Influence of native language phonetic system on audio-visual speech perception

Yue Wang\textsuperscript{a,}\textsuperscript{*}, Dawn M. Behne\textsuperscript{b,1}, Haisheng Jiang\textsuperscript{a}

\textsuperscript{a}Department of Linguistics, Simon Fraser University, RCB 9224, 8888 University Drive, Burnaby, BC, Canada V5A 1S6
\textsuperscript{b}Department of Psychology, Norwegian University of Science & Technology, NO-7491 Trondheim, Norway

Received 19 May 2008; received in revised form 18 February 2009; accepted 20 April 2009

Abstract

This study examines how native language (L1) experience affects auditory–visual (AV) perception of nonnative (L2) speech. Korean, Mandarin and English perceivers were presented with English CV syllables containing fricatives with three places of articulation: labiodentals nonexistent in Korean, interdentals nonexistent in Korean and Mandarin, and alveolars occurring in all three L1s. The stimuli were presented as auditory-only, visual-only, congruent AV and incongruent AV. Results show that for the labiodentals which are nonnative in Korean, the Koreans had lower accuracy for the visual domain than the English and the Mandarin perceivers, but they nevertheless achieved native-level perception in the auditory and AV domains. For the interdentals nonexistent in Korean and Mandarin, while both nonnative groups had lower accuracy in the auditory domain than the native English group, they benefited from the visual information with improved performance in AV perception. Comparing the two nonnative groups, the Mandarin perceivers showed poorer auditory and AV identification for the interdentals and greater AV-fusion with the incongruent AV material than did the Koreans. These results indicate that nonnative perceivers are able to use visual speech information in L2 perception, although acquiring accurate use of the auditory and visual domains may not be similarly achieved across native groups, a process influenced by L1 experience.

\(2009\) Elsevier Ltd. All rights reserved.

1. Introduction

Examining the auditory–visual (AV) perception of English fricative consonants by native Korean and Mandarin Chinese perceivers, this study explores the extent to which AV speech perception in a second language (L2) is affected by a native language (L1) phonetic system.

1.1. Previous findings

It has been well established that visual speech information facilitates speech perception for adults (e.g., Bernstein, Auer, & Takayanagi, 2004; Davis & Kim, 2004; Erber, 1969; Sumby & Pollack, 1954; Summerfield, 1979) as well as for infants (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999, 2003; Weikum et al., 2007). The relative contribution of auditory and visual speech information has also been revealed by what is known as the “McGurk effect”, where an auditory stimulus (e.g., /b/) dubbed onto a different visual stimulus (e.g., /g/) can lead to an intermediate percept that bears the features from both the auditory and the visual input (e.g., /d/, McGurk & MacDonald, 1976).

However, although the ability to integrate AV speech information exists cross-linguistically, language specific patterns have also been observed. For example, whereas the McGurk effect occurs for native perceivers of a variety of languages such as English, Finnish and French (Massaro, 1998, 2001; Massaro, Tsuzaki, Cohen, Gesi, & Heredia, 1993; McGurk & MacDonald, 1976; Sams, Manninen, Surakka, Helin, & Katto, 1998), it appears to be weak for native perceivers of Chinese and Japanese (Sekiyama, 1997; Sekiyama & Tohkura, 1993). In addition, while perceivers across languages (e.g., English, Spanish) demonstrate similar auditory–visual processing patterns perceiving a /ba/-/da/ continuum (Chen & Massaro, 2004;
Chinese perceivers show less use of visual cues in the perception of /b/ and /d/ than Dutch perceivers (de Gelder & Vroomen, 1992). Although this lack of efficient use of visual speech cues has been attributed to the fact that Chinese as a tone language relies more on the less visually distinct prosodic information (Sekiyama, 1997; Sekiyama & Tohkura, 1993), research has also shown a significant effect of visual information on tone perception in Chinese (Burnham, Lau, Tam, & Schoknecht, 2001).

These cross-linguistic differences in L1 AV processing suggest that nonnative (especially inexperienced L2) users’ AV perception may be differentially affected by their L2 and their L1. While nonnatives may be facilitated by general visual speech information (Davis & Kim, 2004; Erdener & Burnham, 2005; Hardison, 1999; Navarra & Soto-Faraco, 2007; Reisberg, McLean, & Goldfield, 1987; Sekiyama, Kanno, Miura, & Sugita, 2003), they also reveal poorer L2 AV perception than in their L1 (de Gelder & Vroomen, 1992; Sekiyama, 1997; Werker, Frost, & McGurk, 1992). For example, with only visual speech information, Italian and English natives cannot discriminate Catalan and Spanish (Soto-Faraco et al., 2007).

Nonnative perceivers, such as L2 perceivers of English and Japanese, are also more susceptible to the McGurk illusion for L2 stimuli (Sekiyama & Tohkura, 1993) than for L1 stimuli, although Chinese perceivers exhibit an equally weak McGurk effect for their L1 and an L2 (Japanese or English) (Sekiyama, 1997; Sekiyama, Tohkura, & Umeda, 1996).

Further research shows that while L2 perceivers do not have difficulty perceiving the visual cues common in their L1 and L2, they may more likely be impeded in correct use of L2 visual cues non-existent in their L1 compared to those occurring both in their L1 and L2 (Hardison, 1999, 2003, 2005a, 2005b; Hazan, Sennema, Faulkner, & Ortega-Llebaria, 2006; Ortega-Llebaria, Faulkner, & Hazan, 2001). For example, Spanish learners outperform Japanese learners in the perception of the visual cues for the English /b-v/ contrast, since /v/ exists allophonically in Spanish as the voiceless counterpart /f/, whereas it is nonexistent in Japanese (Hazan et al., 2006). Likewise, nonproficient French learners of English cannot efficiently perceive the visual information of the English interdental fricatives absent in French (Werker et al., 1992). These studies indicate that L2 learners have difficulty in correctly using the visual cues that do not contrast phonetically in their L1, suggesting that subsequent research should focus on new L2 sounds.

Indeed, while L2 speech perception has been studied extensively in the auditory learning domain, research has not adequately explored how L1 and L2 systems interact to affect the perception of new L2 visual speech categories (Hardison, 1999, 2003; Hazan, Sennema, Iba, & Faulkner, 2005; Hazan et al., 2006). Theories developed on the basis of nonnative or L2 auditory speech learning may inform patterns of visual speech learning. These theories (e.g., Speech Learning Model, SLM, Flege, 1995, 2007; the Perceptual Assimilation Model, PAM, Best, 1995; Best & Tyler, 2007) posit an interactive relationship between L1 and L2 phonetic systems, stating that a nonnative listener’s inability to accurately perceive L2 sounds is due to incorrect assimilation of L2 sounds with L1 phonetic categories and (in the case of L2 learners) a failure to establish new L2 phonetic categories. It has been proposed that category formation is most likely if an L2 speech sound is perceived to be phonetically distant from the closest L1 sound, whereas the L2 sounds that are phonetically similar to those in the L1 are often incorrectly assimilated to an L1 phonemic category, and the formation of separate phonetic categories for those similar L2 sounds can be most challenging (e.g., Flege, 1995, 2007).

L2 visual speech learning may also be a matter of establishing visual speech categories. To classify visual speech categories, previous research has identified various homophonous groups based on the visual speech information in speakers’ articulatory movements (in English, for example, bilabials [p,b,m], labiodentals [f,v], interdentals [θ,ð], alveolars [t,d,n,s,z], etc., Binnie, Montgomery, & Jackson, 1974; Hardison, 1999). In terms of L2, Hazan et al. (2006) specified three types of relationships between L1 and L2 visual speech categories: a visual category occurring in both L1 and L2 (e.g., /s/ in English and Mandarin); a visual category existent in L2 but not in L1 (e.g., English /f/ for Korean perceivers); and a visual category occurring in both L1 and L2, but used for different phonetic distinctions in the L2 (e.g., English labiodental /v/ occurring in Spanish only as voiceless). While the visual categories that are common to an L1 and L2 do not present much difficulty, L2 perceivers may have lost sensitivity to visual cues that are not used in their L1 and need to learn to associate these visual cues with a visual category in the L2 to establish new L2 categories (Hazan et al., 2005, 2006). Further research has indeed shown that new visual categories may be established with linguistic experience, as L2 perceivers’ use of nonnative visual speech information is positively correlated with their length of residence (LOR) in an L2 country (Wang, Behne, & Jiang, 2008) and their overall proficiency in the L2 (Werker et al., 1992), and as sensitivity to visual information in L2 sounds can be enhanced through auditory–visual training (Hardison, 2003, 2005b; Hazan et al., 2005).

It should be noted that L2 categories may be new in terms of their auditory distinctions, visual distinctions or both, where the relative familiarity of audio and visual distinction of a new L2 category do not necessarily present the same challenge for a learner (e.g., Hardison, 1999; Hazan et al., 2006). Thus the connection of the auditorily- and visually-based L1 and L2 phonetic distinctions needs to be examined in greater detail. For instance, similar phones with a voicing distinction such as /p/ and /b/ often cause confusion for L2 learners when perceiving the contrast auditorily (e.g., Flege, 1980), but may not be necessarily difficult in the auditory–visual domain, as has
been shown with Spanish perceivers who can correctly perceive the nonnative /v/ due to the existence of /l/ in their L1 (Hazan et al., 2006), possibly because voicing is not as visually distinctive as other features such as place of articulation. Indeed, previous studies indicate that the difficulty in L2 visual speech perception may lie in the places of articulation that are not used in the L1 (Hardison, 1999; Hazan et al., 2006; Werker et al., 1992). However, while most research compares the visual perception of the place of articulation of L2 consonant contrasts, those contrasts (e.g., /b/ vs. /v/) are often spread across manners of articulation (e.g., Hardison, 2005b; Hazan et al., 2006; Sekiyama & Tohkura, 1993; Werker et al., 1992). Since manner of articulation, as well as place of articulation, may affect visual speech perception (Faulkner & Rosen, 1999; Green & Kuhl, 1989), an extension for further research is to control for these differences in order to further characterize the relationships of L1 and L2 visual categories and thus address L2 auditory–visual learning patterns.

1.2. The current study

This study examines nonnative auditory–visual perception of English fricatives differing in three visually distinct places of articulation: labiodental, interdental, and alveolar. Two nonnative groups, Korean and Mandarin Chinese, as well as native English controls, are included for their differences in L1 phonetic systems, where English contains fricatives and corresponding visual cues that are nonnative to the Korean and Chinese perceivers. In particular, Mandarin does not contain interdental fricatives, but prevocalic voiceless labiodentals and alveolars both exist (Ladefoged & Wu, 1984; Svantesson, 1986; Tse, 1988), whereas Korean contains neither labiodental nor interdental fricatives, but voiceless alveolars occur (Hardison, 1999; Kang & Lee, 2002; Kim, 1972; Schmidt, 1996).

The systematic differences in the phonetic systems of these languages make it possible to explore the questions addressed in this study with regard to the effect of L1 influence. Based on the phonetic inventories of their L1s, the English alveolars which occur in both Korean and Mandarin are expected to be perceived in a comparable manner by the native and nonnative perceivers, while Mandarin natives may be unable to correctly perceive interdentals nonexistent in Mandarin, and Korean natives may fail to accurately perceive both labiodentals and interdentals not occurring in Korean. Moreover, based on the assimilation patterns proposed by auditory speech learning theories where an L2 sound is often incorrectly assimilated to a perceptually close L1 sound category (Best, 1995; Best & Tyler, 2007; Flege, 1995, 2007), Mandarin natives’ perception of English interdental fricatives may be inferior to that of Korean natives, since the front visual categories (thus perceptual space) are more crowded for Mandarin in the vicinity of interdental fricatives (i.e., with labiodentals and alveolars to which interdentals may be assimilated).

Subsequent questions follow from the interaction of auditory and visual perception of these fricatives, with regard to how Korean and Mandarin perceivers weigh L2 auditory and visual cues under the influence of their L1s, and the extent to which auditory or visual cues contribute to the percept in the L2. To tease apart the influence of auditory and visual information in a single stimulus, this study makes use of the McGurk illusion (McGurk & MacDonald, 1976). Thus, in addition to auditory, visual, and congruent auditory–visual input modalities, incongruent auditory–visual input (with labiodental and alveolar fricatives as auditory or visual components, and the intermediate interdentals projected to be one of the major fused percepts) is included to determine the contribution of each component modality. Of particular interest are the extent to which the Mandarin perceivers either remain anchored to the familiar auditory or visual components, or perceive the fused nonnative interdental percept (or other intermediate sounds), and the extent to which the Koreans use these modalities where a nonnative sound is involved in both the input (labiodentals) and one of the expected fused percepts (interdentals). Given the previous finding that nonnatives who do not rely heavily on visual cues in their L1 tend to make more use of visual information in an L2 (e.g., Sekiyama (1994), we predict that both Korean and Mandarin perceivers will be more responsive to the visual component or fused percept bearing the articulatory features of the visual component than the English natives. Furthermore, based on the previous claim that languages with fewer visible place of articulation contrasts are less visually marked (Hazan et al., 2006), we predict that, compared to Mandarin perceivers, Koreans will use less visual information.

2. Methods

2.1. Participants

Three groups of young adults (mean age = 23, balanced for the numbers of males and females) participated in the study: 15 native Canadian English, 15 native Korean, and 20 native Mandarin speakers. All were undergraduate or graduate students at Simon Fraser University (SFU, Vancouver, Canada) at the time of the study. The participants’ language background information is summarized in Table 1. The Korean and Mandarin participants were recruited to have comparable backgrounds in their experience with English. They all had studied English as an L2 in a classroom setting 3–5 h/week since an average age of 10–11 years. As an admission requirement into SFU, they passed a test of English, either the International English Language Testing System (IELTS) with a minimum score of 6.5 (out of 9) or the Test of English as a Foreign Language computer-based test (TOEFL CBT) with a minimum score of 230 (out of 300). None were
English majors. They came to Canada after age 18 years and had been there for a relatively short length of time (1–4 years), but never resided in any other English speaking environments prior to their arrival in Canada. They reported using and being exposed to their native Korean or Mandarin as their more dominant language since their arrival. All three groups of participants reported having normal or corrected vision and no known history of speech and hearing impairments. They were compensated for their participation.

2.2. Stimuli

Stimuli, exemplified in Table 2, were based on 18 English CV syllables having a fricative onset ([f], [v], [θ], [ð], [s], or [z]) followed by a vowel ([i], [a], or [u]). The fricatives included a series of auditory–visual categories (labiodental, interdental, alveolar) differing in place of articulation, from relatively front to relatively less front in the vocal tract. As described earlier, whereas these three places of articulation are contrastive in English, in Mandarin only voiceless labiodental and alveolar fricatives are phonemically distinctive, and in Korean only voiceless alveolar fricatives exist. Both voiced and voiceless fricatives were included for a broad evaluation of place of articulation identification. The vowels ([ɪ], [a], [ʊ]), phonemically distinctive in English, Mandarin, and Korean, represent diverse vocal tract configurations differing in tongue height, advancement, and lip-rounding (Hazan et al., 2005, 2006; Jongman, Wang, & Kim, 2003). Thus each place of articulation was represented by six exemplars (2 voicing conditions /C2/ 3 vowels) to ensure responses to phonetic distinctions rather than acoustic idiosyncrasies.

These syllables were the basis for developing stimuli varying in AV composition of the syllable’s fricative POA: (1) congruent auditory and visual syllables (AVc), (2) auditory-only syllables (A), (3) visual-only syllables (V), and (4) incongruent auditory–visual stimuli (AVi), including two auditory and visual input alternatives differing in fricative POA (\(A_{\text{labiodental}}-V_{\text{alveolar}}\), \(A_{\text{alveolar}}-V_{\text{labiodental}}\)).

In the AVi stimuli, the auditory and visual components had mismatched labiodental and alveolar fricatives such that they were either \(A_{\text{labiodental}}-V_{\text{alveolar}}\) (e.g., auditory [fa] dubbed onto visual [su]) or \(A_{\text{alveolar}}-V_{\text{labiodental}}\) (e.g., auditory [sa] dubbed onto visual [fa]). Interdentals (nonnative to Mandarin and Korean perceivers) were not included in the creation of the mismatched stimuli, but are expected to be one of the major occurrences of AV-fusion, being the only fricative category intermediate to the labiodental and alveolar places of articulation. The interdentals share the “nonsibilant” acoustic feature with the labiodentals and the “nonlabial” visual feature with the alveolars, as is shown by previous research that a fused percept usually combines auditory and visual information (e.g., McGurk & MacDonald, 1976; Sekiyama et al., 2003; Werker et al., 1992). In a given AVi syllable the A and V components always had the same fricative voicing and

### Table 1
Participants’ language background information.

<table>
<thead>
<tr>
<th>Language</th>
<th>Age (years)</th>
<th>AOL (years)</th>
<th>AOA (years)</th>
<th>LOR (years)</th>
<th>English study (years)</th>
<th>L1 input (%)</th>
<th>English input (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English (n = 15)</td>
<td>22 (4)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>66 (23)</td>
<td>34 (23)</td>
</tr>
<tr>
<td>Korean (n = 15)</td>
<td>24 (4)</td>
<td>10 (5)</td>
<td>19 (5)</td>
<td>4 (1)</td>
<td>10 (3)</td>
<td>60 (23)</td>
<td>40 (23)</td>
</tr>
<tr>
<td>Mandarin (n = 20)</td>
<td>24 (4)</td>
<td>11 (3)</td>
<td>21 (4)</td>
<td>2 (1)</td>
<td>12 (4)</td>
<td>66 (23)</td>
<td>34 (23)</td>
</tr>
</tbody>
</table>

Note: AOL = mean age of initial English learning; AOA = mean age of arrival in Canada; LOR = mean length of residence in Canada (years); English study = mean number of years of formal English study (as a school subject); L1 input = mean % daily input in Korean and Mandarin, respectively; English input = mean % daily input in English (L2). The standard deviations are included in parentheses.

### Table 2
The set of English CV syllables containing voiceless fricatives presented as auditory-only (A), visual-only (V), and congruent auditory–visual (AVc) modalities for each of the fricative POAs (labiodental, interdental, alveolar), and as incongruent auditory–visual stimuli (AVi), including two auditory and visual input alternatives differing in fricative POA (\(A_{\text{labiodental}}-V_{\text{alveolar}}\), \(A_{\text{alveolar}}-V_{\text{labiodental}}\)).

<table>
<thead>
<tr>
<th>Modality</th>
<th>A, V, or Avc</th>
<th>AVi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fricative POA</td>
<td>Labiodental</td>
<td>Interdental</td>
</tr>
<tr>
<td>Sample stimuli</td>
<td>fi</td>
<td>0i</td>
</tr>
<tr>
<td>fa</td>
<td>0a</td>
<td>sa</td>
</tr>
<tr>
<td>fu</td>
<td>0u</td>
<td>su</td>
</tr>
<tr>
<td>Occurrence</td>
<td>English Mandarin</td>
<td>English Mandarin Korean</td>
</tr>
</tbody>
</table>

Note: POA = place of articulation; \(A_{\text{labiodental}}-V_{\text{alveolar}}\): auditory input = labiodental, visual input = alveolar; \(A_{\text{alveolar}}-V_{\text{labiodental}}\): auditory input = alveolar, visual input = labiodental.
vowel. Thus 12 AVi stimuli were included (2 AV-input place [labiodental]−[alveolar], [alveolar]−[labiodental]×2 voicing conditions ×3 vowels).

A total of 66 stimuli were used across stimulus conditions (18 A, 18 V, 18 AVc, 12 AVi).

Audio and video recordings were made with an adult male speaker of Canadian English sitting against a white background in the recording studio in the Language and Brain Lab at Simon Fraser University. While the speaker produced six randomized repetitions of the 18 English syllables (6 fricatives × 3 vowels) at a normal speaking rate, recordings of the speaker’s face were made using a digital camcorder (SONY DCR-HC30/40) positioned approximately 2 m away. In addition, separate audio recordings were simultaneously made with a Shure KSM 109 condenser microphone via an audio interface (M-audio MobilePre USB) to a PC at a 44.1 kHz sampling rate. These high quality audio recordings were used to replace the audio track from the camcorder recording.

A best example was selected from among the six repetitions for each syllable such that the durational difference among the 18 selected syllables was under 10%, the approximate just noticeable difference for duration (Lehiste, 1970). For these selected syllables, the audio-video recordings from the camcorder were aligned with the corresponding high quality audio recordings by synchronizing the two waveforms using SoundForge 8.0. The audio track from the camcorder was then deleted. The audio tracks were then normalized to attain the same unweighted RMS value for the resulting stimuli. The audio tracks were edited to have a 1.2 s neutral fade before and after the stimulus; that is, before the frame where mouth opening first occurred and after the frame where the mouth was fully closed. All stimuli had a frame length of .06 s, and a resolution of 640 × 480 pixels.

The resulting AV materials were used as AVc stimuli. They were also the basis for creating the A-only, V-only, and AVi stimuli. The A-only and V-only stimuli were created from the AVc stimuli by removing the video tracks or muting the audio tracks, respectively. The AVi stimuli were based on the same auditory and visual components as those used in the AVc, A-only and V-only conditions. To create the AVi stimuli, the audio and video components were aligned from syllables differing only in place of articulation (e.g., A[fa]−V[sa]). Starting with the AVc stimulus which had the target video component (e.g., AVc [sa]), the audio signal of the target audio component (e.g., [fa]) was aligned with the onset of the fricative audio signal from the AVc, and the original AVc audio (e.g., audio [sa]) was removed, so that the resulting AVi had the original AVc video component (e.g., [sa]) with a new audio signal (e.g., [fa]).

To test the intelligibility of the audio signals and the naturalness of the AV signals, the final AVc and AVi stimuli were evaluated by two phonetically trained native speakers of English. An identification task testing intelligibility of the audio signals showed 95% correct responses (errors were 1.5% labiodental, 2% interdental, .5% alveolar, and 1% voicing and other errors) for the audio signals in the AVc stimuli and 100% correct responses for the audio signals in the AVi stimuli. These scores are comparable (and exceed) those in previous AV studies of auditory perception of English fricatives by native English listeners (e.g., Jongman et al., 2003; Werker et al., 1992). The naturalness of the AV stimuli was tested in a 5-point goodness rating task (5 being the most natural) where the same two evaluators judged whether the audio and video signals were naturally synchronized, regardless of what they heard or saw. AVc stimuli were rated 4.4 and the AVi stimuli rated 4.6.

2.3. Procedure

A perception experiment was generated using E-prime 1.0 (Psychology Software Tools, integrating video clips imported from Microsoft Powerpoint files) to present stimuli and log participant responses. Participants were tested in an identification task using the full set of stimuli blocked by modality: A, V, AV (with AVi and AVc stimuli in the same block). Two randomized repetitions of each stimulus were included in each block. The presentation order of the blocks was counter-balanced across participants. The test began with instructions for the task, familiarization with the stimuli (e.g., matching the symbols and the sounds they represent), and five practice trials for each modality (A, V, AVc/AVi). The practice trials were presented with the same speaker as used in the test, but no feedback was provided to avoid any learning effect. The test was followed by a debriefing session, which included a post-test questionnaire. The full experiment, including a short break after three test blocks, lasted about one hour for a participant.

Stimuli were presented auditorily over loudspeakers, visually on a computer monitor, or both. Participants were tested individually, sitting approximately 1 m from a 20 in LCD flat panel computer monitor and two loudspeakers (Altec Lansing) positioned on each side of the monitor, so that the audio and video sources were approximately the same distance from the perceiver. Loudspeakers were used instead of headphones to avoid any bias for the audio component. The audio signal had a comfortable level of approximately 70 dB which has previously been shown to be an appropriate level for speech perception experiments (Nábelek & Robinson, 1982; Takata & Nábelek, 1990). For each trial, a visual fixation point was displayed in the middle of the monitor for one second, followed by the target stimulus.

Six response alternatives [f, v, θ, s, z] were then shown on the monitor, together with an “other” option to allow participants to type in an alternative response. “Th” and “dh” were used to represent [θ] and [ð], respectively (Jongman et al., 2003). The participants were given the “other” option since previous research has shown that participants’ responses are not limited to the given type of
responses, and in the case of incongruent stimuli, they are not limited to one (fused) response (e.g., Hardison, 1999; Sekiyama et al., 1996; Werker et al., 1992). However, these studies have also revealed that most of the alternative responses involved a difference in manner of articulation and/or voicing in addition to POA. For instance, the responses involved a difference in manner of articulation studies have also revealed that most of the alternative responses were typically analyzed in terms of POA regardless of voicing and manner (e.g., Burnham & Dodd, 2004; Hayashi & Sekiyama, 1998; Werker et al., 1992). Thus, as used in other similar AV studies (e.g., Chen & Hazan, 2007; Hazan et al., 2006; Jongman et al., 2003), the current study employed the “quasi-closed-set” task (i.e., 6 target fricatives and “other”) with the focus on POA.

The participants were to identify the fricative closest to the one they perceived and respond by pressing the corresponding key on the keyboard. Participants were allowed a maximum of 4 s to respond in each trial. In the blocks where visual stimuli were involved, participants were instructed to always watch the speaker’s face. In addition to the visual fixation mark they were to attend to at the beginning of each trial, visual notes were also placed on the computer monitor and keyboard to remind them to fix their eyes to the monitor. Only one participant was tested at a time which allowed the experimenter to observe the testing process, ensuring that the perceiver’s eyes focused on the speaker’s face.

3. Results

Correct responses were tabulated according to L1 group, AV-input modality, place of articulation of the fricatives, voicing of the fricatives, vowel context, as well as AV-input place (for the AVi condition). Separate data analyses were carried out for the A, V, and AVc conditions, and for the AVi condition.

3.1. Auditory, visual, and congruent auditory–visual conditions

The percent correct identification of the fricatives was analyzed using a 5-way mixed analysis of variance (ANOVA), with Group (English, Korean, Mandarin) as a between-subjects factor, and Modality (A, V, AVc), Place of articulation (POA) (labiodental, interdental, alveolar), Voicing (voiceless, voiced), and Vowel ([i, a, u]) as repeated measures. The dependent variable was perceivers’ correct identification for POA (Hazan et al., 2005, 2006; Jongman et al., 2003; Werker et al., 1992). Fig. 1 presents mean percent correct responses for the English, Korean, and Mandarin groups as a function of Modality and POA.

A significant main effect of Group was found $F(2,47) = 18.9, p < .0001$, with the post hoc analysis (Tukey-HSD) showing that native English perceivers had a higher overall percent correct (84%) than the Korean group (80%), which in turn had a higher percent correct than the Mandarin group (70%). Significant main effects were also observed for Modality $F(2,47) = 179.6, p < .0001$ and POA $F(2,47) = 29.1, p < .0001$, along with significant interactions of Group × Modality $F(5,47) = 6.1, p < .001$, Group × POA $F(5,44) = 5.2, p < .001$, Group × Modality × POA $F(8,41) = 4.1, p < .001$. However, the results showed no significant main effect of Voicing $F(1,48) = .4, p = .517$. Furthermore, although a significant main effect of Vowel was observed $F(2,47) = 22.5, p < .0001$, there were no reliable interactions involving L1 group (Group × Vowel: $F(4,45) = 1.2, p = .308$), Group × Vowel × Modality $F(8,41) = 1.7, p = .110$, Group × Vowel × POA $F(8,41) = 1.1, p = .395$, Group × Vowel × Voicing $F(4,45) = 1.4, p = .228$, Group × Modality × POA × Voicing × Vowel: $F(16,33) = .9, p = .463$), indicating that vowel context affected the three native groups in a similar fashion. Since the goal of the present study was to compare native group difference as a function of AV-input modality and fricative place of articulation, further analyses were thus conducted to focus only on these factors which revealed significant interactions.

Given the significant interactions of Group, Modality and POA, sets of two-way ANOVAs were conducted for each POA to compare how the AV modalities were perceived by different native groups. Significant Group and Modality interactions were observed only for labiodentals $F(4,45) = 3.3, p < .015$ and interdentals $F(4,45) = 4.7, p < .002$, but no interaction occurred for alveolars where all native groups exhibited significantly poorer V than A or AVc perception (see Fig. 1). Subsequent one-way repeated measures ANOVAs and post hoc analyses (Bonferroni adjustment) were carried out with Modality as a factor for each Group for labiodentals and interdentals. As shown in Fig. 1, for labiodentals, whereas the native English group did not reveal any difference in input modality (with equally high scores across the three modalities) $F(2,42) = .9, p = .413$, both the Korean $F(2,42) = 7.8, p < .002$ and Mandarin $F(2,57) = 5.3, p < .009$ perceivers’ identification in the V modality was significantly poorer than in the AVc modality. For interdentals, while the English natives showed poorer identification in V than in both A and AVc conditions $F(2,42) = 15.7, p < .0001$, the Koreans had lower accuracy in the A and V modalities compared to the AVc modality $F(2,42) = 11.2, p < .0001$, and the Mandarin perceivers showed lower identification accuracy in the A than AVc modality $F(2,57) = 5.3, p < .009$. These results indicate that the three native groups’ perception of labiodentals and interdentals differed depending on the input modality.

To compare these group effects directly, sets of one-way ANOVAs and post hoc analyses (Tukey-HSD) were carried out with Group as a between-subjects factor for the labiodentals and interdentals at each modality.
As shown in Fig. 1, significant group differences were observed in the following conditions. For the labiodentals, in the V condition, only the Korean perceivers showed lower identification accuracy than did the English natives \[ F(2,47) = 5.1, \ p < .010 \]. For the interdentals in the A condition, the Mandarin perceivers exhibited a significantly lower identification accuracy than the Korean perceivers, who in turn had a significantly lower identification accuracy than the native English perceivers \[ F(2,47) = 18.9, \ p < .0001 \]. In the AVc condition, only the Mandarin perceivers' identification was lower than that of the native English \[ F(2,47) = 10.2, \ p < .0001 \]. No other conditions revealed significant group differences.

To summarize, (1) for the labiodentals, both the Korean and Mandarin perceivers showed an increase in identification accuracy in the AVc than in V modality. However, only the Korean perceivers' V identification did not reach the native English level; (2) for the interdentals, both the Korean and Mandarin perceivers showed better identification accuracy in the AVc than A or V modalities. Whereas the Koreans showed poorer than native English identification only in the A condition, the Mandarin perceivers showed poorer A and AVc identification than did both the English and Korean groups, and (3) for the alveolars, the three groups show similar patterns, all having much poorer scores in the V than in A or AVc condition, probably due to the alveolars being least visually accessible among the three types of fricatives. Overall, these results reveal that the Korean and Mandarin perceivers perceived the AVc condition better than A or V conditions for their corresponding nonnative sounds (labiodental and/or interdental), although they still could not reach native performance identifying these sounds, particularly the labiodental V identification for the Koreans, the interdental AVc identification for the Mandarin participants, and the interdental A identification for both nonnative groups.

### 3.2. Confusion patterns

The perceivers’ response patterns for POA were analyzed for the A, V and AVc materials, as shown in Table 3. To compare group differences in perceptual confusion patterns, 2-way mixed ANOVAs were conducted for each Modality and POA, with Group (English, Korean, Mandarin) as the between-subject factor and Response pattern (perceived as labiodental, interdental, or alveolar) as the repeated measure, and the post hoc analyses being Tukey-HSD (for between-group comparisons) and Bonferroni adjusted (for within-group comparisons). Significant group differences in response patterns were found for...
the following conditions: labiodental responses in the V condition \(F(4,45) = 3.5, p < .038\), interdental responses in the A \([F(4,45) = 11.4, p < .0001]\), V \([F(4,45) = 3.9, p < .026]\), and AVc \([F(4,45) = 8.0, p < .0001]\) conditions, and alveolar responses in the A \([F(4,45) = 7.5, p < .001]\), V \([F(4,45) = 3.1, p < .020]\), and AVc \([F(4,45) = 22.7, p < .0001]\) conditions.

Post hoc analyses show that, for the labiodentals in the V condition, while all groups to some degree misperceived the labiodentals as interdentals, the Koreans (for whom the labiodentals were nonnative) gave a greater percentage of alveolar responses (8%) than the Mandarin group (5%) which in turn showed more alveolar responses than did the English group (2%) \((p < .040)\), indicating that for the nonnatives, particularly for the Koreans, the labiodentals and alveolars were less visually distinguishable. The three groups did not differ significantly in the A and AVc conditions in terms of the confusion patterns, all more likely misperceiving labiodentals as interdentals than as alveolars \(p < .05\) for all groups.

For the interdentals in the V condition, both the Mandarin and Korean groups’ interdental misperception was more biased towards the alveolars than labiodentals (Mandarin: 30% versus 9%, \(p < .001\); Korean: 25% versus 5%, \(p < .005\)), while the English natives misperceived the interdentals as labiodentals or alveolars to a more similar degree (16% versus 11%, \(p > .716\)). In both the A and AVc conditions, the Mandarin group misperceived 18–19% of the interdentals as alveolars and for the Koreans this misperception was 2–3% (A: \(p < .0001\), AVc: \(p < .001\)), while the English natives never confused interdentals and alveolars. Consistently, it is also noted that for the alveolars across conditions, Mandarin perceivers more often (than the English and Koreans) perceived alveolars as interdentals (A: \(p < .0001\), V: \(p < .031\), AVc: \(p < .0001\)).

These results indicate that across modalities, the nonnatives, particularly the Mandarin group, more likely mixed interdentals with alveolars than did the native English perceivers.

3.3. AV incongruent condition

Responses to AVi stimuli were analyzed based on whether they matched the fricative in the stimulus’ auditory component (A-match), visual component (V-match) or were intermediate (i.e., interdental, since no relevant “other” responses applied) to A and V components (AV-fusion). Sets of 2-way mixed ANOVAs were carried out for each of these response types, with Group as a between-subjects factor, and AV-place input \((A_{labiodental} + V_{alveolar})\) as a repeated measure. Fig. 2 displays the mean percent responses for each response type as a function of AV-place input.

Results for the AV-fusion (interdental) responses show significant main effects of Group \([F(2,97) = 4.9, p < .011]\) and Input place \([F(1,98) = 24.4, p < .0001]\), and a Group × Input interaction \([F(5,94) = 5.6, p < .007]\). Overall, the Mandarin group had a higher mean percent of AV-fusion responses than the Korean group whose responses were in turn greater than the English group (Tukey-HSD, \(p < .05\)), indicating that the nonnative perceivers more easily fused the incongruent A and V components of the stimuli (McGurk effect) despite the intermediate fricatives being nonnative. Further separate one-way ANOVAs for each input place revealed that this pattern of native group difference only reliably occurred with the \(A_{alveolar} + V_{labiodental}\) input \([F(2,47) = 18.9, p < .0001]\), that is to say, with stimuli where the visual component was more visually accessible.

For the A-match responses, significant results were observed for Group \([F(2,97) = 8.5, p < .001]\) and Input place \([F(1,98) = 29.5, p < .0001]\), as well as a Group × Input interaction \([F(5,94) = 3.6, p < .034]\). Separate one-way ANOVAs with each input place revealed a significant group difference for the \(A_{alveolar} + V_{labiodental}\) input \([F(2,47) = 17.5, p < .0001]\), showing that the Mandarin perceivers gave fewer responses matching the auditory component than English and Korean participants (Tukey-HSD, \(p < .05\)).

Finally, for the V-match responses, there was no significant main effect of Group \([F(2,97) = 2.3, p = .109]\) or Input place \([F(1,98) = .23, p = .632]\). However, a significant Group × Input interaction was observed \([F(5,94) = 4.4, p < .020]\). Further analyses showed a reliable difference only for the \(A_{labiodental} + V_{alveolar}\) input \([F(2,47) = 5.0, p < .011]\), with a greater number of responses for the Mandarin than Korean and English perceivers.

Overall, all groups more predominantly responded to the auditory than the visual input. The nonnatives, Mandarin in particular, had fewer A-match responses compared to the native English perceivers. While the native English group only demonstrated AV-fusion when the auditory component was labiodental and the visual component was alveolar but not the reverse, the nonnatives, especially the Mandarin group, showed similar effects for both input places, demonstrating that the Mandarin perceivers more easily fused the incongruent auditory and visual components of the stimuli. These results showed that whereas the Mandarin perceivers were more affected by the visual input, the pattern of results for the Koreans is more like that of the native English perceivers, with more use of auditory information.

4. Discussion

4.1. Perception of auditory and visual modalities

The results show that native English perceivers can in general accurately perceive the fricatives in AVc and A conditions, agreeing with previous results of the perception of English fricatives (e.g., Jongman et al., 2003) and other consonants (e.g., Chen & Hazan, 2007; Hardison, 1999; Massaro, 1998; Sekiyama & Tohkura, 1993; Werker et al., 1992). The results are also consistent
with the previous findings that even native perception of the interdentals and alveolars (particularly the latter) is poorer in the V than in the A and AVc conditions (Chen & Hazan, 2007a, 2007b; Cienkowski & Carney, 2002; McGurk & MacDonald, 1978), and that consonants involving labials are most visually distinctive (Hardison, 1999; Woodward & Barber, 1960). Thus, it appears that the native English perceivers actively use the auditory component across fricative places of articulation. Their use of the visual information decreases with the decrease of visual salience (from labiodentals to alveolars) and with the increase of auditory salience (from the nonsibilant labiodentals to alveolars).

Agreeing with the previous findings (e.g., Hardison, 1999, 2003; Hazan et al., 2005, 2006; Navarra & Soto-Faraco, 2007; Sekiyama, 1997), group comparisons also reveal differences for the English, Korean, and Chinese natives in the use of auditory and visual cues when perceiving English. The results show that native and nonnative perceivers vary in their accuracy perceiving the different POAs and in weighing the auditory and visual input, depending on whether the auditory and/or visual cues are linguistically distinctive in their L1.

Of particular interest is the perception of the interdental fricatives nonnative to both Mandarin and Korean perceivers. The results support the first hypothesis raised in this study that the nonnative perceivers would not reach native performance (particularly in the A and/or AVc modalities). However, both nonnative groups performed better in the AVc than the A condition, suggesting an ability to exploit visual information in L2 speech perception. Given previous findings that perceivers rely more on visual speech information when intelligibility is poor (Erber, 1969; Sumby & Pollack, 1954; Summerfield, 1979), these nonnative perceivers conceivably made use of the visual information as an additional channel of input in perceiving the difficult nonnative sounds. Indeed, previous research has consistently shown that the perception of L2 stimuli (such as English /f, r/ by Japanese) increases with additional visual information (Hardison, 1999) in that auditory perception of the most difficult sounds benefits most from visual cues (Hardison, 2005a, 2005b), and that visual cues enhance L2 speech comprehension (Navarra & Soto-Faraco, 2007; Reisberg et al., 1987; Soto-Faraco et al., 2007).

Results for labiodental perception show that for the Koreans whose L1 does not contain labiodentals, perception reached the native English level in the A and AVc conditions, but not the V condition. Similar patterns have previously been reported, with Korean learners demonstrating 92% correct perception of the English /f/ in an A condition and 97% in an AV condition (Hardison, 1999). The relatively accurate auditory perception of the labiodentals may be because, as nonsibilants, labiodentals are acoustically and perceptually very distinctive from the closest L1 sounds in Korean which are the sibilant alveolar fricatives (e.g., Jongman, Wayland, & Wong, 2000; Jongman et al., 2003). In the V condition, perceptual accuracy with labiodentals is lower for the Korean learners than for the other language groups, indicating that they have greater difficulty fully working out the visual cues for the English labiodentals. Their lack of accurate use of the visual input may be due to the relatively distinctive auditory input, and resembles the native English group’s performance where the visual benefit is not apparent when the auditory perception reaches ceiling. From an alternative angle, the Korean perceivers’ unbalanced performance in the auditory and visual domain indicates that
grasping auditory cues for the labiodentals may have preceded visual cues. Moreover, the confusion patterns for visual perception show that, compared to native English and Mandarin perceivers, the Koreans more often confused labiodentals with alveolars, a place of articulation native for them but not immediately adjacent to the labiodentals. This may have been due to the lack of place of articulation contrasts for front fricatives in their native Korean, leading to their incorrect assimilation of the L2 visual cues to an L1 category.

The visual perception of alveolars has a low accuracy by the nonnatives and natives alike. As sibilants, alveolars are more acoustically robust than the nonsibilant labiodentals and interdentals (e.g., Jongman et al., 2003). It has been claimed that the degree of visual contribution is inversely related to the ambiguity of auditory information (e.g., Chen & Massaro, 2004). Thus, since the auditory (acoustic) cues of alveolars are robust, they are less dependent on visual cues. The results consistently show very high scores for auditory perception of alveolars. This supports previous observations that a visual benefit is in part determined by the visual salience of the place of articulation, and that consonants which are less front are less likely to have visual influence than relatively front ones (e.g., Dodd, 1977; Hardison, 1999; Hazan et al., 2006).

4.2. Incongruent auditory–visual condition

All three language groups captured auditory information more often than visual information or AV-fusion (i.e., the McGurk illusion), which is consistent with previous findings showing more auditory responses for the AV incongruent /p-/ (Hardison, 1999) and /b-g/ (Chen & Hazan, 2007) combinations. However, as has also been found previously, (e.g., Burnham & Dodd, 1998; Chen & Hazan, 2007; Sekiyama & Tohkura, 1993; Sekiyama et al., 2003), the nonnative perceivers (Mandarin in particular), compared with the native perceivers, have more occurrences of AV-fusion, possibly because the nonnative auditory and visual cues of these sounds are less stable, and therefore more susceptible to cross-modal influence. Compared to the Mandarin perceivers, the Koreans show a lower degree of fusion and their high accuracy in the A-match condition indicates that they have come further in working out auditory cues of the nonnative labiodentals.

Moreover, a direction effect with AV-fusion is also observed with the native English perceivers, with more fused responses for A<sub>labiodental</sub> + V<sub>alveolar</sub> stimuli than A<sub>alveolar</sub> + V<sub>labiodental</sub> stimuli. Given that interdentals share the acoustic feature of “nonsibilant” with labiodentals and that they share the visual feature of “nonlabial” with alveolars, it is natural that fusion (i.e. the interdental response) more easily occurs for native perceivers when the auditory input is labiodental and the visual input is alveolar. Findings consistent with these results have previously been reported where native perceivers demonstrate greater fusion when A-labials are combined with V-nonlabials (McGurk & MacDonald, 1978; Sekiyama & Tohkura, 1991). It is worth noting that this direction effect is less apparent for the Mandarin perceivers than for the Korean and native English perceivers, even though for the Koreans both the input component (labiodental) and the expected fused percept (interdental) are nonnative. This may be due to the Mandarin perceivers’ less stable use of auditory information (compared to the Korean and English natives’ exclusive use of the auditory input), leading them to be more susceptible to the auditory–visual fusion.

4.3. General discussion

The present results from L2 AV speech perception suggest the possibility of bridging learning patterns across speech input modalities. Indeed, the acquisition of nonnative visual cues may be analogous to that of auditory L2 speech learning (Hazan et al., 2005, 2006; Ortega-Llebaria et al., 2001; Sekiyama, 1997). Just as with auditory L2 perception, difficulty with the perception of L2 visual information may also be attributed to influence from an L1. For L2 visual information which is nonassimilable to an L1 category, learners need to learn to associate L2 specific visual cues to corresponding L2 phones in order to establish new L2 categories.

Theories based on auditory perception (e.g., the SLM) suggest that whether new L2 categories can be created depends on the perceived phonetic distance of an L2 sound from the closest L1 speech category: category formation may be blocked under the mechanism of category assimilation, where similar L1 sounds exist in the vicinity of the target L2 sounds (Flege, 2007). Given this account, the current study hypothesizes that perception of the nonnative interdentals can be more difficult for the Mandarin perceivers than for the Koreans, since the Mandarin perceptual space is more crowded in the neighborhood of the interdentals (with both labiodentals and alveolars). Indeed, the results show that the Koreans outperform the Mandarin perceivers in the AV<sub>c</sub> condition, demonstrating more efficient use of visual information when integrating with the auditory input.

Recent AV speech learning theories also suggest that, in addition to similarities in visual cues between an L1 and L2, L2 speech learning models should include further factors to account for AV L2 speech processing, such as the relative weighting of auditory and visual cues, and the distinctiveness of visual cues (Hazan et al., 2006). Indeed, nonnative perceivers may differently weigh the auditory and visual input (Sekiyama, 1997; Ortega-Llebaria et al., 2001). In the present study, the nonnative Mandarin group may have increased visual weighting in the perception of the new L2 interdentals, although they still have difficulty assimilating them to the appropriate L2 categories (e.g., the increased performance in the AV<sub>c</sub> condition compared to the A condition, despite the interdentals still often being confused with the adjacent alveolars familiar to the
Mandarin perceivers in their L1). On the other hand, for the Korean perceivers, the auditory perception of these nonnative sounds is relatively good, which may lead to their lack of use or less accurate use of the visual domain. These findings suggest that the perception and acquisition of L2 sounds in the auditory and visual domain may not occur in parallel, and may even take place in a complementary manner.

5. Concluding remarks and future directions

In sum, while L2 learners make use of both auditory and visual information in perceiving nonnative speech sounds, their perception is influenced by the interaction of the AV speech categories in their L1 and L2. In future research, the correlation of visual perceptual distance between L1 and L2 visual categories, and the corresponding level of difficulty in acquisition (e.g., the more dissimilar the L1 and L2 visual categories are, the easier the formation of an L2 category) need to be more extensively studied and even quantified, as also suggested by the auditory speech learning research (e.g., Flege, 2007; Strange, 1999; Strange, Yamada, Kubo, Trent, & Nishi, 2001). Furthermore, the current study indicates that L2 auditory and visual speech cues may not be acquired simultaneously; rather, AV learning may even occur in a complementary manner. Subsequent research and theories should take into account visual and auditory relationships of L2 speech sounds with corresponding L1 sounds.

Additionally, as the current study focuses on the effect of L1, it has left unaddressed a number of factors that may also affect the perception of L2 speech contrasts, such as L2 phonetic context and variability (Hardison, 2005b), length of residence (LOR) in an L2 country and L2 input (Flege, 2007), attention (Guion & Pederson, 2007), etc. One factor worth noting is phonetic context. For example, previous research has shown that vowel context influences the neighboring sounds not only for native AV perception (e.g., Benguerel & Pichora-Fuller, 1982; Daniloff & Moll, 1968) but also for nonnative perception (e.g., Hardison, 2005b). Since the vowels used in the current study exist in the native phonetic inventories of all three native groups, no interaction of L1 group and vowel context was observed. A possible future direction is to examine whether unfamiliar (nonnative) and familiar vowel context would similarly influence the perception of the target L2 sounds. Second, although the current nonnative participants all had a LOR between 1 and 5 years, the average LOR for the Korean participants (4 years) was longer than that for the Mandarin participants (2 years). Given that research has demonstrated the effect of LOR on L2 AV (Wang et al., 2008) as well as auditory speech learning (e.g., Flege, Yeni-Komshian, & Liu, 1999; McAllister, Flege, & Piske, 2002; Riney & Flege, 1998), the 2 years’ LOR difference could have led the Koreans to outperform the Mandarin participants in the current results. On the other hand, the Koreans’ reported mean daily English use (34%) was a bit lower than that of the Mandarin natives (40%), which may have balanced out the slight LOR difference. Further longitudinal research would be needed to examine the dynamic auditory and visual learning patterns as a function of LOR (e.g., from 0 to 5 years) and L1/L2 input. Moreover, to compare the relative auditory and visual contribution, it would be necessary to evaluate the degree of attention to the speech input from these two modalities. Techniques such as eye-tracking which had previously been adopted in native AV processing research (e.g., Lansing & McConkie, 1999) would provide quantitative means of where exactly the perceivers’ eyes focus.

Together, the findings of the current study along with the previous L2 AV research suggest a shared mechanism in L2 speech learning for the auditory and visual domains, providing supporting evidence to extend L2 speech learning theories developed on the basis of auditory speech perception to include visual speech perception, and thus bridging the L2 perceptual learning patterns across AV modalities.

Acknowledgments

We thank Nicole Carter, Chad Danyluck, Angela Feehan, Lan Kim, Lindsay Shaw, Rebecca Simms, and Jung-yueh Tu at Simon Fraser University (SFU) for their assistance in stimuli development, data collection and analysis. We would also like to thank Dr. Ocke-Schwen Bohn and the two anonymous reviewers for their valuable comments. Portions of this research were presented at the 2007 International Conference on Auditory–Visual Speech Processing (AVSP 2007), and published in the Proceedings of AVSP 2007 (Eds.: J. Vroomen, M. Swerts and E. Krahmer). Portions of the English and Mandarin data have been included in a previous article investigating the effect of length of residence in an L2 speaking environment on AV speech perception (Wang et al., 2008). This project was supported by a standard research grant from the Social Sciences and Humanities Research Council of Canada (SSHRC 410-2006-1034).

References


