Linguistic experience and audio-visual perception of non-native fricatives

Yue Wang
Department of Linguistics, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada

Dawn M. Behne
Department of Psychology, Norwegian University of Science and Technology, N7491 Trondheim, Norway

Haisheng Jiang
Department of Linguistics, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada

(Received 7 December 2007; revised 13 June 2008; accepted 13 June 2008)

This study examined the effects of linguistic experience on audio-visual (AV) perception of non-native (L2) speech. Canadian English natives and Mandarin Chinese natives differing in degree of English exposure [long and short length of residence (LOR) in Canada] were presented with English fricatives of three visually distinct places of articulation: interdental nonexistent in Mandarin and labiodentals and alveolars common in both languages. Stimuli were presented in quiet and in a café-noise background in four ways: audio only (A), visual only (V), congruent AV (AVc), and incongruent AV (AVi). Identification results showed that overall performance was better in the AVc than in the A or V condition and better in quiet than in café noise. While the Mandarin long LOR group approximated the native English patterns, the short LOR group showed poorer interdental identification, more reliance on visual information, and greater AV-fusion with the AVi materials, indicating the failure of L2 visual speech category formation with the short LOR non-natives and the positive effects of linguistic experience with the long LOR non-natives. These results point to an integrated network in AV speech processing as a function of linguistic background and provide evidence to extend auditory-based L2 speech learning theories to the visual domain.

© 2008 Acoustical Society of America. [DOI: 10.1121/1.2956483]

PACS number(s): 43.71.Hw [AJ]

I. INTRODUCTION

Language experience often involves face-to-face interaction with simultaneous perception of a speaker’s voice and facial movements. For native (L1) perceivers speech perception is enhanced with visual information (e.g., Erber, 1969; Sumby and Pollack, 1954), especially when auditory distinctiveness decreases, such as in a noisy environment (Bernstein et al., 2004; Erber, 1969; Sumby and Pollack, 1954; Summerfield, 1979). The relative contribution of audio and visual information has also been revealed by what is known as the “McGurk effect,” where an audio [ba] dubbed onto a visual [ga] may produce a [da] percept (McGurk and MacDonald, 1976). These results suggest an ability to integrate auditory and visual speech information (Massaro, 1987, 1998; Sekiyama et al., 2003). A question that arises is whether this ability reflects an innate capacity to process multimodal information or is developed by learning and experience such that the information processing is instantiated by language-specific patterns.

A hardwired ability to process audio and visual information has been shown for prelinguistic infants: 2–4 month olds prefer speech information matching, compared to non-matching, in auditory and visual information (Kuhl and Meltzoff, 1982; Patterson and Werker, 1999, 2003) and can integrate auditory and visual speech information with incongruent input modalities (Burnham and Dodd, 2004; Desjardins and Werker, 2004; Rosenblum et al., 1997). However, while 4–6-month-old infants can visually distinguish their L1 speech from an unfamiliar language, 8 month olds cannot (Weikum et al., 2007), suggesting a shift from language-universal to language-specific multimodal speech processing as infants are exposed to a specific language. Cross-linguistic studies with adults have also shown language specific, as well as language-universal patterns. For example, while Chinese, English, Japanese, and Spanish perceivers demonstrate similar audio-visual processing patterns when perceiving a [ba]-[da] continuum (Chen and Massaro, 2004; Massaro et al., 1993), Chinese perceivers show worse visual perception of [ba] and [da] than Dutch perceivers (de Gelder and Vroomen, 1992). Similarly, although the McGurk effect has consistently been observed across languages (Sams et al., 1998; Sekiyama and Tohkura et al., 1993; Werker, Frost, and McGurk, 1992), native speakers of some languages such as Chinese and Japanese (Sekiyama, 1997; Sekiyama and Tohkura, 1993) show a weaker magnitude of this effect than do English or Spanish perceivers (Massaro et al., 1993). It has been speculated that Chinese speakers’ relatively diminished use of visual speech information may be due to Chinese being a tonal language and thereby relying on the visually less distinct prosodic information (Sekiyama, 1997; Sekiyama and Tohkura, 1993), although significant visual in-
formation for tones in Chinese has also been observed (Burnham et al., 2001). More recent research has nevertheless demonstrated that Mandarin Chinese perceivers can use the visual speech information in their L1 to the same degree as English perceivers do (Chen and Hazan, 2007a). These discrepancies suggest the need for further cross-linguistic research to address the issue of language specificity of audio-visual speech processing.

The perception of non-native (L2) speech provides a unique case. On the one hand, visual speech information when available may enhance the perception and production of an L2 (Massaro et al., 1993; Massaro, 1998). For example, with the addition of visual information, English perceivers demonstrate more accurate perception of French (Reisberg et al., 1987), Korean (Davis and Kim, 2004), Irish, and Spanish sounds (Erdener and Burnham, 2005). Similarly, Spanish-dominant Spanish-Catalan bilinguals can distinguish the Catalan vowel contrast [ɛ-ɛ] in an audio-visual condition but not with audio-only or visual-only input (Navarra and Soto-Faraco, 2007). For Korean and Japanese learners of English, perception of non-native [f] and [l] improves with matched auditory and visual input compared to audio-visual mismatched stimuli (Hardison, 1999), and the identification of the most difficult auditory stimuli receives the greatest benefit from visual cues (Hardison, 2005a).

Nevertheless, non-natives are less efficient at using visual information than natives (de Gelder and Vroomen, 1992; Hazan et al., 2005, 2006; Ortega-Llebaria et al., 2001; Sekiyama, 1997). For example, unlike native Spanish and Catalan perceivers, Italian and English natives cannot discriminate Catalan and Spanish when only visual information is available (Soto-Faraco et al., 2007). Non-natives, such as L2 perceivers of English and Japanese, appear to be more affected by the McGurk illusion for stimuli presented in an L2 than in an L1 (Sekiyama and Tohkura, 1993). Exceptions to this are Chinese perceivers showing an equally weak McGurk effect for their L1 and L2 (Japanese or English), suggesting language-specific differences based on a perceiv-er’s L1 (Sekiyama, 1997; Sekiyama, Tohkura, and Umeda, 1996). Notably, in these studies with native Japanese and Chinese perceivers, all the target stimuli involve audio-visual phonemes common to both the L1 and L2, which may not correspond to perceptual patterns of new L2 audio-visual stimuli. Indeed, visual information would more likely be beneficial for phonemes occurring both in L1 and L2, whereas the addition of visual information would not so easily reduce confusion for visual cues specific to new L2 phones (Ortega-Llebaria et al., 2001). For example, Spanish learners show more sensitivity than Japanese learners to the visual cues to English [v], because the visual category for the labiodental [v] is allophonically existent in Spanish and the voiceless counterpart [f], whereas it is totally missing in Japanese (Hazan et al., 2006). These findings indicate that L2 learners have difficulty in correctly using the visual cues that do not contrast phonetically in their L1.

The patterns of audio-visual perception are reminiscent of the L2 speech learning theories developed on the basis of auditory perception [e.g., speech learning model (SLM), Flege, 1995; the perceptual assimilation model (PAM), Best, 1995], stating that L2 learners’ inability to accurately perceive L2 sounds is due to incorrect assimilation of L2 sounds with L1 phonetic categories and therefore failure in establishing new L2 phonetic categories (Best et al., 1988; Best, 1995; Flege, 1987, 1995). In this respect, the acquisition of L2 visual categories (termed “visemes,” Brooke and Summerfield, 1983; Walden et al., 1977) may be comparable to that of auditory L2 sound learning (Erdener and Burnham, 2005; Hardison 1999, 2003; Hazan et al., 2005, 2006). L2 perceivers may have lost their sensitivity to visual cues that are not used in their L1 and need to learn to associate these visual cues with a viseme in the L2 (Hazan et al., 2005, 2006) to establish new L2 categories.

Visual speech learning may occur with linguistic experience. Sensitivity to visual information in L2 sounds can be enhanced through audio-visual training, with the degree of improvement affected by one’s L1 (Hardison, 2003, 2005b; Hazan et al., 2005). Research has also shown a positive correlation between overall L2 proficiency and proficiency in visual L2 speech perception. Werker et al. (1992) examined the visual perception of incongruent audio-visual English fricatives by native French perceivers with varying English proficiency and found that for the interdental fricative absent in French, French speakers’ degree of fluency in English was positively related to their correct use of the English visual information.

In sum, these findings indicate that the difficulty in L2 visual perception of speech may lie in visual speech categories that are not used in the L1, suggesting that subsequent research should focus on new L2 visual categories. Furthermore, the presence of cross-linguistic differences also suggests the need for further research to include additional L2 learners (such as Chinese learners who have been found to use less visual information in their L1) to examine possible language-specific differences. In addition, if linguistic experience plays a role in L2 visual perception, further studies are necessary to identify the kinds of linguistic experience that may be facilitative. Moreover, nonlinguistic factors may also affect L2 audio-visual processing. For example, given that L1 visual speech perception is particularly effective in non-optimal conditions, such as in noise (e.g., Sumby and Pollack, 1954; Summerfield, 1979), how noise affects the perception of new L2 visual speech information will further illuminate the role of language background in audio-visual processing.

A. The current study

On the basis of these previous findings, the current research examines auditory and visual perception of the English interdental fricatives in quiet and noisy backgrounds by native Mandarin Chinese perceivers varying in length of residence (LOR) in Canada.

Firstly and primarily, this study focuses on the visual perception of interdental fricatives and hypothesizes that English interdental non-existent in Mandarin perceivers’ L1 are more difficult to perceive than the place-adjacent labioden-tals and alveolars familiar to them in Mandarin. Previous research on Mandarin natives’ perception of speech visemes


Wang et al.: Non-native audio-visual fricative perception
common to their L1 and L2 (e.g., Sekiyama et al., 1996, Sekiyama, 1997) may not reflect the perception of new L2 visemes. Since the difficulty of L2 visual speech perception lies in the new L2 visual speech categories (Hazan et al., 2005, 2006), interdental fricatives are used in the current study. Furthermore, since Mandarin perceivers demonstrate less use of visual speech information in their L1 compared to perceivers of other languages (e.g., Burnham et al., 2001; de Gelder and Vroomen, 1992; Sekiyama et al., 1996; Sekiyama, 1997), questions arise as to how this affects their visual perception of L2 sounds.

The second focus is the effect of L2 experience indexed by the LOR in an L2 environment. LOR, which has been used as an index in L2 auditory speech learning research (e.g., Flege, 1995; 1998; Flege et al., 1995; Flege et al., 1999; McAllister et al., 2002; Riney and Flege, 1998), may be particularly relevant for visual speech perception since L2 learners typically do not have extensive exposure to L2 visual cues until they arrive in an L2 country. Based on findings that L2 visual perception improves with short-term training (Hardison, 2003; Hazan et al., 2005) and overall L2 proficiency (Werker et al., 1992), the current study hypothesizes that the short LOR Mandarin perceivers outperform the short LORs in perceiving the English visemes nonexistent in Mandarin. Furthermore, differences observed between the two Mandarin groups will inform patterns of visual speech learning in progress.

Additionally, this study addresses the influence of background noise. For native perceivers, the contribution of visual input is especially effective in a noisy environment (Bernstein et al., 2004; Sumby and Pollack, 1954; Summerfield, 1979). L2 research on the auditory domain shows that non-native listeners are more affected by noise than native listeners, with a reduction in intelligibility when speech is degraded in noise (Hazan et al., 2005; Mayo and Florentine, 1997; Sekiyama et al., 2003). On this basis, L2 perceivers may particularly need to rely on visual speech information. However, if they cannot effectively pick up visual speech information in the L2, their performance would be similar in quiet and noisy backgrounds. In other words, if audio-visual processing is not language-specific, noise would be expected to affect non-native and native perceptions in the same way.

Alternatively, language-specific audio-visual processing would result in different processing patterns in L1 and L2 as a function of listening backgrounds.

II. METHODS

A. Participants

Three groups of young adults (mean age = 22) participated in the study: one group of native Canadian English participants (n = 15), and two groups of native Mandarin Chinese differing in their LOR in Canada (long LOR: n = 15; short LOR: n = 20). The participants’ background information is summarized in Table I. The “long LOR” Mandarin group consisted of native Mandarin English users who had resided in the country for an average of 10 years. They moved to Canada, on average, at age 11, when they were initially exposed to English. All reported having English as their dominant language since their arrival and having attended Canadian secondary schools where English was the language of education. The Mandarin “short LOR” group had an average of two years’ residency in Canada. Before arriving in Canada, they had studied English as an L2 in a classroom setting (5 hours/week) since they were on average 12 years old and had not resided in any other English-speaking environments prior to their arrival in Canada. Based on their reported language background, they continued to use Mandarin as their more dominant language since their arrival. These selection criteria ensure a comparable age range between the two Mandarin groups at the time of testing (21–24 years old), comparable age of initial English learning (11–12 years old), and comparable length of English learning (12 years). All three groups reported having normal or corrected vision and no known history of speech and hearing impairments. All were undergraduate or graduate students at Simon Fraser University (SFU), Canada at the time of the study. They were compensated for their participation.

B. Stimuli

Stimuli were based on 18 English consonant-vowel (CV) syllables having a fricative onset followed by a vowel: [fi, fa, fu, vi, va, vu, ði, ða, ðu, ði, ða, ðu, si, sa, su, zi, za, ...]

TABLE I. Participants’ language background information. [LOR: mean length of residence in Canada (years); AOA: mean age of arrival in Canada; AOL: age of initial English learning; English study: mean number of years of formal English study; Mandarin input: mean percent daily input in Mandarin (L1); English input: mean percent daily input in English (L2). Means are based on averages across participants, and the range across participants is included in parentheses.]

<table>
<thead>
<tr>
<th>Language</th>
<th>Age (yrs)</th>
<th>Gender (F/M)</th>
<th>LOR (yrs)</th>
<th>AOA (yrs)</th>
<th>AOL (yrs)</th>
<th>English study (yrs)</th>
<th>Mandarin input (%)</th>
<th>English input (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>22 (n = 15)</td>
<td>9F/6M</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Mandarin long LOR (n = 15)</td>
<td>21 (18–25)</td>
<td>10F/5M</td>
<td>10</td>
<td>11</td>
<td>11</td>
<td>12</td>
<td>38</td>
<td>62</td>
</tr>
<tr>
<td>Mandarin short LOR (n = 20)</td>
<td>24 (19–25)</td>
<td>15F/5M</td>
<td>2</td>
<td>21</td>
<td>11</td>
<td>12</td>
<td>60</td>
<td>40</td>
</tr>
</tbody>
</table>
The fricatives differed in place of articulation (labiodental, interdental, alveolar), representing a sequence of audio-visual categories involving articulators from more front to further back in the vocal tract. While all three places of articulation are phonemically distinctive in English, the labiodental and alveolar fricatives occur in Mandarin (as voiceless), whereas the interdentals do not occur in Mandarin. The voiced and voiceless counterparts of each syllable were included to allow a broad test of place of articulation identification. The vowels [i, a, u] occur in Mandarin as well as in English, representing a range of vocal tract configurations varying in tongue height, advancement, and lip rounding (Hazan et al., 2006; 2005; Jongman et al. 2003). Thus each target fricative place of articulation was represented by six different exemplars, to ensure responses to phonetic categories rather than acoustic idiosyncrasies.

On this basis, stimuli were developed which had (1) congruent audio and visual components (AVc), (2) only an audio component (A), (3) only a visual component (V), and (4) incongruent (mismatched) audio and visual components (AVi) (see McGurk and MacDonald, 1976). In the AVi stimuli, the fricative place of articulation was an audio-visual combination of labiodental and alveolar ([A]\textsubscript{labiodental}−[V]\textsubscript{alveolar}], e.g., audio [fa] dubbed onto video [sa], [A]\textsubscript{fa}−[V]\textsubscript{sa}, or [A]\textsubscript{alveolar}−[V]\textsubscript{labiodental}], e.g., audio [sa] dubbed onto video [fa], [A]\textsubscript{sa}−[V]\textsubscript{fa}). If AV-fusion occurs, the percept is expected to have a place of articulation intermediate to labiodental and alveolar, which is the interdental place non-native to Mandarin perceivers (e.g., [8a]). The fricative voicing and vowel were always the same for the A and V components of a given AVi syllable (e.g., [A]\textsubscript{fa}−[V]\textsubscript{sa}, [A]\textsubscript{fa}−[V]\textsubscript{sa}, and [A]\textsubscript{sa}−[V]\textsubscript{fa}). A total of 12 AVi stimuli were used (two AV input place [A]\textsubscript{labiodental}−[V]\textsubscript{alveolar}, [A]\textsubscript{alveolar}−[V]\textsubscript{labiodental]×two voicing conditions [voiceless, voiced]×three vowels [i, a, u]).

1. Production

Audio and video recordings were made in the Language and Brain Lab at SFU with an adult male speaker of Canadian English producing six repetitions of the 18 syllables (six fricatives×three vowels) at a normal speaking rate. Video recordings of the speaker’s face were made using a digital camcorder (Sony DCR-HC30/40) placed about 3 m from the speaker against a white background. Separate audio recordings were simultaneously made using a Shure KSM 109 condenser microphone to a PC via an audio interface (M-audio MobilePre USB) at a 44.1 kHz sampling rate. These high quality audio recordings were used to replace the audio track from the camcorder recording.

2. Editing

The video recordings with their camcorder audio recordings were transferred to a PC. From among the six repetitions of each syllable, a best example of each syllable was selected 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 video recordings were aligned with corresponding high quality audio recordings using SOUNDFORGE 8.0. This was done by synchronizing the waveforms of the syllable recorded from the camcorder and the Shure microphone. The audio signal from the camcorder was then deleted. The final set of audio stimuli was normalized so that all the resulting stimuli had the same unweighted rms value. Visual stimuli were edited such that each clip had a 1.2 s neutral face before and 1.2 s after the stimulus, with the length of the stimulus measured from the frame where mouth movement was first detected (mouth opening) to the frame where the mouth movement ended and the mouth was fully closed. The frame length was 0.06 s, and the resolution was 640×480 pixels.

The resulting AV stimuli were used as AVc stimuli and as the basis for creating the A-only, V-only, and AVi stimuli. The A and V sets of stimuli were created from the AVc files by removing the video tracks or muting the audio tracks, respectively. For the AVi stimuli, the A and V components were the same as those used in the AVc, A-only, and V-only conditions. Each AVi stimulus was created by aligning audio and video components from different syllables (e.g., [A]\textsubscript{fa}−[V]\textsubscript{fa}). Thus each AVi [A]\textsubscript{fa}−[V]\textsubscript{fa}

3. Goodness evaluation

Stimulus goodness evaluation tests were carried out on the final sets of AVc and AVi stimuli to confirm the intelligibility of the audio signals and the naturalness of the audio-visual signals, each with two phonetically trained native speakers of English. Intelligibility of the audio signals was tested with an identification task, showing 95% correct responses for the audio signals in the AVc stimuli and 100% correct responses for the audio signals in the AVi stimuli. The naturalness of the AV stimuli was determined by asking the evaluators whether the audio and video signals were naturally synchronized, regardless of what they heard or saw. This was carried out with a 5-point goodness rating task (5 being the most natural), showing that the AVc stimuli were rated 4.4 and the AVi stimuli rated 4.6.

4. Stimulus background

Finally, cafeteria noise recorded at SFU was added to a copy of all the stimuli, resulting in a quiet and café-noise (S/N=0 dB) version for all the syllables. The signal-to-noise (S/N) ratio was established empirically by testing two English listeners at each of the following S/N values: 10, 0, and −10 dB. A S/N ratio of 0 dB was found to generate an error rate of about 30%. At 10 dB, identification was similar to the quiet background condition, whereas at −10 dB, signals, especially the fricatives, were completely masked. Therefore, cafeteria noise at 0 dB S/N ratio was embedded in each of the CV syllables, with the duration of the noise corresponding to the syllables’ duration, to create the set of stimuli in the noisy condition (A-noise, V-noise, AVc-noise, AVi-
noise). In total, 132 stimuli were used, including 66 in quiet (18 A, 18 V, 18 AVc, 12 AVi) and 66 in a café-noise background.

C. Procedure

The perception experiment was generated in E-PRIME 1.0 (Psychology Software Tools), allowing stimulus presentation (integrating video clips imported from Microsoft Powerpoint files) and response log. An identification task was carried out where each participant was tested on the full set of stimuli in the quiet and noisy backgrounds. Stimuli were blocked by modality (A, V, and AV) and background (quiet and noise), with AVi and AVc stimuli in the same block. Each of these six blocks included two repetitions of each stimulus. Stimuli were randomized within a block. The order of background and modality presentation was counterbalanced across participants.

For each trial a stimulus was presented auditorily over loudspeakers, visually on a computer monitor, or both. The participants were tested individually, seated about 1 m from a 20 in. liquid crystal display (LCD) flat panel computer monitor and two loudspeakers (Altec Lansing) positioned directly at each side of the monitor, such that the sound and image source had approximately the same distance to the perceivers’ ears and eyes. Loudspeakers rather than headsets were used to prevent perceivers from being biased toward the audio input. The audio signal was presented at a comfortable level of approximately 70 dB, found to be sufficient to achieve an optimal score (Nábelek and Robinson, 1982; Takata and Nábelek, 1990).

For each trial, a fixation point was displayed in the center of the monitor for 1 s, followed by the target stimulus. Response alternatives were presented on the monitor, where, for a given trial, the alternatives were the six fricatives [f, v, θ, δ, s, z], as well as the option to type in an alternative response. The symbols “th” and “dh” were used to represent [θ] and [ð], respectively (Jongman et al., 2003). The participants’ task was to identify the fricative and respond by pressing a key on the computer keyboard for the corresponding symbol displayed on the monitor. Participants had up to 4 s to respond.

The test was preceded by a short introduction session containing task instructions, stimulus familiarization (e.g., matching the symbols and the sounds they represent), and five practice trials for each modality (A, V, and AV). The test ended with a debriefing session during which participants filled out a post-test questionnaire. The whole test lasted about 1 h, including a short break after three test blocks.

III. RESULTS

Data were analyzed separately for the audio, visual, and audio-visual congruent conditions (A, V, and AVc), and for the audio-visual incongruent condition (AVi).

A. Audio, visual, and congruent audio-visual conditions

The percent correct identification of the fricatives was analyzed using a four-way mixed analysis of variance (ANOVA), with group (English, Mandarin long LOR, and Mandarin short LOR) as a between-subject factor, and place of articulation (POA) (labiodental, interdental, and alveolar), modality (A, V, and AVc), and background (quiet and noise) as repeated measures. The dependent variable was the perceivers’ correct identification of POA regardless of consonant voicing and vowel context (Hazan et al., 2005, 2006; Jongman et al., 2003; Werker et al., 1992) since POA was the focus of interest.

Table II presents mean percentage of correct responses for the English, Mandarin long, and short LOR groups as a function of POA, modality, and background. A significant main effect of group was found \[F(2, 47) = 17.6, p < .001]\, with the post hoc analysis (Tukey HSD), showing that native English perceivers had a higher overall percentage of correct responses than the Mandarin long LOR group, which in turn had a higher percentage of correct responses than the Mandarin short LORs. Significant main effects were also observed for POA \[F(2, 47) = 30.2, p < .001]\, modality \[F(2, 47) = 183.1, p < .001]\, and background \[F(1, 47) = 381.1, p < .001]\.

Across participant groups, identification was better with labiodentals than with interdentals or alveolars, better in the AV than in the A or V condition, and better in quiet than in noise.

Moreover, native group differences were observed with significant interactions of group \(\times\) modality \[F(4, 47) = 2.9, p = .027]\, group \(\times\) POA \(\times\) modality \[F(8, 47) = 2.5, p = .013]\, and group \(\times\) POA \(\times\) modality \(\times\) background \[F(8, 47) = 6.1, p < .001]\.

Further analyses were thus conducted to compare group differences for POA, modality, and background. Figure 1 displays the results.
First, sets of two-way (modality and group) repeated measure ANOVAs were conducted for each POA in quiet to compare how the AV modalities were perceived as a function of group. Only interdental identification showed a modality and group interaction \( F(4,44)=4.0, p=.005 \). Post hoc analyses show that whereas the English natives’ identification in the V condition was poorer than the A and AVc conditions, the Mandarin long LOR group’s A and V identifications were poorer than AVc identification, and the short LOR group did poorer in the A (than V than AVc) modality. To compare how different fricative POAs were perceived, sets of two-way ANOVAs were carried out in each modality in quiet. A POA and group interaction was observed in the A [\( F(4,44)=6.8, p<.001 \)] and AV [\( F(4,44)=3.4, p<.001 \)] conditions. In the A condition, whereas the English natives’ identification of both the interdents and labiodentals was moderately lower than for the alveolars, both the Mandarin long and short LOR groups’ identification for the interdents was significantly lower than for the labiodentals and alveolars. For the AV stimuli, while the native English and Mandarin long LOR groups did not differ in their performance for the three POAs, the Mandarin short LOR group had a much lower percentage of correct responses for the interdentals than labiodentals and alveolars.

To compare group differences directly, sets of one-way ANOVAs were carried out with group as a between-subject factor in each modality and POA in quiet. Differences were observed for the interdents and alveolars in the A and AVc conditions. For the interdents, the A condition revealed a decreasing identification accuracy from the native English to Mandarin long LOR to short LOR groups \( F(2,47)=23.6, p<.001 \), and in the AVc condition, both the English and the Mandarin long LOR group’s accuracy was higher than that of the short LOR group’s \( F(2,47)=8.3, p<.001 \). For the alveolars, both the native English and Mandarin long LOR group’s percentages of correct responses were higher than the Mandarin short LOR group’s \( [A: F(2,47)=10.1, p<.001; AVc: F(2,47)=14.0, p=.001] \). The V condition, however, did not reveal a significant group difference for interdents (all being moderately low) or alveolars (all being low). For the labiodentals, no difference was observed across modalities, all of which were relatively high.

Given that most of the between group differences occurred with interdents, confusion patterns of interdental identification were analyzed (Table III). In both A and AV conditions, the Mandarin short LOR group misperceived 18–19% of the interdents as alveolars, while the English natives never confused these two POAs in quiet. Similarly, in the V condition, the short LOR group was also more biased toward the alveolars (30%) than toward the labiodentals (9%), while the English natives misperceived the interdents as labiodentals or alveolars to a similar degree. The Mandarin long LOR group generally followed the native patterns. It is also noted that although the participants were given the option to type in an alternative to the six fricatives, this “other” option was not much used (1%–3% across conditions and participant groups).

For stimuli in noise, cross-modality comparisons show that significant group differences only occurred with interdents: while an increase in accuracy from A-noise to V-noise to AVc-noise was observed for both the English natives \( F(2,42)=12.1, p<.001 \) and the Mandarin long LOR group \( F(2,42)=8.6, p=.001 \), the short LOR group only
showed a marginal difference across modalities \([F(2, 57) = 3.1, p = .051]\), all being low. Direct comparisons of the noise and quiet conditions for each POA and modality show that while the perception was generally poorer in noise than in quiet \((p < .05)\), the Mandarin short LOR group’s auditory perception of interdental did not differ between the noise and quiet conditions \([F(1, 19) = 1.5, p = .240]\), both being relatively low.

### B. AV incongruent condition

For each AVi stimulus, corresponding responses were tabulated based on whether the consonant in the response matched the consonant in the audio component of the stimulus \((A\text{-match})\), the video component \((V\text{-match})\), or the fused \((A\text{~V-fusion})\) and background as repeated measures. Figure 2 displays the mean percent responses for each response type as a function of AV place input in quiet and noisy backgrounds.

Results for AV-fusion showed a significant effect of AV-place \([F(1, 47) = 31.3, p < .001]\) and a group \(\times\) AV-place interaction \([F(1, 47) = 6.3, p = .004]\). A moderate degree of

<table>
<thead>
<tr>
<th>Perceived as</th>
<th>Interdental stimuli</th>
<th>Mandarin long LOR</th>
<th>Mandarin short LOR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>V</td>
<td>AVc</td>
</tr>
<tr>
<td>Quiet</td>
<td>Labiodental</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Interdental</td>
<td>89</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>Alveolar</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>Noise</td>
<td>Labiodental</td>
<td>24</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Interdental</td>
<td>54</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>Alveolar</td>
<td>22</td>
<td>11</td>
</tr>
</tbody>
</table>

FIG. 2. Mean percent responses for incongruent AV stimuli in (a) quiet and (b) noise backgrounds by English (black bars), Mandarin Chinese long LOR group (dark gray bars), and short LOR group (light gray bars). “A-match” and “V-match” indicate correct responses matching A and V components, respectively. “AV-fusion” refers to interdental responses matching neither the A nor the V component, corresponding to the McGurk effect. “*” represents significant \((p < .05)\) statistical results comparing among the three participant groups for each place of articulation and input modality.
fused interdental responses (25%) was observed across groups for the $A_{labiodental} + V_{alveolar}$ condition where one would expect the McGurk effect. However, in the $A_{alveolar} + V_{labiodental}$ input condition, the Mandarin short LOR group had a higher mean percent than the long LOR group whose responses were in turn greater than the English group [$F(2,47) = 12.2, p < .001$], indicating that the Mandarin perceivers more easily fused the incongruent stimuli despite the intermediate fricatives being non-native. Furthermore, a reliable effect was observed across groups for background [$F(1,47) = 55.3, p < .001$], with greater fusion in noise than in quiet.

For percent responses matching the A component with $A_{alveolar} + V_{labiodental}$ stimuli, the English group had a higher mean score than the Mandarin long LOR group, which in turn was higher than the short LOR group [quiet: $F(2,47) = 5.1, p = .001$] [noise: $F(2,47) = 4.6, p = .015$], indicating that the native group was more accurate in using the auditory component. For the V-match responses, on the other hand, a group difference was observed with $A_{labiodental} + V_{alveolar}$ stimuli, where in both quiet and noise backgrounds, the short LOR group scored higher than long LOR and English perceivers [quiet: $F(2,47) = 5.03, p = .010$], [noise: $F(2,47) = 3.9, p = .027$].

C. Summary of results

Results across participant groups show identification in general to be better with AV stimuli than with A- or V-only stimuli and better in quiet than in noise. Responses to the incongruent AV stimuli were dominated by the A input, but the likelihood to use V information increased in noise. Significant group differences were nonetheless evident, particularly with the interdental fricatives nonexistent in Mandarin, showing decreasing identification accuracy from the native English to Mandarin long LOR to short LOR groups. While the English natives’ interdental identification in the A condition was poorer than that in the V condition only in noise, the Mandarin short LOR group showed poorer use of A than V information in quiet as well as in noise. Although both the Mandarin long and short LOR groups’ performances improved in the AVc condition, only the long LOR group achieved the native level. In the incongruent condition, the Mandarin perceivers, particularly the short LOR group, were more likely to give an AV-fused interdental response than the English perceivers. These results indicate that the native English group relied more on the A information, with the V information coming in to facilitate perception in noise. The Mandarin short LOR group, on the other hand, appears to have made more use of V than A information in the perception of the non-native interdentals. The results also reveal the effect of LOR, with the Mandarin long LOR group’s performance being intermediate to that of the English and Mandarin short LOR groups.

IV. DISCUSSION

A. Native and universal patterns

Consistent with previous findings of similar consonant and fricative perception (Chen and Hazan, 2007a, 2007b; Hardison, 1999; Jongman et al., 2003; Massaro, 1998; Sekiyama and Tokhura, 1993; Werker et al., 1992), results show that the English natives can accurately perceive the fricatives in quiet in the A-only and AVc condition. That the perception (especially of the interdentals and alveolars) is poorer in the V-only than in the other conditions also agrees with previous research (Chen and Hazan 2007a, 2007b; Cienkowski and Carney, 2002; McGurk and McDonald, 1976). Similarly, as has been reported (Erber, 1969; Hardison, 1999; Summerfield, 1979), native English perceivers are able to effectively adopt the V cues when the signal is masked in noise.

It has been argued that some aspects of AV speech perception may be neutral across languages (Chen and Massaro, 2004; Massaro, 1998). The present findings show that the native and non-native groups all actively use the audio component, especially in the incongruent AV condition and in the perception of alveolars, which as sibilants are auditorily (acoustically) more robust than the nonsibilant labiodentals and interdentals. On the other hand, all groups use visual cues when available, but reveal poor performance for the visually presented sounds that are not visually distinct (e.g., alveolars), suggesting that a visual benefit is in part determined by the visual salience of the place of articulation (e.g., Dodd, 1977; Hardison, 1999; Hazan et al., 2006). Additionally, according to the fuzzy logical model of perception (FLMP) (Massaro, 1998), although the input information may differ for specific languages, the operational processes of AV perception are universal across languages, similar to those for the processing of low-level nonlinguistic auditory and visual information. Thus, visual speech information may similarly affect perceivers of different languages, with the degree of its influence being largest to the extent that the audio information is ambiguous (Chen and Massaro, 2004). In the present results, supporting evidence is shown by the similarity between the AV perception of non-native sounds (by Mandarin perceivers) and that of the native sounds in noise (by English perceivers), where perceivers tend to pay more attention to the visual speech input in nonoptimal listening conditions, be it involving new non-native auditory information, or in a poor listening environment.

B. Non-native patterns

The non-native perception of labiodentals and alveolars has a general pattern similar to the English natives’, supporting that visemes common to the L1 and L2 do not present problems in L2 perception (e.g., Hardison, 1999; Hazan et al., 2005, 2006; Werker et al., 1992).

Of particular interest is the Mandarin perceivers’ identification of the non-native interdentals. It has been reported that native Mandarin perceivers use less visual speech information in their L1 compared to perceivers of other languages, thus the prediction that Mandarin perceivers may not be sensitive to the L2 visual speech information (Burnham et al., 2001; de Gelder and Vroomen, 1992; Hazan et al., 2005; Sekiyama et al., 1996; Sekiyama, 1997). The current results, however, show that Mandarin perceivers’ (especially the short LOR group’s) correct interdental identification in
the V-only condition is greater than that in the A-only condition, and they show a greater use of V information and greater magnitude of AV-fusion in the incongruent condition than native English perceivers. These results suggest that the Mandarin perceivers made greater use of L2 visual speech information, despite not weighing the visual input heavily in their L1. Indeed, previous research has shown that the perception of L2 stimuli improves with additional visual information (Hardison, 1999) and that visual cues enhance L2 speech comprehension (Navarra and Soto-Faraco, 2007; Reisberg et al., 1987; Soto-Faraco et al., 2007). Consistently, Japanese perceivers who do not weigh visual speech information heavily in their L1 also demonstrate more AV-fusion in perceiving English than in perceiving Japanese sounds (Sekiyama et al., 1996). Given that perceivers rely more on visual speech information when auditory intelligibility is poor (Erber, 1969; Sumby and Pollack, 1954; Summerfield, 1979), L2 perceivers conceivably resort to the visual information as an additional channel of input in perceiving the difficult non-native sounds (Hattori, 1987; Hardison, 2003).

The Mandarin perceivers’ interdental identification is nevertheless poor across input modalities, leading to the critical question of whether non-native perceivers simply make greater use of visual information or if they can adopt the L2 specific visual cues. For the incongruent condition, the Mandarin (short LOR) perceivers have more occurrences of AV-fusion (with the fused interdental being non-native) than the English perceivers, indicating that non-natives are more vulnerable to the AV illusion. Even though they make greater use of the visual input, they cannot effectively use these visual cues in a linguistically meaningful manner. In most of the earlier studies with similar results (e.g., Burnham and Dodd, 1998; Chen and Hazan, 2007a, 2007b; Sekiyama and Tokhura, 1993; Sekiyama et al., 2003), the fused sound often has a place of articulation existent in perceivers’ L1; thus the present results extend these findings showing consistent patterns for an AV-fused (interdental) sound non-existent in the perceivers’ L1. Moreover, the natives but not the (short LOR) non-natives make more fused responses for $A_{labiodental} + V_{alveolar}$ stimuli than $A_{alveolar} + V_{labiodental}$ stimuli. This direction effect has also been reported previously, with fusion more easily occurring when the visual input is not visually salient (McGurk and McDonald, 1976; Sekiyama and Tokhura, 1991). However, that this effect in the current study decreases from the natives to the long and short LOR groups suggests that the non-natives are less sensitive to the difference of the L2 visual input. Overall, the results suggest that although L2 perceivers tend to use as much information as possible to compensate for the difficulty in the perception of non-native sounds, awareness of the visual speech domain does not necessarily lead to an accurate perception of L2 visual cues.

C. Linguistic experience

Extending previous findings showing a facilitative role of experience on L2 AV perception (e.g., Hardison, 2003; Sekiyama, 1997; Werker et al., 1992), the results reveal the effect of linguistic exposure indexed by the length of residence in an L2 country. Compared to the Mandarin short LOR group, the long LOR group can more correctly identify the non-native interdental and integrate the AV information, as well as being less susceptible to the AV-fusion. They approximate the native English patterns with a greater degree of accuracy in perceiving auditory than visual information, but with visual information becoming particularly facilitative in noise when the listening condition is poor. Together, these results reveal a pattern of learning in progress. As less experienced perceivers’ (the short LOR group) auditory perception of L2 sounds is poor, they focus on visual information as an additional channel of input, yet cannot use the visual information correctly or efficiently integrate the auditory and visual information. In contrast, experienced non-native perceivers (the long LOR group) make a nativelike use of AV information. These results not only reveal that AV perception of L2 sounds changes as a function of linguistic experience, but also suggest a dynamic learning pattern whereby auditory learning may precede and is accompanied by visual learning, resulting in an effective integration of AV speech information.

D. Noise background

A comparison of the native and non-native perceivers’ performance in quiet and noise reveals that a significant group difference exists only in the interdental identification. Visual information facilitates the English natives’ perception in noise, as they can efficiently weigh the audio and visual input channels based on need, relying more on visual input in a nonoptimal listening condition. The Mandarin (short LOR) natives, however, show poor audio perception in both quiet and noise, and the addition of visual information does not facilitate perception to the same degree as for the natives. These patterns indicate that noise differentially affects native and non-native AV perception for the L2 sounds, suggesting a language-specific AV processing.

E. General discussion

Evidence from the present study points to language-specific processing integrated with universal aspects of AV processing. The results of interdental perception suggest language-specific patterns whereby native and non-native speakers weigh the auditory and visual input differently. While native English perceivers are primarily dependent on the auditory component, non-native perceivers (the short LOR group in particular) tend to make greater use of the visual information, although their performance remained poor in the manner in which they used the L2 visual cues and in effectively integrating the L2 AV speech information. While previous research indicates that L2 perceivers do not use visual information to the same degree as natives (e.g., Hazan et al., 2006; Ortega-Llebaria et al., 2001; Sekiyama, 1997; Werker et al., 1992), this study further shows that focusing on the visual speech domain does not necessarily lead to an accurate perception. Furthermore, that the Mandarin long LOR group outperformed the short LOR group indicates that L2 visual speech cues may be learned and improved with L2 experience. Despite these language-specific
patterns, the similarity between the native and non-native visual perception with degraded auditory information (as in noise, and when non-native, respectively) also suggests universal aspects in the underlying integration of AV speech information (consistent with the FLMP; Chen and Massaro, 2004).

Particularly promising from the present results along with previous L2 AV research is the possibility of bridging the learning patterns across speech input modalities. Indeed, the acquisition of non-native visual cues is assumed to be analogous to that of auditory L2 speech learning (Hazan et al., 2005, 2006; Ortega-Llebaria et al., 2001; Sekiyama, 1997). The difficulty in the perception of L2 visual information may be attributed to the interference from an L1. In particular, L2 visual cues may be classified as “identical,” “similar,” or “new,” depending on whether they have gestural counterparts in the L1, as has been proposed in L2 auditory speech learning theories (e.g., SLM, Flege, 1995; PAM, Best, 1995). For L2 visual information which is non-assimilable 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. In the present visual perception results, the short LOR Mandarin perceivers predominantly confuse the interdentals with the place-adjacent alveolars familiar in their L1 (see Table III), indicating that L2 category formation may be blocked not only if an L2 phone is auditorily perceived as similar to the closest L1 category (Flege, 2007), but also if L2 and L1 phones are visually close. Recent theories also suggest that in addition to the visual similarity between L1 and L2, L2 speech learning models should include further factors to account for audio-visual L2 speech processing, such as the relative weighing of auditory and visual cues and the distinctiveness of visual cues (Hazan et al., 2006). The non-native short LOR group may have increased visual weighing in the perception of the new L2 interdentals but still have difficulty assimilating them to the appropriate L2 categories. On the other hand, as is the case with the acquisition of auditory speech, the long LOR group’s nativelike performance indicates the formation of new L2 visual categories with exposure, suggesting a similar role of linguistic experience on L2 speech perception across auditory and visual domains.

V. CONCLUDING REMARKS

The language-specific as well as universal aspects of AV processing revealed in this paper and in previous research suggest an integrated network in cognitive processing and learning that involves and goes beyond individual modalities and domains. Results also suggest that the L2 speech learning theories developed on the basis of auditory speech perception may be extended to include visual speech perception. These findings point to directions for future research. First, the similarities and dissimilarities of L1 and L2 visual/gestural categories deserve to be further explored and eventually quantified. Subsequent research should include more extensive types of L2 speech sounds that characterize various viseme relationships with corresponding L1 sounds (e.g., identical, similar, and new). Second, the present study reveals the effect of linguistic experience indexed by length of residence in the L2 environment. As previous research suggests that other factors (such as age of L2 acquisition, amount of L2 input, and L2 proficiency) may vary and covary to influence auditory speech perception (e.g., Flege, 1995, 1998; Flege et al., 1999; McAllister et al., 2002), these factors should also be taken into account in L2 visual speech learning. Finally, the present findings may have pedagogical implications: that is, training and/or teaching of L2 speech should not only emphasize the awareness of the visual domain (to increase visual weighting), but should also direct learners to focus on L2 specific visual speech cues (to establish correct L2 categories).

ACKNOWLEDGMENTS

We thank Nicole Carter, Angela Cooper, Chad Danyluck, Angela Feehan, Vivian Hsing, Nina Leung, Elaine Pang, Lindsay Shaw, and Kristy Stefanucci at Simon Fraser University (SFU) for their assistance in stimulus development, data collection, analysis, and preparation of the manuscript. We would also like to thank Dr. Allard Jongman and the three anonymous reviewers for their valuable suggestions. Portions of this research were presented at Interspeech 2006 in Pittsburgh and at the Fourth Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan in Honolulu in 2006. This project was supported by a standard research grant from the Social Sciences and Humanities Research Council of Canada (SSHRC 410-2006-1034) and a SFU Institutional SSHRC grant.

1Here and elsewhere Tukey HSD was carried out for post hoc analyses where there are three levels of a factor. The accepted significance level was $p < .05$.


