AUDIOVISUAL INTEGRATION IN NATIVE AND NON-NATIVE SPEECH PERCEPTION

A Thesis Presented to
The Honors Tutorial College
Ohio University

In Partial Fulfillment
of the Requirements for Graduation
from The Honors Tutorial College
with the degree of
Bachelor of Science in Communication Sciences and Disorders

by
Margaret Elizabeth Harrison
April 2016
This thesis has been approved by

The Honors Tutorial College and Division of Communication Sciences and Disorders

________________________

Chao-Yang Lee
Associate Professor and Coordinator,
Communication Sciences and Disorders
Thesis Advisor

________________________

Chao-Yang Lee
Honors Tutorial College, DOS
Communication Sciences and Disorders

________________________

Jeremy Webster
Dean, Honors Tutorial College
Abstract

Lip rounding enhances the distinction between the fricative sounds /s/ (as in “save”) and /ʃ/ (as in “shave”) in English. However, this distinction is not common throughout world languages. How well are learners of English able to utilize lip rounding to distinguish these two sounds in their production and perception? Is the ability to utilize lip rounding in this distinction related to overall audiovisual integration in speech perception, as measured by the McGurk effect? Does this ability develop over time? Fourteen English language learners and 14 native speakers of English participated in three tasks to explore these questions. A fricative production task and fricative perception task were used to evaluate the ability to use lip rounding to distinguish /s/ and /ʃ/. A McGurk perceptual task was utilized to determine the degree of audiovisual integration in speech perception. Non-native speaker participants also completed a second session to evaluate development over time. Overall, the English language learners showed high proficiency in both production and perception, suggesting effective use of lip rounding in making the distinction. Development of perceptual ability also occurred between the two sessions. Reaction time patterns differed between the learners and native participants in all perception tasks, suggesting distinct strategies in dealing with audiovisual integration in speech perception.
ACKNOWLEDGMENTS

I would like to thank my thesis adviser Chao-Yang Lee for his dedication and guidance throughout this project. The idea for this thesis began with what I learned in his Phonetics and Speech Science classes, so it is only fitting that we accomplished this project together. His encouragement and patience have been invaluable throughout not only this project, but my overall undergraduate experience. This has truly been a team effort, and I couldn’t have asked for a better mentor.

I would also like to thank Seth Wiener for his contributions to the project. His assistance with Superlab 5, experimental design details, statistics, graphs, insights, and so much more have made this project all the more successful and enjoyable. Ken Dobo’s assistance with stimulus recording and creation—particularly his help with Final Cut Pro X—is also greatly appreciated.

A sincere thank-you goes out to the entire Division of Communication Sciences and Disorders at Ohio University. The faculty, staff, and my peers have played such an important role in this project through their support and encouragement. I would also like to thank the Ohio University English Language Improvement Program instructors for their assistance in recruiting participants.

I would not be where I am today without my family and friends. They have all been such a wonderful source of encouragement not only as I have completed this project, but throughout my entire life. I would especially like to thank my parents, who have lovingly shaped me into the person I am today—they are both such a blessing.
and I will be forever grateful for the principles of faith, love, and dedication that they have instilled in me through their examples. Finally, I would like to acknowledge Megan Smith and Haley Trottier—I am so glad we got to go through this thesis experience together.

This project was funded in part through a Provost’s Undergraduate Research Fund Grant from Ohio University and a College of Health Sciences and Professions Student Research Grant.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>3</td>
</tr>
<tr>
<td>List of Tables</td>
<td>7</td>
</tr>
<tr>
<td>List of Figures</td>
<td>8</td>
</tr>
<tr>
<td>Introduction</td>
<td>10</td>
</tr>
<tr>
<td>Fricative Consonants: Articulation and Acoustics</td>
<td>10</td>
</tr>
<tr>
<td>Second-Language Speech Acquisition</td>
<td>12</td>
</tr>
<tr>
<td>The Role of Visual Information in Speech Perception</td>
<td>15</td>
</tr>
<tr>
<td>Current Study</td>
<td>18</td>
</tr>
<tr>
<td>Methods</td>
<td>23</td>
</tr>
<tr>
<td>Participants</td>
<td>24</td>
</tr>
<tr>
<td>Materials</td>
<td>25</td>
</tr>
<tr>
<td>Procedure</td>
<td>27</td>
</tr>
<tr>
<td>Results</td>
<td>30</td>
</tr>
<tr>
<td>Fricative Production</td>
<td>30</td>
</tr>
<tr>
<td>Fricative Perception</td>
<td>33</td>
</tr>
<tr>
<td>Accuracy</td>
<td>34</td>
</tr>
<tr>
<td>Reaction Time</td>
<td>37</td>
</tr>
<tr>
<td>Correlation of Fricative Production and Fricative Perception</td>
<td>41</td>
</tr>
<tr>
<td>McGurk Task</td>
<td>43</td>
</tr>
<tr>
<td>Correlation of Fricative Perception and McGurk</td>
<td>48</td>
</tr>
<tr>
<td>Discussion</td>
<td>49</td>
</tr>
<tr>
<td>Limitations</td>
<td>54</td>
</tr>
</tbody>
</table>
Future Directions ................................................................. 54
Conclusion .............................................................................. 56
References ............................................................................. 57
Appendix A: 21 Fricative Word Pairs .................................... 60
Appendix B: Fricative Perception Task—On-Screen Instructions ........... 61
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Spectral Center of Gravity (in Hz with SD) Measured in Fricative Production Task</td>
<td>32</td>
</tr>
<tr>
<td>2.</td>
<td>Correlations between /s/-/ʃ/ Acoustic Difference and Accuracy Measures of Fricative Perception for Non-native Participants</td>
<td>41</td>
</tr>
<tr>
<td>3.</td>
<td>Correlations Between /s/-/ʃ/ Acoustic Difference and Reaction Time Measures of Fricative Perception for Non-native Participants</td>
<td>43</td>
</tr>
<tr>
<td>4.</td>
<td>Accuracy of Consonant Identification (in % with SD) in the McGurk Task</td>
<td>43</td>
</tr>
<tr>
<td>5.</td>
<td>Percentage of McGurk Responses</td>
<td>45</td>
</tr>
<tr>
<td>6.</td>
<td>Reaction Time of Consonant Identification (in ms with SD) in the McGurk Task</td>
<td>47</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.</td>
<td>Spectral slices of /s/ and /ʃ/</td>
<td>32</td>
</tr>
<tr>
<td>Figure 2.</td>
<td>Accuracy of fricative identification (by % with SE) in fricative perception task</td>
<td>35</td>
</tr>
<tr>
<td>Figure 3.</td>
<td>Reaction time data (in ms with SE) for correct fricative identification responses in fricative perception task</td>
<td>38</td>
</tr>
</tbody>
</table>
Introduction

Learning a second language is a challenging process. Each language has its own phonology, or set of sounds. A difference in phonological sets is one of the characteristics of language that make learning a new one difficult. Not only do second language learners have to master new arrangements of sounds they know, but, in some cases they also have to learn to articulate sounds they have never heard before and then use them correctly. For example, English makes use of 24 consonant sounds while Mandarin Chinese includes 22 consonants, many of which do not exist in both languages (American Speech-Language-Hearing Association, 2015).

Fricative Consonants: Articulation and Acoustics

Fricative consonants are produced when a narrow constriction is made by two articulators. As air is forced through the narrow opening, turbulent airstream is generated to produce a noisy quality. The English language includes eight fricative consonants. This relatively large number provides another example of cross-linguistic phonological differences. The two most commonly spoken world languages, Mandarin and Spanish, for example, utilize five and three fricatives respectively (American Speech-Language-Hearing Association, 2015).

The means by which speakers are able to produce so many acoustically distinct consonants from a turbulent airstream is best described by the source-filter theory of speech production (Müller, 1848; as cited in Lieberman & Blumstein, 1988).
According to this theory, the supralaryngeal vocal tract is essentially a uniform tube through which a sound source (laryngeal, noise, and/or transient) is passed and filtered, producing the acoustic characteristics of speech. In the case of fricatives, the source of sound is the turbulent airstream created by passing air through a narrow constriction in the vocal tract (Johnson, 2012). Additionally, the airstream may strike an obstacle in front of the constriction (i.e. lower teeth), which would amplify the sound, as is the case for sounds such as /s/. The place of constriction divides the supralaryngeal vocal tract into two cavities: the area of the vocal tract between the narrow constriction and the larynx is known as the back cavity, while the area between the narrow constriction and the lips is the front cavity. The front cavity most influences the acoustic properties of fricative consonants.

The resonant properties of the front cavity will determine the resulting acoustics of the sound source passed through it. Resonance, as it applies to human speech production, is the reinforcement, or amplification, of specific frequencies of a sound source by the vocal tract. As all fricatives are produced with a turbulent airstream, it is the filtering of this sound source through the front cavity which produces distinct consonants. Place of articulation is the main difference in the physical production of fricative consonants. Movement of the place of articulation further back along the vocal tract results in a longer front cavity. As the front cavity is lengthened, the perceived pitch of the fricative consonant is lowered. This inverse relationship between front cavity length and the acoustic properties of fricatives is
expressed by the formula used to calculate resonant frequencies of the front cavity (Johnson, 2012):

$$F_n = \frac{(2n - 1)c}{4L}$$

In this equation, $F_n$ represents the $n$-th resonant frequency, $c$ is the speed of sound, and $L$ is the length of the front cavity. The first resonant frequency of the front cavity (i.e., $n = 1$) is most important for fricative consonants, as higher resonant frequencies are not relevant to human speech perception. So, for consideration of fricative consonant perception, the equation can be reduced to:

$$F_1 = \frac{c}{4L}$$

As the length of the front cavity increases, then, the first resonant frequency will decrease proportionally.

These predictions from the formula can be verified by measuring resonant frequencies on the power spectrum, which shows the distribution of acoustic energy along the frequency domain. Specifically, the first resonant frequency can be measured on the spectrum by identifying the most prominent spectral peak.

Second-Language Speech Acquisition

Traditional language pedagogy focuses on drilling specific sounds. On the surface, this appears to be the most effective approach as there will likely be new sounds to learn. The effectiveness of this approach in producing proficient, meaningful
communication, however, is unclear. To determine whether this is the most appropriate and successful approach, it is necessary to explore how sounds are produced and perceived, both as a whole and, specifically, in second language acquisition. Understanding these underlying processes could provide a knowledge-based instructional approach that will most benefit language learners. Second language acquisition literature has identified factors that contribute to success in language proficiency, such as similarities between the native and second language; however, the role of visual information in second language acquisition has not been fully explored.

While multiple theories of second language acquisition exist, most agree on at least two general principles: the interference, or imposition, of the native language on the second language and a connection between production and perception of the second language. Two models of second language speech learning, Best and colleagues’ Perceptual Assimilation Model (PAM; Best, 1995) and Flege and colleagues’ Speech Learning Model (SLM; Flege, 1995), are two such examples. These models differ in various aspects of second language acquisition theory, but agree that characteristics of the native language affect second language perception and production, although at times in different ways (as cited in Cutler, 2012 & Flege, 2003).

Best’s PAM focuses on the way in which sounds of the second language are assimilated to the learner’s existing phonemic categories, which were developed for the native language (as cited in Cutler, 2012). This model suggests six possible outcomes for a second language sound. According to this model, second language
sounds which are assimilated to separate phonemic categories of the native language will be more accurately discriminated than multiple sounds which are assimilated to a single native category. How well a sound fits into a native category is also suggested to influence a learner’s ability to discriminate sounds. Second language sounds which fit native categories equally well are expected to be more difficult to distinguish than sounds that differ in degree of fit (as cited in Cutler, 2012 & Flege, 2003).

Flege’s SLM is based, in part, on a similar foundation: “…bilinguals cannot fully separate the L1 and L2 phonetic subsystems.” (Flege, 2003, p. 8) This model views second language acquisition as a learning process that takes time and significant native speaker input which shapes and changes perception and production of the second language sounds. The components of the SLM model agree with and are supported by suggestions from other researchers in regards to the influence of feature weighting on second language processing and acquisition. Researchers such as Trubetzkoy (1939/1958; as cited in Flege, 2003) and Michaels (1974; as cited in Flege) have suggested that different languages assign different levels of importance to distinctive features of speech sounds and that these language-specific perceptual “filters” influence second language perception. This perceptual interference is likely to carry over to production accuracy. As Flege suggests, “…production of an L2 phonetic segment will typically be no more native-like than its perceptual representation and might, in early stages of learning, be less native-like” (Flege, 2003, p.4). Research such as that by Gottfried and Beddor (1988; as cited in Flege 2003), which found that advanced second language speakers closely resembled native
speakers in perception of second language sounds and led to the suggestion that “L2 learning might result in a change in feature weighting” (as cited in Flege, 2003, p. 9), supports the notion of the SLM that speech learning slowly becomes more successful over time, becoming more native-like and less constrained by the native language.

*The Role of Visual Information in Speech Perception*

These models of second language acquisition focus on auditory distinction of phonemes; however, speech perception is not a purely auditory process. As Harry McGurk and John MacDonald suggest in their 1976 research article, humans can “hear lips and see voices” (McGurk & Macdonald, 1976). At the time, this was a rather novel idea. McGurk and MacDonald found that dubbing a sound track of the syllable /ba/ onto a video of a woman saying /ga/ resulted in participants perceiving /da/. When the audio and visual syllables were reversed (audio-/ga/, visual-/ba/), some participants reported hearing /bga/. These findings, now known as the “McGurk effect”, demonstrate that visual cues influence speech perception in humans.

Since this discovery, researchers have continued to work towards an accurate model of human speech perception that incorporates both auditory and visual information. Several theories have been proposed and tested, but a final conclusion has yet to be drawn. What has been proven by this work is that visual information undoubtedly influences speech perception. The McGurk-MacDonald (1976) study has been replicated several times with similar, and therefore reliable, results. Researchers have also altered the details of the original study to explore the effect of aspects such
as talker gender differences (Green, Kuhl, Meltzoff, & Stevens, 1991), individual differences among listeners (Strand, Cooperman, Rowe, & Simenstad, 2014), vowel context (Shigeno, 2000), foreign language effects (Sekiyama & Tohkura, 1991; Sekiyama, 1997), speech production knowledge (D’Ausilio, Bartoli, Maffongelli, Berry, & Fadiga, 2014), and more, on the magnitude of the McGurk effect.

While many of these studies have found compelling results that add to the greater understanding of human speech perception, the findings regarding foreign language effects and speech production knowledge are most important to the current study. Studies on the McGurk effect have shown that speakers of certain languages show a reduced McGurk effect when tested. Sekiyama found that speakers of both Japanese (Sekiyama & Tohkura, 1991) and Chinese (Sekiyama, 1997) exhibit reduced levels of McGurk effect when compared with speakers of American English. The authors offered suggestions as to why this is the case, many of which are culturally based; however, this study raises several questions, as the influence of culture on audiovisual speech perception has not been thoroughly explored.

Taking the idea of influence of visual information on speech perception beyond visible articulatory cues, D’Ausilio, Bartoli, Maffongelli, Berry, and Fadiga discovered in their 2014 study that showing participants sagittal (side) views of tongue movements during speech influenced perceptual responses much like standard visual cues do in classic McGurk studies. Speech sounds were presented to participants along with congruent, incongruent, or meaningless control visuals of the tongue; participants were asked only to identify what they heard. Results reported that while identification
accuracy was high in all conditions, response was quickest when the audio and visual presentations were congruent. From these results, the researchers concluded, “visual experience of articulatory movements is not necessary for the generation of audiovisual mismatch effects” (D’Ausilio et al., 2014). These findings led the authors to suggest that, “audiovisual integration during speech may benefit from speech production learning,” (D’Ausilio et al., 2014). These results further support a strong connection between the abilities to produce and perceive spoken language, while adding a further detail relevant to second language learners: perception can be influenced even by unfamiliar visual cues.

Since visual information has been proven to influence speech perception of (at least) speakers of English, one might question the role of visual information in learning English as a second language. As Sekiyama and colleagues discovered, visual information does not seem to be equally influential in speech perception across languages and/or cultures. As second language acquisition research has suggested, not all languages weight distinguishing characteristics, whether auditory or visual in nature, in the same ways. It could be that for some languages, visual cues hold less valuable, or distinctive, information than do other sources, such as tone for speakers of tonal languages. For a language such as English, which utilizes several different sounds rather than multiple tones to distinguish meaning, correctly perceiving several sounds as distinct phonemes is crucial to successful communication. In early stages of language learning when, according to the SLM, phoneme mapping and distinction is most heavily influenced by the native language, it may be helpful for English language
learners to recognize salient visual characteristics while learning to recognize auditory distinctions. These visual cues also provide insight for correct articulation and, therefore, potential for improved production of new sounds.

**Current Study**

As the role of visual information in second language acquisition is yet undetermined, the overarching goal of the current study is to evaluate the role of visual information in speech perception by non-native speakers learning English as a second language. This will be accomplished by using one sound distinction found in English but not all major world languages: the contrast between /s/ (the first sound in “save”) and /ʃ/ (the first sound in “shave”). These two fricative consonants provide an opportunity to study audiovisual integration, as they differ not only acoustically, but also visually. The most noticeable difference acoustically is that the frication noise for /s/ has a higher frequency than /ʃ/. Along with a difference in place of articulation, this is accomplished in part by the articulatory action of lip rounding in the production of /ʃ/. Lip rounding lengthens the front resonant cavity of this fricative consonant, therefore further lowering the resonant frequency of the frication noise. This lip rounding also serves to provide a salient visual cue which is specific to /ʃ/, therefore further distinguishing this fricative from /s/.

Exploiting this difference will allow conclusions to be drawn regarding the use of visual information in both speech production and perception. Results from this study will contribute to further understanding of the role of visual information in
speech production and perception, as well as the process and contributing factors of second language acquisition. From this knowledge, more effective language pedagogy may be developed, allowing language students to become more effective, successful communicators.

To achieve this goal, this study will address two specific research questions. The first question to be explored is: Is the ability to attend to visual information in speech perception related to the ability to produce speech sounds? It is hypothesized that sensitivity to visual information is associated with the ability to produce the /s/-/ʃ/ distinction. In other words, participants who are more sensitive to and aware of lip rounding during speech perception are more likely to use this lip rounding in their own production, and will show a more accurate /s/-/ʃ/ distinction. As discussed previously, the McGurk effect proved that visual information incongruent with auditory information can result in perceptions not represented by either source. Additionally, it has been found that visual cues never before experienced can influence perception of speech (D’Ausilio et al., 2014). If the visual information available from speech proves to affect non-native speakers of English in a similar way, this concept can be used to inform improved language instruction strategies.

The second question to be addressed by this study is: Is the magnitude of McGurk effect experienced by participants associated with their ability to produce and perceive the /s/-/ʃ/ distinction? While magnitude of McGurk effect has proven to be a robust measure of audiovisual perception, studies such as those by Sekiyama and colleagues (1991, 1997) have shown that magnitude depends on the participant’s
native language; in both studies, it was found that Chinese and Japanese participants experienced a significantly decreased magnitude compared to participants whose native language was American English. These findings suggest that the use of visual information in speech perception may be modulated by language and/or cultural background. As non-native speakers of English come from various language and cultural backgrounds, it is of interest to explore whether magnitude of McGurk effect can be used to predict ability to produce the /s/-/ʃ/ distinction. As magnitude of McGurk effect measures audiovisual perception, it may be that this magnitude could indicate to what degree visual information is utilized to differentiate /s/ and /ʃ/. In the case of this fricative distinction, this would mean that the visual cue of lip rounding is recognized and used to a greater extent by those who exhibit higher magnitudes of McGurk effect. It is hypothesized that magnitude of McGurk effect is positively associated with the ability to perceive and produce the /s/-/ʃ/ distinction. A higher magnitude of McGurk effect is expected to correlate with greater awareness of lip rounding as a salient characteristic of /ʃ/ and therefore more prominent use of lip rounding during speech production, as well.

These questions are explored using three types of tasks: fricative production, fricative perception, and McGurk; both native and non-native speakers of English will be tested. In the fricative production task, participants will be recorded while reading word pairs which differ in initial sound—either /s/ or /ʃ/. These recordings will be acoustically analyzed to determine how well participants are able to distinguish between the two fricatives during speech production. Native speaker data serves as the
target frequency range for each fricative and the overall difference or distinction between the two sounds. Non-native speaker participants’ data will be compared to the native speaker data in order to determine accuracy of production, with more native-like production being considered more accurate. As discussed previously, the English fricatives /s/ and /ʃ/ differ acoustically and visually, with lip rounding serving as a significant factor in both types of distinction. A lack of significant frequency distinction between the two sounds likely indicates that the speaker is not rounding his or her lips during production.

Fricative perception will be tested using four conditions. An audio-only condition will be utilized to determine how well participants are able to acoustically distinguish between /s/ and /ʃ/ (in the absence of any visual input). The audiovisual-congruent condition will evaluate fricative perceptual ability when provided with both auditory and visual information. An audio-visual incongruent condition will also be used to measure what effect, if any, incongruent visual information has on fricative perception. Finally, a video-only (or visual-only) condition will be presented to participants to determine their awareness and utilization of the characteristic of lip-rounding in distinguishing /s/ from /ʃ/. The McGurk task will be used to measure magnitude of McGurk experienced by each participant. Both accuracy and reaction time will be collected for the fricative production and McGurk tasks.

Native speakers of English are expected to generate high accuracy in both fricative production and fricative perception tasks. Specifically, responses to the audio-only and audiovisual-congruent conditions are predicted to be the most accurate
and the quickest. Responses to visual-only stimuli are predicted to be accurate because of participants’ knowledge of the association between lip rounding and the /s/-/ʃ/ distinction, but may have slower reaction times. Reaction time to visual-only stimuli may suffer because of lack of auditory information. Although responses are predicted to remain accurate, increased reaction times are also predicted in the audiovisual-incongruent condition because of the interference of the incongruent visual signal. This follows the results of the 2014 study by D’Ausilio et al. discussed previously.

Non-native speaker participants would be tested twice over the course of one college semester. This is done with the understanding that performance may change over time. During this time, participants would be involved in English language improvement courses and immersed in an environment—an American college campus—in which English is the predominant language. It is predicted that performance in both production and perception will improve in the second session, after receiving continued direct English instruction and further native speaker input. Specific predictions for each of the fricative perception conditions differ slightly from those for the native participants. In general, more variance in results, as modulated by perceptual ability, is predicted for the non-native speaker group because they would still be in a period of development. The results overall are predicted to be less accurate, to varying degrees, than the native speaker data because non-native speaker participants are still in the process of mastering the English language. Slower reaction times are also predicted, as English is not the native language and is therefore less natural perceptually for these participants. As discussed in the first hypothesis,
participants who are able to more accurately distinguish between the fricatives during speech production are more likely to respond accurately and with quicker reaction times to perception tasks. The main difference for this group of participants is expected to occur in the reaction time data for the audio-visual incongruent condition. While a difference in reaction time (that is, slower than the other 3 conditions) is expected for native participants, this difference is not predicted to be present for all non-native participants. Those who are not utilizing visual information to accurately distinguish /s/ and /ʃ/ are expected to display consistent reaction times in both the audiovisual-congruent and –incongruent conditions, because the visual difference present in the audiovisual-incongruent stimuli would presumably not be noticed.

McGurk task results are expected to follow those of previous McGurk studies. That is, native magnitudes are expected, overall, to be higher than those of the non-native speaker group, as McGurk magnitude has been found to be reduced in speakers of certain languages other than English. If the second hypothesis holds true, non-native speaker participants who exhibit higher magnitudes of McGurk effect are expected to also be those whose performance in the fricative perception tasks indicate use of lip rounding to make the /s/-/ʃ/ distinction.

Method

This project utilized one production task and four perceptual tasks. An objective of the current study is to explore speech production and perception in a way which will allow clear, applicable conclusions to be drawn regarding English language
learning; details of this study were developed with this in mind. Considerations for non-native speakers (i.e. limited English vocabulary, limited spelling abilities, etc.) were significant factors in the development of stimuli and tasks.

Participants

Two groups of adult participants were recruited from Ohio University students. The experimental group includes 14 non-native speakers of English who were enrolled in an English Language Improvement Program (ELIP) oral communication course at Ohio University at the time of their participation. The non-native speaker group consisted of 7 males and 7 females between the ages of 22 and 33 \((M = 26.6, SD = 2.7)\). The control group consisted of 14 native speakers of English. This group also consisted of 7 males and 7 females, between the ages of 19 and 35 \((M = 22.7, SD = 4.9)\). The experimental group participated in the experiment twice in the semester in order to evaluate the development of their performance. The second test session occurred within approximately 4 to 9 weeks of the first session \((M = 47 \text{ days}; SD = 10 \text{ days})\). Between the two test sessions, these participants received English instruction through the ELIP class and were immersed in an environment where English is the predominant language. The control group participated only once because, as native speakers, their performance is not expected to change over time.

Within the experimental group, there were differences in language background. Native languages reported include Arabic \((n = 1)\), Bengali \((n = 2)\), Chinese \((n = 4)\), Farsi \((n = 4)\), Nepali \((n = 1)\), and Thai \((n = 2)\). Of the 14 non-native speaker
participants, 5 reported native languages with phonologies that include the /ʃ/ sound (Speech Accent Archive). For the remaining 9 participants, the /s/-/ʃ/ distinction is not present in their native phonologies.

**Materials**

For the fricative production and fricative perception tasks, 21 pairs of single syllable words contrasting in initial /s/-/ʃ/ (e.g., “save”-“shave”) were selected (see Appendix A for complete list of word pairs). Considerations for non-native speaker participants such as word frequency, spelling, etc. were influential in the development of the stimulus set. Fricative pairs included front and mid vowels. For this study, back vowels were excluded due to considerations of co-articulation. That is, when followed by a rounded back vowel, the fricative /s/ may appear rounded, making it less distinct visually from /ʃ/.

For the McGurk task, the stop consonants /b/ and /g/ were paired with the vowels /a/ and /i/, resulting in four syllables: /ba/, /ga/, /bi/, and /gi/. Two different vowels were chosen because Shigeno (2000) showed that vowel context could affect the magnitude of McGurk effect. Using two vowels allows exploration of any possible vowel effect on measured magnitude of McGurk.

A female native speaker of American English was audio-video recorded reading the stimuli in a sound proof booth with a video recorder (Canon Vixia HF G20 HD) and an omnidirectional microphone (Electro-Voice) at 48 kHz sampling rate and 16-bit quantization. The recordings were edited using the program Final Cut Pro X on
a MacBook Air computer. Two experimental tasks, fricative perception and McGurk, were created from this recording.

Specifically, four conditions were created for the fricative perception task: audiovisual congruent, audiovisual incongruent, audio-only, and video-only. Each individual word was cut from the recording and trimmed to approximately 3 seconds long, with one second before the fricative onset, then the stimulus word and time remaining to reach 3 seconds. Audiovisual congruent, audio-only, and video-only conditions were saved from the same formatted file. The audiovisual incongruent condition was made by dubbing the audio of one word onto the audiovisual congruent stimulus of its minimal pair (e.g., audio of “save” dubbed onto video of “shave”). The audio was detached from the audiovisual file and the vowels of the two audio components were aligned. (The fricative-vowel boundary provided a reliable point of alignment for the incongruent stimuli, as this is the point where articulation of the word pairs becomes relatively identical; this allowed a seamless presentation.) The audio portion of the audiovisual file was then disabled, leaving the video of one stimulus with the aligned audio of its minimal pair. The McGurk condition was created in the same way, by dubbing syllables onto their same-vowel counterparts (e.g., audio-/ba/ with video-/ga/). The audio files for these stimuli were aligned at the stop consonant burst. Following the manipulation of the original recordings, the stimulus set included the four original syllables /ba/, /ga/, /bi/, and /gi/ and four incongruent syllables: audio-/ba/ with video-/ga/, audio-/bi/ with video-/gi/, audio-/ga/ with video-/ba/, and audio-/gi/ with video-/bi/.
The program SuperLab 5 (Cedrus Corporation, 2014) was used to create the experimental presentation of the edited stimuli and collect response and reaction time data. A PowerPoint presentation of the 21 word pairs was also created for presentation of words during the production task. Pairs of words were presented on one slide to highlight the difference in spelling. All stimuli were presented using a 21-inch iMac computer.

Procedure

This study was conducted in the sound proof booth in Grover W241. An FM wireless intercom was utilized to allow the experimenter to hear and monitor participation and conditions within the sound-proof room. Once participants completed the consent process, basic demographic information was collected on age, sex, and languages spoken. A simple hearing screening, to determine normal hearing, was also conducted prior to task completion. Normal hearing for this study was defined as pure-tone, air-conducted thresholds of \( \leq 20 \) dB HL at octave frequencies from 1000 to 4000 Hz. Participants who passed the hearing screening then completed the production and perceptual tasks in the following order. One participant failed the hearing screening; no data was collected for this participant.

Fricative production task. Participants read aloud the 42 words presented in a PowerPoint presentation via an iMac computer. The transition between slides was controlled by the experimenter; slides were changed once the participant had read both words presented on the slide. Participants were audio recorded using a Sony PCM-
M10 portable audio recorder as well as audio-video recorded using the computer’s built-in camera. Words were presented in minimal pairs (differ in only one sound; in this case initial /s/ or /ʃ/) to highlight the difference in the fricative contrast.

**Fricative perception task.** Participants were presented with the recordings of the 42 fricative words in three conditions: audiovisual, audio-only, and visual-only. Conditions were presented through the program SuperLab 5 via an iMac computer. Participants were given a general introduction and instructions in regards to the perception tasks, similar to the following script:

“The main question for this part of the study is, “What sound does the word start with?” You will be presented with different tasks, which each ask this same question. For each task, you will see instructions, which are helpful to read because there are only slight differences between the tasks. You will also have some practice trials before the test session begins. As you can see, your options for this part of the experiment are /s/ as in “save” and /ʃ/ as in “shave”. When the instructions ask you to press the space bar to begin, use the one on the keyboard; otherwise, only use the labeled buttons to respond. Please remember to keep your attention on the screen throughout each of the tasks.”

Participants were also informed that they were free to ask questions at any point during the study for clarification of instructions. Specific instructions for each individual task were presented for participants to read before the task began (see Appendix B). Ten practice trials of each condition were also presented before data collection began to familiarize participants with the condition (these practice trials utilized single syllable, /s/- or /ʃ/-initial words not present in the stimuli set).

The audiovisual condition included congruent and incongruent presentations as discussed in the “Stimuli” section above, for a total of 84 tokens. In the audio-only
condition, participants were presented with the audio portion of the 42 fricative words, and were presented with a “+” symbol on the screen to maintain attention as a standard practice in psychological research. For the video-only condition, participants were presented with the video portion of the 42 fricative word recordings without sound. In each of these tasks, participants were asked to respond whether the word presented began with /s/ or /ʃ/ by pressing the corresponding button labeled with /s/ or /ʃ/ on a response pad (Cedrus RB series). The order of presentation for the fricative perception tasks was counterbalanced across participants by the use of six presentation order groups. The placements of the /s/ and /ʃ/ buttons on the response pad were counterbalanced as well. Within each condition, the presentation of the tokens was randomized for each participant, meaning each participant received a unique, randomized order of presentation.

**McGurk task.** Participants were presented with the McGurk stimuli described in the “Stimuli” section above via the iMac computer and SuperLab 5 program used in previous tasks.

Before this task began, the experimenter changed the labels on the response pad to “b”, “d”, and “g”. Participants received oral instructions similar to the following:

“This is the last task. The basic idea is the same, “what sound does it start with?” however for this task you will hear syllables instead of words. As you can see, your options are now “b” as in “boy”, “d” as in “dog”, and “g” as in “girl”. Everything else is the same as before—use only the labelled buttons to respond, and please keep your focus on the screen.”
The eight McGurk stimuli were presented three times each for a total of 24 tokens. The presentation of the tokens was randomized for each participant, meaning each participant received a unique, randomized order of presentation. Participants were asked to indicate whether the syllable presented began with /b/, /d/, or /g/ by pressing the corresponding button labeled “b”, “d”, or “g” on the same response pad used in previous tasks. As in previous tasks, instructions were presented for participants to read and practice trials (4 congruent syllables) were presented before data collection began.

The second session of non-native speaker participant testing followed the same procedure as the first session (demographic data collection was excluded in the second session). Participants were presented with tasks in the same order as the first session. Positioning of /s/ and /ʃ/ on the response pad remained the same, as well.

Results

Fricative Production

Data from the fricative production task was analyzed acoustically. In particular, the spectral center of gravity was obtained using Praat (Boersma & Weenink, 2016) for each fricative produced. Spectral center of gravity is the measure of the central frequency of a sound’s energy distribution. In other words, this measure is calculated using the energy distribution of a sound to determine the “average” frequency. To produce a fricative, a narrow constriction is made in the vocal tract and air is forced through this constriction. This forcing of air causes frication, which
produces the perceived sound. As the point of constriction moves further back in the vocal tract, the front resonant cavity (the space between the point of constriction and the end of the vocal tract) becomes longer. Lip rounding also lengthens the front resonant cavity, as discussed previously. When passed through a longer front resonant cavity, the lower-frequency components of a sound are amplified. This amplification means that the sound’s acoustic energy is more concentrated around lower frequencies, which will result in a lower spectral center of gravity than that measured in sounds produced with short front resonant cavities. Therefore, accurate productions in the Fricative Production task would be indicated by the average spectral center of gravity for /ʃ/ being lower than that of /s/.

To obtain this measure, acoustic analysis was conducted on audio recordings from the built-in microphone of the iMac. The audio files were extracted using Final Cut Pro X. Praat (Boersma & Weenink, 2016) was then used to resample the audio files to 22.05 kHz, remove noise using Praat’s default settings, and finally convert the files from stereo to mono. For each fricative, a spectral slice was taken from approximately the middle of the frication noise (with a window size of 0.025 second) and the program’s Centre of Gravity function was utilized to obtain the acoustic measure. These numbers were then used to find the average center of gravity for each fricative and the average difference between the two fricatives (average /s/-average /ʃ/) for each participant.
Figure 1. Spectral slices of /s/ and /ʃ/ sampled from native Fricative Production task, obtained from Praat. Energy (amplitude) of /s/ is more concentrated in the high-frequency range compared to the energy distribution of /ʃ/; this results in a higher spectral center of gravity measured for /s/.

Results of the fricative production task analysis are summarized in Table 1, which shows the average spectral centers of gravity for each fricative (/s/ and /ʃ/) as well as differences in spectral centers of gravity between the two fricatives for the native participants and non-native participants in sessions 1 and 2. The substantial separations between /s/ and /ʃ/ indicate that all participants (native, non-native session 1, and non-native session 2) were able to accurately produce the /s/-/ʃ/ distinction. Further, no significant difference in native and non-native production was observed, indicating that non-native participants were able to produce the distinction as proficiently as native speakers. Results from the second session actually indicate that
the non-native participants were able to make an even greater acoustic distinction
between /s/ and /ʃ/ than did the native speakers.

Table 1

*Spectral Center of Gravity (in Hz with SD) Measured in Fricative Production Task*

<table>
<thead>
<tr>
<th>Group/Session</th>
<th>Fricative</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>/s/</td>
<td>/ʃ/</td>
</tr>
<tr>
<td>Native</td>
<td>8166 (824)</td>
<td>4406 (752)</td>
</tr>
<tr>
<td>Nonnative 1</td>
<td>8601 (651)</td>
<td>4580 (1247)</td>
</tr>
<tr>
<td>Nonnative 2</td>
<td>8694 (396)</td>
<td>4389 (747)</td>
</tr>
</tbody>
</table>

An analysis of variance (ANOVA) supports these observations. This ANOVA was conducted on spectral center of gravity with fricative (/s/, /ʃ/) as a within-subject factor, group (natives/non-natives in session 1/non-natives in session 2) as a between-subject factor, and participant as a random factor. The main effect of fricative is significant \[F(1, 39) = 671.712, p < .0001, \eta^2_p = .945\], supporting the finding that participants were able to accurately distinguish the two sounds in their production. A main effect of group was not found to be significant, indicating no difference across the three groups. These two findings suggest that non-native participants were able to proficiently produce a native-like distinction between /s/ and /ʃ/. The lack of fricative by group interaction suggests that the difference between groups is uniform between the two fricatives.

*Fricative Perception*

Accuracy of responses as well as reaction time data were collected during the fricative perception tasks. High accuracy was expected in the control group (native
speakers) for all conditions. Therefore, reaction time provides a more sensitive measure which may reveal differences not shown by accuracy data. Accuracy in the audio-only condition serves as a control factor by indicating whether or not the participant is able to identify /s/ and /ʃ/ based on the auditory signal alone. Influence of visual information during speech perception would be indicated by a decrease in accuracy and/or increase in reaction time from congruent to incongruent audio-visual stimulus responses based on the findings of D’Ausilio et al. (2014) discussed previously. Accuracy in the video-only condition will also indicate awareness of the visual cue of lip rounding, which distinguishes the English fricatives /s/ and /ʃ/.

**Accuracy.** Figure 2 summarizes the fricative identification accuracy results for the four conditions (audio-only, audiovisual-congruent, audiovisual-incongruent, and visual-only) of the fricative perception task. Results are divided by group: natives, non-natives in session 1, and non-natives in session 2.
The native participants responded with high accuracy in all four conditions. Responses were equally accurate between the two fricatives as well. Accuracy remained consistent for the audio-only, audiovisual-congruent, and audiovisual-incongruent conditions. This equality in accuracy across conditions suggests that this group was able to proficiently distinguish the /s/-/ʃ/ distinction based on the auditory
signal alone—visual information therefore did not enhance perceptual accuracy.

Further, native speakers were able to override incongruent information and correctly identify the fricatives in the audiovisual-incongruent condition. Taken as a whole, these observations indicate that when an auditory signal is present, visual information does not significantly aid or impair the perception of /s/ and /ʃ/ for this group.

This nearly perfect performance does not carry over into the visual-only condition. Responses maintained high accuracy in this condition, but, not at the levels seen in the other three. This condition was expected to be more difficult, as communication rarely relies solely on visual speech cues. Based on this difference in results, it appears that the auditory signal holds more significance in speech perception for native speakers, though the visual signal does hold enough information to allow them to discriminate /s/ and /ʃ/ accurately.

Non-native participants also responded with high accuracy in all 4 fricative perception conditions. As expected, participants demonstrated highest accuracy in the audio-only and audio-visual congruent conditions. These findings remained consistent across both sessions. As seen in Figure 2, accuracy appears to have improved in session two; this observation, however, is not supported by statistical analysis. In session one, accuracy for both /s/ and /ʃ/ stimuli were similar in the audio-only and audio-visual congruent conditions. In the audio-visual incongruent and video-only conditions, there is a visible separation between fricatives, with higher accuracy for /ʃ/ in both conditions (Figure 2). In the second session, /ʃ/ stimuli received more accurate responses in all four conditions than did /s/ stimuli. Overall accuracy improved from
session one to session two, with the exception of /s/-audio-only stimuli, though not by statistically significant amounts. This overall increase in accuracy suggests that within the time between session 1 and session 2, non-native participants showed evidence of development of perceptual abilities.

To statistically evaluate these observations, accuracy data from the fricative perception task was analyzed using an ANOVA with presentation (audio-only, audiovisual-congruent, audiovisual-incongruent, and visual-only) and fricative (/s/, /ʃ/) as within-subject factors, group (native, non-native session 1, non-native session 2) as the between subject factor, and participant as a random factor. The ANOVA revealed a main effect of presentation \([F(3, 117) = 9.738, \ p < .0001, \ \eta^2_p = .200]\). Pair-wise means comparisons with Bonferroni corrections show significant differences between audio-only and visual-only conditions \(p=.004\) and between audio-visual incongruent and visual-only conditions \(p<.0001\). The difference between audio-visual incongruent and visual-only conditions approached significance \(p = .055\), as well. These findings further support the conclusion that the auditory signal holds the strongest source of identification for participants. Additionally, a main effect of fricative is significant \([F(1,39) = 4.080, \ p = .05, \ \eta^2_p = .095]\). As shown in Figure 2, the source of the accuracy difference between /s/ and /ʃ/ is most likely attributed to the the non-native participant’s higher accuracy for /ʃ/ stimuli.

**Reaction time.** Figure 3 summarizes the reaction time data for the four conditions (audio-only, audiovisual-congruent, audiovisual-incongruent, and visual-only) of the fricative perception task. Only reaction times for correct response were
used in calculations. Results are divided by group: natives, non-natives in session 1, and non-natives in session 2.

![Figure 3. Reaction time data (in ms with SE) for correct fricative identification responses in fricative perception task in four conditions: audio-only, audiovisual-congruent, audiovisual-incongruent, and visual-only, by native participants (n=14) and non-native participants session 1 (n=14) and session 2 (n=14).](image)

Native results for reaction time complement those observed in the accuracy data. Specifically, reaction times remained consistent across the audio-only, audiovisual-congruent, and audiovisual-incongruent conditions. Reaction times
deviated only in the visual-only condition, with a noticeable increase. The consistency of reaction time across the two audiovisual conditions further indicates that for native speakers, discrimination ability is not compromised by incongruent visual information. Unlike previous results, a difference in reaction times between fricatives exists, with responses to /ʃ/ stimuli showing to be faster than those to /s/ stimuli. This difference in perceptual speed suggests that while the visual cue of lip rounding may not affect native listener’s accuracy in perceiving the /s/-/ʃ/ distinction, it may serve to aid in recognizing /ʃ/ more quickly.

In the experimental group, reaction time data revealed changes across sessions. In the first session, reaction times for /ʃ/-stimuli were overall shorter than those for /s/-stimuli, as observed in the native results. Across conditions, data for both fricatives followed similar patterns. In the second session, reaction time data between the two fricatives showed a much smaller gap. Reaction times were greatest for the audiovisual incongruent condition in both sessions, indicating that the incongruent visual cues had some impact on perceptual speed for the non-native speaker participants. In session one, the shortest reaction time was that for the audio-only condition.

Notably, the shortest reaction time in the non-native participant’s second session was that for the video-only condition. This contrasts sharply with native results, indicating differences in perception in at least this condition between the two groups. For the native group, the visual-only condition resulted in the lowest accuracy and longest reaction times. In the second non-native session, short reaction times in the visual-only condition occur with lower accuracy, as well. Both groups show
differences between the two fricatives in accuracy and reaction time for this condition, with /ʃ/ stimuli receiving more accurate responses and overall quicker responses. These findings suggest that the lip rounding present in the production of /ʃ/ allows quicker, more accurate identification overall in the absence of auditory information. The difference observed between native and non-native (session 2) reaction times to visual-only conditions suggest that the non-native group may be utilizing this signal to a greater degree in overall perception, which then allowed them to utilize this heightened awareness to more quickly identify the distinction.

To statistically evaluate these observations, reaction time of fricative identification was analyzed using an ANOVA in which presentation (audio-only, audiovisual-congruent, audiovisual-incongruent, visual-only) and fricative (/s//ʃ/) were within-subject factors, group (native/non-native session 1/non-native session 2) was a between-subject factor, and participant was a random factor. The main effect of fricative was found to be significant \([F(1, 39) = 20.514, p < .0001, \eta_p^2 = .345]\], supporting the interpretation that the visual cue of lip rounding serves to decrease reaction time in the perception of /ʃ/ compared to /s/. The interaction between fricative and presentation was also significant \([F(3, 117) = 3.072, p = .031, \eta_p^2 = .073]\), suggesting that the visual information available for /ʃ/ enhanced reaction times for certain presentation conditions. In other words, lip rounding is likely responsible for the difference in reaction times between /s/ and /ʃ/, particularly in the visual-only condition.
Correlation of Fricative Production and Fricative Perception

It was predicted that the ability to utilize the visual signal of lip rounding in perception of the /s/-/ʃ/ would be associated with the ability to use lip rounding to accurately produce this distinction. Pearson’s correlation coefficients were calculated between performance measures of fricative production (difference in spectral center of gravity between the two fricatives) and the eight measures of fricative perception (4 presentations by 2 fricatives) to test this hypothesis. These calculations were completed for accuracy and reaction time, as well as each participant group (native, non-native session 1, non-native session 2). Table 2 shows the correlation coefficients and the results of significance tests.

Table 2

<table>
<thead>
<tr>
<th>Presentation</th>
<th>Session 1</th>
<th></th>
<th></th>
<th>Session 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audio-only /s/</td>
<td>.835</td>
<td>&lt; .0001</td>
<td>.123</td>
<td>.676</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audio-only /ʃ/</td>
<td>.853</td>
<td>&lt; .0001</td>
<td>.495</td>
<td>.072</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AV congruent /s/</td>
<td>.860</td>
<td>&lt; .0001</td>
<td>.184</td>
<td>.529</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AV congruent /ʃ/</td>
<td>.723</td>
<td>.0003</td>
<td>.326</td>
<td>.255</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AV incongruent /s/</td>
<td>.652</td>
<td>.011</td>
<td>.087</td>
<td>.767</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AV incongruent /ʃ/</td>
<td>.448</td>
<td>.108</td>
<td>.034</td>
<td>.908</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video-only /s/</td>
<td>.784</td>
<td>.001</td>
<td>.418</td>
<td>.137</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video-only /ʃ/</td>
<td>.879</td>
<td>&lt; .0001</td>
<td>.316</td>
<td>.271</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. n = 14

For the native participant group there were no significant correlations between production and accuracy of perception for either accuracy or reaction time. For the non-native participants, there were significant correlations in session 1, but no
significant correlations in session 2. The absence of significant correlations observed in session 2 may be due to development of production and perception by the non-native participants between the two sessions. As the native speaker participants showed no correlations between the ability to produce and perceive the distinction, it is likely that as non-native participant’s abilities further develop, reaching more native-like levels of proficiency, correlations will be lost, matching the native results.

As in the accuracy results, no correlations were found between production measures and reaction time measures of perception for the native participants. For the non-native participants, the reaction time data revealed several significant correlations, shown in Table 3. In session 1, the audio-only /s/ presentation correlation approached significance \((r = -.521, p = .056, n = 14)\). In session 2, this presentation reached significance \((r = -.593, p = .026, n = 14)\) and several others approached significance, including audio-only /ʃ/ \((r = -.461, p = .097, n = 14)\), video-only /s/ \((r = -.507, p = .064, n = 14)\), and video-only /ʃ/ \((r = -.498, p = .070, n = 14)\). Interestingly, these conditions, audio-only and video-only, contain the least amount of information available across conditions (i.e. each condition only contains one source of information). It may be that over time, English language learners become better able to identify the /s/-/ʃ/ distinction based on one signal alone. Put another way, experience with accurately producing the fricative sounds, which includes utilizing lip rounding in the production of /ʃ/ and provides examples of correct acoustic signals, may have allowed the non-native participants to more quickly identify the accurately produced fricatives based on only one signal.
Table 3

**Correlations Between /s/-/ʃ/ Acoustic Difference and Reaction Time Measures of Fricative Perception for Non-native Participants**

<table>
<thead>
<tr>
<th>Presentation</th>
<th>Session 1</th>
<th></th>
<th></th>
<th>Session 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>r</td>
<td>P</td>
<td></td>
<td>r</td>
</tr>
<tr>
<td>Audio-only /s/</td>
<td>-.521</td>
<td>.056</td>
<td>-.593</td>
<td>.026</td>
<td></td>
</tr>
<tr>
<td>Audio-only /ʃ/</td>
<td>-.326</td>
<td>.255</td>
<td>-.461</td>
<td>.097</td>
<td></td>
</tr>
<tr>
<td>AV congruent /s/</td>
<td>-.274</td>
<td>.343</td>
<td>-.377</td>
<td>.185</td>
<td></td>
</tr>
<tr>
<td>AV congruent /ʃ/</td>
<td>-.260</td>
<td>.369</td>
<td>-.397</td>
<td>.160</td>
<td></td>
</tr>
<tr>
<td>AV incongruent /s/</td>
<td>-.268</td>
<td>.354</td>
<td>-.267</td>
<td>.357</td>
<td></td>
</tr>
<tr>
<td>AV incongruent /ʃ/</td>
<td>-.210</td>
<td>.470</td>
<td>-.335</td>
<td>.242</td>
<td></td>
</tr>
<tr>
<td>Video-only /s/</td>
<td>-.085</td>
<td>.772</td>
<td>-.507</td>
<td>.064</td>
<td></td>
</tr>
<tr>
<td>Video-only /ʃ/</td>
<td>-.275</td>
<td>.341</td>
<td>-.498</td>
<td>.070</td>
<td></td>
</tr>
</tbody>
</table>

*Note. n = 14*

McGurk Task

Table 4 summarizes accuracy of consonant identification in the McGurk task. Identification is highly accurate in all but two conditions: audio-/ba/ video-/ga/ and audio-/bi/ video-/gi/. The low accuracy in these two conditions is most likely due to presence of McGurk effect, i.e., /d/ response to audio /b/ - video /g/ stimuli.

Table 4

**Accuracy of Consonant Identification (in % with SD) in the McGurk Task.**

<table>
<thead>
<tr>
<th>Match</th>
<th>Native</th>
<th>Nonnative1</th>
<th>Nonnative2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AV congruent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audio /ba/ – video /ba/</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
</tr>
<tr>
<td>Audio /bi/ – video /bi/</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
</tr>
<tr>
<td>Audio /ga/ – video /ga/</td>
<td>100 (0)</td>
<td>100 (0)</td>
<td>100 (0)</td>
</tr>
<tr>
<td>Audio /gi/ – video /gi/</td>
<td>100 (0)</td>
<td>93 (19)</td>
<td>95 (12)</td>
</tr>
<tr>
<td>AV incongruent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audio /ba/ – video /ga/</td>
<td>64 (42)</td>
<td>7 (14)</td>
<td>24 (33)</td>
</tr>
<tr>
<td>Audio /bi/ – video /gi/</td>
<td>69 (42)</td>
<td>5 (12)</td>
<td>2 (9)</td>
</tr>
<tr>
<td>Audio /ga/ – video /ba/</td>
<td>100 (0)</td>
<td>94 (15)</td>
<td>88 (28)</td>
</tr>
<tr>
<td>Audio /gi/ – video /bi/</td>
<td>93 (14)</td>
<td>79 (34)</td>
<td>79 (31)</td>
</tr>
</tbody>
</table>

*Note. Conditions which could elicit McGurk response are reported in bold*
Native participants showed high accuracy across all conditions, even in the two conditions where McGurk response (and thus incorrect identification) was expected. The low percentage of McGurk response for this group of participants was unexpected. Non-native participants also showed high identification accuracy for each of the congruent presentations, establishing that they were able to perceptually distinguish between consonants in the syllable pairs /ba/-/ga/ and /bi/-/gi/. Responses to incongruent stimuli, which should elicit McGurk response, showed extremely low accuracy, suggesting a strong McGurk effect. This degree of visual influence in perception was not carried over to the other incongruent presentations (i.e., audio /g/ paired with video /b/), suggesting that in a situation where an implausible sound is perceived, (i.e. /bg/ is not a sound combination used in English), the information available in the auditory signal is most relied upon. Incongruent results as a whole, then, indicate that while non-native participants do utilize visual information in speech perception, the auditory signal still remains the predominant source of information. However, visual signals carry enough weight to produce realistic misperceptions (i.e. perception of /d/ in situations such as those caused by McGurk stimuli.

To statistically evaluate these observations, an ANOVA was conducted on accuracy of consonant identification with match (congruent/incongruent), consonant (/b//g/), and vowel (/a///i/) as within-subject factors, group (native/non-native session 1/non-native session 2) as a between-subject factor, and participant as a random factor. All four main effects were found to be statistically significant: match \[F(1, 39) = 250.254, \ p < .0001, \ \eta_p^2 = .865\], consonant \[F(1, 39) = 131.655, \ p < .0001, \ \eta_p^2 = .771\],
vowel \( F(1, 39) = 11.394, p = .002, \eta^2_p = .226 \), and group \( F(2, 39) = 18.162, p < .0001, \eta^2_p = .482 \). Significant interactions were also identified: match by group \( F(2, 39) = 18.475, p < .0001, \eta^2_p = .487 \), consonant by group \( F(2, 39) = 8.099, p < .0001, \eta^2_p = .293 \), match by consonant \( F(1, 39) = 183.701, p < .0001, \eta^2_p = .825 \), match by consonant by group \( F(2, 39) = 12.962, p < .0001, \eta^2_p = .399 \), and match by vowel \( F(1, 39) = 6.467, p = .015, \eta^2_p = .142 \).

While low consonant identification accuracy in incongruent McGurk conditions is likely an indication of McGurk effect, individual responses, specifically, responding /d/, to audio-/ba/ with video-/ga/ and audio-/bi/ with video-/gi/, must be examined to determine actual magnitude of McGurk effect. Table 5 shows the magnitude of McGurk effect for native participants and non-native participants in session 1 and session 2. Magnitude was determined for each participant by dividing the number of McGurk responses (/d/) by the total number of McGurk stimuli (audio-/b/ with visual-/g/ presentations).

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Response</th>
<th>Native</th>
<th>Nonnative1</th>
<th>Nonnative2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio /ba/ –</td>
<td>/da/</td>
<td>17 (28)</td>
<td>45 (43)</td>
<td>48 (45)</td>
</tr>
<tr>
<td>video /ga/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audio /bi/ –</td>
<td>/di/</td>
<td>29 (43)</td>
<td>88 (28)</td>
<td>88 (31)</td>
</tr>
<tr>
<td>video /gi/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A closer examination of responses in the McGurk task supports earlier observations. Low accuracy in consonant identification for audio-/b/ with video-/g/ corresponds to McGurk response of /d/ for these stimuli. Non-native participants show
a strong McGurk effect, which remains consistent across sessions. Native participants show relatively low magnitudes of McGurk effect. For both groups, presentations with the /i/ vowel elicited stronger McGurk effect than the vowel /a/. This finding is consistent with previous research by Shigeno (2000), on the effect of vowel context on perception of voiced stop consonants (e.g. /b/ and /g/).

To statistically evaluate these observations, McGurk-specific responses were analyzed using an ANOVA with vowel (/a//i/) as the within-subject factor, group (native/non-native session 1/non-native session 2) as the between-subject factor, and participant as a random factor. Main effects of vowel $[F(1, 39) = 22.366, p < .0001, \eta^2_p = .364]$ and group $[F(2, 39) = 10.357, p < .0001, \eta^2_p = .347]$ were found to be significant. Pair-wise means comparisons with Bonferroni corrections show significant differences between native and non-native session 1 groups ($p < .0001$) and native and non-native session 2 groups ($p < .0001$). The main effect of vowel supports the observation that the context of the vowel /i/ elicited stronger McGurk effect than did the vowel /a/. A main effect of group supports the difference in McGurk magnitudes observed between native and non-native participants. This presence of significantly stronger McGurk effect in non-native participants is in contrast to findings of previous studies in which speakers of languages other than American English showed reduced magnitude of McGurk effect when compared to their American counterparts (Sekiyama & Tohkura, 1991; Sekiyama, 1997).

Table 6 summarizes the reaction time data for consonant identification in the McGurk task for the natives, non-natives in session 1, and non-natives in session 2.
Reaction times are greater overall for incongruent presentations across all participant groups. This suggests an awareness of the incongruent visual signal, regardless of whether the response was accurate or inaccurate.

Table 6

*Reaction Time of Consonant Identification (in ms with SD) in the McGurk Task.*

<table>
<thead>
<tr>
<th>Match</th>
<th>Native</th>
<th>Nonnative1</th>
<th>Nonnative2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AV congruent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audio /ba/ – video /ba/</td>
<td>2053 (205)</td>
<td>2194 (314)</td>
<td>2246 (484)</td>
</tr>
<tr>
<td>Audio /bi/ – video /bi/</td>
<td>2053 (244)</td>
<td>2194 (190)</td>
<td>2184 (317)</td>
</tr>
<tr>
<td>Audio /ga/ – video /ga/</td>
<td>1981 (187)</td>
<td>2113 (216)</td>
<td>1953 (186)</td>
</tr>
<tr>
<td>Audio /gi/ – video /gi/</td>
<td>1994 (281)</td>
<td>2053 (240)</td>
<td>2036 (360)</td>
</tr>
<tr>
<td>AV incongruent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audio /ba/ – video /ga/</td>
<td>2865 (725)</td>
<td>2937 (596)</td>
<td>2798 (900)</td>
</tr>
<tr>
<td>Audio /bi/ – video /gi/</td>
<td>2628 (610)</td>
<td>2425 (508)</td>
<td>2160 (277)</td>
</tr>
<tr>
<td>Audio /ga/ – video /ba/</td>
<td>2349 (378)</td>
<td>2462 (591)</td>
<td>2507 (782)</td>
</tr>
<tr>
<td>Audio /gi/ – video /ba/</td>
<td>2364 (468)</td>
<td>2628 (865)</td>
<td>2511 (436)</td>
</tr>
</tbody>
</table>

An ANOVA was conducted on reaction time of consonant identification with match (congruent/incongruent), consonant (/b/ or /g/), and vowel (/a/ or /i/) as within-subject factors, group (native/nonnative1/nonnative2) as a between-subject factor, and participants as a random factor. This analysis indicated two significant main effects: match \(F(1, 39) = 86.241, p < .0001, \eta^2 = .689\) and consonant \(F(1, 39) = 7.776, p = .008, \eta^2 = .166\). The effect of vowel only approached significance \(F(1, 39) = 3.991, p = .053, \eta^2 = .093\). Significant interactions include match by consonant by group \(F(2, 39) = 3.667, p = .035, \eta^2 = .158\), match by vowel \(F(1, 39) = 6.353, p = .016, \eta^2 = .140\), consonant by vowel \(F(1, 39) = 11.515, p = .002, \eta^2 = .228\), and match by consonant by vowel \(F(1, 39) = 10.344, p = .003, \eta^2 = .210\).
Correlation of Fricative Perception and Magnitude of McGurk

It was predicted that the ability to perceive the /s/-/ʃ/ distinction would be associated with magnitude of McGurk effect. Results from the fricative perception task indicated that visual information can influence the perception of the /s/-/ʃ/ distinction. As the presence of McGurk effect depends on the use of visual information, which combines with the auditory signal to produce a new perceived signal, the presence of this effect was hypothesized to be related to audiovisual integration in fricative perception. Two sets of measures, one for fricative perception and one for McGurk effect, were derived to test this hypothesis.

Sensitivity to visual information in fricative perception is defined in this study as the difference in accuracy and reaction time between audiovisual-congruent and audiovisual-incongruent conditions. Based on D’Ausilio et al. (2014), which showed that incongruent visual information did not reduce accuracy but did increase response time, it was expected that reaction time would be a more sensitive measure than accuracy. In other words, an increase in reaction time would indicate recognition of the incongruent visual signal, even if that signal did not compromise fricative identification. Six measures were derived to determine sensitivity to visual information: AVC-AVI accuracy for /s/, AVC-AVI accuracy for /ʃ/, average accuracy of AVC-AVI for both fricatives, AVC-AVI reaction time for /s/, AVC-AVI response time for /ʃ/, and average reaction time of AVC-AVI for both fricatives. On the other hand, McGurk measures included the percent of /da/ responses to audio-/ba/

---

1 AVC = audiovisual-congruent, AVI = audiovisual-incongruent
with visual-/-ga/ stimuli, percent of /di/ responses to audio-/bi/ with visual-/gi/ stimuli, and overall percent of McGurk (/d/) responses to McGurk stimuli.

These McGurk and fricative perception measures were calculated for each participant. Pearson’s correlation coefficients were calculated between the two sets of measures, separately for each participant group (native/nonnative1/nonnative2). Fisher’s $r$ to $z$ transformation was then carried out to test whether the correlation coefficients were significantly different from zero. There were no significant correlations between visual sensitivity in fricative perception and magnitude of McGurk effect. That is, there is no evidence that sensitivity to visual information in the perception of the /s/-/ʃ/ distinction is associated with McGurk effect.

**Discussion**

The overarching goal of this study was to evaluate the role of visual information in speech perception by English language learners in order to better understand the role of visual information in second-language speech acquisition. Two questions were posed to explore this: Does a correlation exist between the abilities to produce and perceive the /s/-/ʃ/ distinction, and, is magnitude of McGurk effect correlated with the ability to produce and perceive the /s/-/ʃ/ distinction? Results of fricative production, fricative perception, and McGurk tasks for native and non-native speakers of English were analyzed to answer these questions. Overall, significant correlations between the abilities to produce and perceive this distinction were observed for non-native participants in their first experimental session, though this did
not carry over into the second test session. No correlation was found between
magnitude of McGurk and the ability to perceive the /s/-/ʃ/ distinction. These findings,
as well as specific task results, provide insight into the role of visual information in
speech perception by English language learners.

Non-native speech perception is not expected to perfectly match that of native
speakers, especially in earlier stages of language acquisition. Many factors influence
second-language speech acquisition, including the learner’s native phonology.
Although non-native results are not expected to mirror those of native participants in
this study, comparisons within the non-native group as well as comparisons between
the two groups (native and non-native) provides insight into second-language speech
acquisition.

Taken as a whole, native results suggest little reliance on visual information to
make the /s/-/ʃ/ distinction when productive and perceptive abilities are highly
developed. This is shown in the equality of accuracy and reaction time data for native
speakers across the audio-only, audiovisual-congruent, and audiovisual-incongruent
conditions of the fricative perception task. With a lifetime of experience, native
speakers of English likely have developed an acute awareness of the auditory
differences between these two fricative sounds from accurate models of others, as well
as listening to their own accurate productions. Speech perception typically occurs in a
setting which provides either the auditory and visual signal, such as a face to face
conversation, or just the auditory signal, as is the case for a conversation over the
phone. It follows, then, that for native speakers of English, the auditory signal carries
the greatest source of information in speech perception and provides enough information to accurately make the distinction between the two fricative sounds. The observation that both accuracy and reaction time suffer in the absence of an auditory signal further support this by indicating that the visual signal is not equally informative in perception of the \(/s/-/ʃ/\) distinction.

One the other hand, English language learners may have a native phonology which does not include the \(/s/-/ʃ/\) distinction; this situation is actually rather likely, as this distinction is not common among world languages. Without several years of experience both hearing and producing this distinction, the English language learner will need to develop an acute awareness of the acoustic distinction between \(/s/\) and \(/ʃ/\). As these sounds are both voiceless fricatives, the acoustic difference between these two sounds may not be as easily recognized as that between two sounds of different manner and/or voicing, such as a voiceless fricative and a voiced stop. Development of this awareness will take time along with experience with accurate models of both sounds in English. While non-native speakers may not differ in general from native speakers in relying on the auditory signal as the primary source of information during speech perception, a less developed sense of acoustic distinction would make relying on this signal alone difficult. In other words, if non-native speakers do not have a sufficiently developed acoustic understanding of the \(/s/-/ʃ/\) distinction, then relying on auditory information to make this distinction is likely to result in inaccurate perceptions.
While the non-native speakers in this study performed with high accuracy in all tests of fricative perception, their reaction time patterns indicated that the visual signal of lip rounding in the production of /ʃ/ facilitated the recognition of this fricative more quickly than audio-only presentations, which lacked the visual signal, and presentations of /s/, which also lacks a consistent, distinct visual signal. This pattern then, suggests that the non-native speakers do in fact utilize visual information in perceiving the /s/-/ʃ/ distinction. Comparing this to native data, it is observed that visual information appears to be utilized to a greater degree in non-native speech perception. This difference may be a result of the non-native speakers’ development of the /s/-/ʃ/ distinction. Further, it may be that the non-native speakers did rely significantly on the visual signal of lip rounding to accurately perceive the distinction in the early stages of acquisition, and have maintained this perceptual strategy, even as acoustic awareness improved. This would account for the differences in perceptual strategies suggested by distinct reaction time results observed between native and non-native speakers, even when exhibiting nearly equal levels of accuracy.

Although no direct correlation was found between the magnitude of McGurk effect and the use of visual information in perceiving the /s/-/ʃ/ distinction, it is worth noting that non-native participants differed from their native counterparts in observed McGurk magnitude. The presence overall of a higher magnitude of McGurk effect in the non-native group further supports the observation that non-native participants utilized visual information to a greater degree in speech perception than did native participants. Both groups experienced an increase in reaction time for incongruent
McGurk presentations, indicating perception of the incongruent visual signal. The non-native group, however, integrated the incongruent auditory and visual signals to a greater degree, as indicated by the stronger McGurk effect observed in this group. This provides further evidence of the significant use of visual information in non-native speech perception.

Second language instruction may benefit from consideration of this unique perceptual strategy. If English language learners are explicitly taught about both the visual and auditory distinctions accomplished by correct articulation of the /s/-/ʃ/ distinction, it stands to reason that this may allow them to be more effective communicators in the early stages of learning. Recognizing this visual distinction may aid perception of the sound difference, providing more opportunities early in the learning process to develop accurate acoustic representations. Explicit instruction regarding lip rounding would also be expected to improve production of the /ʃ/ fricative. Based on the correlation between production and perception observed in the non-native results from session 1—which was not found to carry over to the second session, suggesting these abilities are connected only in stages of lower proficiency—the improvement in production and/or perception of the /s/-/ʃ/ distinction, facilitated by instruction regarding lip rounding, may in turn improve both production and perception over time. Put another way, if, for example, direct instruction on lip rounding improved the production of /ʃ/ for an English language learner, this improvement would be expected to carry over into perception of /ʃ/, as these abilities were found to correlate in earlier stages of language acquisition. This correlation,
along with observations of use of visual information in speech perception by non-native speakers, suggests a potential benefit in training both production and perception (auditory and visual) in second language instruction.

Limitations

The results of this study are limited by the variance among non-native participants. Non-native participants were recruited for this study from a limited number of international students enrolled in oral communication courses at Ohio University. As the population from which participants were recruited was small, factors such as language background, experience with English, proficiency level, etc. could not be controlled for. For this reason, data from this study only applies in a broad, general sense to non-native speakers of English and cannot be interpreted to provide implications for individual language backgrounds. These general observations do, however, provide a good basis for further, more controlled explorations of the role of visual information in second language acquisition.

Future Directions

As the current study is exploratory, several directions could be taken in future research. As noted previously, this study was limited by the wide variance in participant characteristics such as language background, experience with English, and proficiency level. It may be of interest to repeat this study with a more homogenous group of less proficient English language learners. This would allow a clearer observation of productive and perceptual development over time and the resulting
correlations, if any, with use of visual cues. As results of the current study indicated a connection between productive and perceptual abilities during periods of lower proficiency, tracking this relationship throughout the course of second-language speech acquisition would provide a deeper understanding of the factors which led to the findings of this study.

A larger sample size would also potentially allow for comparisons to be made between English language learners whose native phonologies include the /s/-/ʃ/ distinction and those who do not use this distinction in their native language. This would then provide opportunity for more language-specific observations and implications. Effects of factors such as a tonal native language on perception of second-language speech could be explored, as well.

The current study excluded /s/-/ʃ/ minimal pairs that utilize back vowels from the stimulus set to avoid an impact of co-articulation. Extending the stimulus set of to include these back vowels may provide further insight into the use of visual speech cues by non-native speakers of English. As the current study demonstrated awareness of the visual cue of lip rounding in non-native production and perception of the /s/-/ʃ/ distinction, exploring how this group perceptually manages the co-articulatory impact of a rounded vowel on the visual cue of /s/ may be of interest.

Additionally, evaluating the perception of other languages by native and non-native speakers may also provide a clearer understanding of the role of visual cues in second-language speech perception. Specifically, Mandarin has a three-way fricative
distinction much like the English /s/-/ʃ/ distinction which utilizes neutral lips, spread lips (similar to English /s/), and rounded lips (similar to English /ʃ/). Exploring how both native and non-native speakers of Mandarin perceive this three-way distinction would provide a set of data which would complement that of the current study. Particularly, a comparison between results may provide further insight into differences between native and non-native speaker perception of language, as was observed in the current study.

**Conclusion**

Results of the current study provide evidence that visual information plays a significant role in second language speech perception. Non-native participant accuracy and response time data show that the visual signal of lip rounding facilitates more accurate and quicker perception of /ʃ/ compared to /s/. Correlation results reveal that in stages of lower proficiency, the abilities to produce and perceive the /s/-/ʃ/ distinction are connected, indicating that a more proficient understanding of lip rounding promotes the development of this fricative consonant distinction. Although McGurk results did not prove to correlate with fricative perception, high magnitudes observed in the non-native speaker group further support the role of visual information in second-language speech perception.
References


Appendix A: 21 Fricative Word Pairs

sack-shack
sad-shad
sag-shag
said-shed
sake-shake
same-shame
save-shave
sear-shear
seat-sheet
see-she
seen-sheen
seep-sheep
self-shelf
sell-shell
sift-shift
sigh-shy
sign-shine
sin-shin
sip-ship
suck-shuck
sun-shun
Appendix B: Fricative Perception Task—On-Screen Instructions

The following instructions were presented on the iMac screen for the Fricative Perception and McGurk tasks.

<table>
<thead>
<tr>
<th>Task</th>
<th>Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main Instruction for Fricative Perception and McGurk Tasks</td>
<td>You will complete four tasks throughout the experiment. Please keep your focus on the screen during each task. Press the space bar to begin.</td>
</tr>
<tr>
<td>Fricative Perception: Audio-Only Condition</td>
<td>You will listen to a list of English words. The words will begin with the “s” sound (as in “save”) or the “sh” sound (as in “shave”). Your task is to decide whether the first sound is “s” or “sh” by pressing the buttons labeled on the response box. You will first practice this task. Press the space bar to begin.</td>
</tr>
<tr>
<td>Fricative Perception: Audiovisual Condition (congruent and incongruent)</td>
<td>You will watch videos of a person speaking English words. The words will begin with the “s” sound (as in “save”) or the “sh” sound (as in “shave”). Your task is to decide whether the first sound is “s” or “sh” by pressing the buttons labeled on the response box. You will first practice this task. Press the space bar to begin.</td>
</tr>
<tr>
<td>Fricative Perception: Video-Only Condition</td>
<td>You will watch videos of a person speaking English words. The words will begin with the “s” sound (as in “save”) or the “sh” sound (as in “shave”). However, you will not hear the actual sounds. Your task is to guess from the video whether the first sound is “s” or “sh” by pressing the buttons labeled on the response box. Press the space bar to begin.</td>
</tr>
<tr>
<td>McGurk Task: Transition (allowed researcher to change response pad labels to “b”, “d”, “g”)</td>
<td>Please wait for the experimenter.</td>
</tr>
</tbody>
</table>
Appendix B: Fricative Perception Task—On-Screen Instructions (continued)

<table>
<thead>
<tr>
<th>McGurk Task</th>
<th>You will watch videos of a person speaking some English syllables. Your task is to decide what the first sound of each syllable is by pressing the buttons labeled on the response pads. The buttons include “B” as in ba/bi, “D” as in da/di, and “G” as in ga/gi. Press the space bar to begin.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transition from Practice to Test Tokens (presented after each set of practice presentations)</td>
<td>If you have any questions, please ask the experimenter. To begin the test, press the space bar.</td>
</tr>
</tbody>
</table>