An fMRI study. *NeuroImage*, 109: 418-428.

Burkholder, R. *et al*., (2004). Perceptual learning and nonword repetition using a cochlear implant simulation. *International Congress Series*, 1273: 208-211.

Casserly, E. (2015). Effects of a real-time cochlear implant simulation on speech production. *Acoustical Society of America*, 137(5): 2791-2800.

Casserly, E. *et al*., (in prep.). The mechanism behind speaker responses to cochlear implant simulation of acoustic feedback: Correction or disengagement? 1-14. in prep.

Dadarlat, M. *et al*. (2015). A learning-based approach to artificial sensory feedback leads to optimal integration. *Nature Neuroscience*, 18(1): 138-144.

Deakin, H. & Wakefield, K. (2014). SKYPE interviewing: Reflections of two PhD researchers. *Qualitative Research*, 0(0): 1-14.

Fisher, C. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11: 796-804.

Gangarosa, L. *et al*. (1989). Use of verbal descriptors, thermal scores and electrical pulp testing as predictors of tooth pain before and after application of benzocaine gels into cavities of teeth with pulpitis. *American Dental Society of Anesthesiology*, 36: 272-275.

Guenther, F. (2014). Auditory feedback control is involved at even sub-phonemic levels of speech production. *Language, Cognition, and NeuroscienceI,* 29(1): 44-45.

Guenther, F. & Vladusich, T. (2012). A neural theory of speech acquisition and production. *Journal of Neurolinguistics*, 25: 408-422.

Guleyupoglu, B. *et al*. (2014). Reduced discomfort during high-definition transcutaneous stimulation using 6% benzocaine. *Frontiers in Neuroengineering*, 7(28): 1-3.

Hear-It. *Central Auditory Processing Disorder*. Retrieved from http://www.hear-it.org/Central-Auditory-Processing-Disorders.

Hersh, E., *et al*. (2013). 10- and 20-Percent benzocaine gels are effective for the temporary relief of toothache and are well-tolerated. *Journal of Evidence-Based Dental Practice*, 144(5): 17-26.

Hickok, G. (2014). The architecture of speech production and the role of phoneme in speech processing. *Language, Cognition, and Neuroscience*, 29(1): 2-20.

Holt, R. *et al*. (2011). Assessing multimodal spoken word-in-sentence recognition in children with normal hearing and children with cochlear implants. *Journal of Speech, Language, and Hearing Research*, 54: 632-657.

Houde, J. & Nagarajan, S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience*, 5(82): 1-14.

Jones, J. & Munhall, K. (2003). Learning to produce speech with an altered vocal tract: The role of auditory feedback. *Acoustical Society of America*, 113: 532.

Kasuga, S. *et al*. (2015). Learning feedback and feedforward control in a mirror-reversed visual environment. *Journal of Neurophysiology*, 114: 2187-2193.

Kruger, R. *et al*. (2001). Relationship patterns between central auditory processing disorders and language disorders, learning disabilities, and sensory integration dysfunction. *Communication Disorders Quarterly*, 22(2): 87-98.

Ladefoged, P. & Johnson, K. (2014). A Course in Phonetics. *Cengage Learning*, 1- 45.

Lane, H. *et al*., (2005). Effects of bite blocks and hearing status on vowel production. *Acoustical Society of America*, 118: 1636.

Lane, H. *et al*., (2007). Effects of short- and long-term changes in auditory feedback on vowel and sibilant contrasts. *Journal of Speech, Language, and Hearing Research*, 50: 913-927.

Lane, H. & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech, Language, and Hearing Research*, 14: 677-709.

Lesner, S. & Kricos, P. (1981). Visual vowel and diphthong perception across speakers. *J.A.R.A.*, 16: 252-258.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. *Speech Production and Speech Modeling*, 403-439.

Meekings, S. *et al*., (2015). Do we know what we're saying? The roles of attention and sensory information during speech production. *Psychological Science*, 26(12): 1975-1977.

Nusbaum, H, *et al*. (1984). "Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words," Progress Report No. 10, Research on Speech.

Peele, J. & Sommers, M. (2015). Prediction and constraint in audiovisual speech perception. *Elsevier Limited*, 1-13.

Petersen, E, *et al*., (1988). Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature*, 331: 585-589.

Petrini, K, *et al*. (2015). Hearing where the eyes see: Children use an irrelevant visual cue when localizing sounds. *Child Development*, 86(5): 1449-1457.

Powers, A. *et al*. (2012). Neural correlates of multisensory perceptual learning. *The Journal of Neuroscience*, 32(18): 6263-6274.

Schwartz, J. *et al*. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition*, 93(2): B69-B78.

Stuart, A. & Kalinowski, J. (2015). Effect of delayed auditory feedback, speech rate, and sex on speech production. *Perceptual & Motor Skills: Learning & Memory*, 120(3): 747-765.

Turgeon, C. *et al*., (2015). Exploring consequences of short- and long-term deafness on speech production: A lip-tube perturbation study. *Clinical Linguistics & Phonetics*, 29(5): 378-400.

Tye-Murray, N. *et al*. (2014). The self-advantage in visual speech processing enhances audiovisual speech recognition in noise. *Psyhonomic Bull Revision.*

Tye-Murray, N. *et al*. (2012). Using patient perceptions of relative benefit and employment to assess auditory training. *Journal of the American Academy of Audiology*, 23(8): 623-634.

Xu, J. *et al*. (2014). Noise-rearing disrupts the maturation of multisensory integration. *European Journal of Neuroscience*, 39: 602-613.