Ling istic e perience and a dio- is al perception of non-nati e fricati es

Yue Wang[a)]
D , , B , B C 5A 1 6, C

Dawn M. Behne
D , , 7491 ,

Haisheng Jiang
D , , B , B C 5A 1 6, C

(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: interdentals 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.

(e.g., Erber, 1969; l Pollack, 1954), especially when auditory distinc- creases, such as in a noisy environment (Bernstein 4; Erber, 1969; Sumby and Pollack, 1954; Sum- 979). The relative contribution of audio and visual n has also been revealed by what is known as the effect," where an audio [ba] dubbed onto a visual roduce a [da] percept (McGurk and MacDonald, se results suggest an ability to integrate auditory l speech information (Massaro, 1987, 1998; ., 2003). A question that arises is whether this ects an innate capacity to process multimodal in- or is developed by learning and experience such formation processing is instantiated by language- tterns.

lwired ability to process audio and visual informa- een shown for prelinguistic infants: 2–4 month speech information matching, compared to non- in auditory and visual information (Kuhl and

formation for tones in Chinese has also been observed (Burnham ., 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 aiversrheless whenno48 0

common to their L1 and L2 (e.g., Sekiyama ., 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 ., 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 ., 2001; de Gelder and Vroomen, 1992; Sekiyama ., 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 ., 1995; Flege ., 1999; McAllister ., 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;

zu]. 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 ., 2006; 2005; Jongman . 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 video 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 ([$A_{labiodental} - V_{alveolar}$], e.g., audio [fa] dubbed onto video [sa], $A_{fa} - V_{sa}$; or [$A_{alveolar} - V_{labiodental}$], e.g., audio [sa] dubbed onto video [fa], $A_{sa} - V_{fa}$). 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., [θa]). The fricative voicing and vowel were always the same for the A and V components of a given AVi syllable (e.g., $A_{fa} - V_{sa}$, $A_{fu} - V_{su}$, and $A_{va} - V_{za}$). A total of 12 AVi stimuli were used (two AV input place [$A_{labiodental} - V_{alveolar}, A_{alveolar} - V_{labiodental}$] □ two voicing conditions [voiceless, voiced] □ three vowels [i, a, u]).

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 [(4,44)=4.0, =.005]. 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 [(4,44)=6.8, .001] and AV [(4,44)=3.4, .001] conditions. In the A condition, whereas the English natives' identification of both the interdentals and labiodentals was moderately lower than for the alveolars, both the Mandarin long and short LOR groups' identification for the interdentals 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 interdentals and alveolars in the A and AVc conditions. For the interdentals, the A condition revealed a decreasing identification accuracy from the native English to Mandarin long LOR to short LOR groups [(2,47)=23.6, .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 [(2,47)=8.3, .001]. For the

TABLE III. Confusion matrix for the A, V, and AVc conditions. Mean percent responses (%

showed a marginal difference across modalities [(2,57) =3.1, =.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 (.05), the Mandarin short LOR group's auditory perception of interdentals did not differ between the noise and quiet conditions [(1,19)=1.5, =.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-match), the video component (V-match), or the fused A and V components (AV-fusion, i.e., interdental). Sets of three-way mixed ANOVAs were carried out for each of these response types, with group as a between-subject factor and AV-place input ($[A_{labiodental}+V_{alveolar}]$, $[A_{alveolar}+V_{labiodental}]$) 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 [(1,47)=31.3, .001] and a group AV-place interaction [(1,47)=6.3, =.004]. A moderate degree of

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 [$(2, 47) = 12.2$, $\square .001$], indicating that the Mandarin perceivers more easily fused the incongruent stimuli despite the

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 ., 1987; Soto-Faraco ., 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 ., 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 interdentals 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, 207b; Sekiyama and Tohkura, 1993; Sekiyama ., 2003), the fused sound often has had a place of articulation existent in perceivers' L1; thus the present results extend these findings showing consistent patterns for an AV-fused (interdental) sound nonexistent 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 Tohkura, 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 ., 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 interdentals 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 experieeT*(.)entlde353.7ggest

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 ., 2005, 2006; Ortega-Llebaria ., 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)