Native and Nonnative Audio-visual Perception of English Fricatives in Quiet and Café-noise Backgrounds

Yue Wang[1], Dawn Behne[2], Haisheng Jiang[1], and Chad Danyluck[1]

[1] Department of Linguistics, Simon Fraser University, British Columbia, Canada
[2]Department of Psychology, Norwegian University of Science and Technology, Trondheim, Norway

yuew@sfu.ca

Abstract

This study examines audio-visual perception of second-language (L2) speech, with the goal of investigating the extent to which the auditory and visual input modalities are integrated in processing unfamiliar L2 speech. Native (Canadian English) and nonnative (Mandarin) perceivers responses were collected for a set of fricative-initial syllables presented with a quiet and a cafe-noise background, and presented in four ways: congruent audio-visual (AVc), incongruent audio-visual (AVi), audio-only (A) and visual-only (V). Results show that for both native groups, performance was better in the AVc condition than A or V condition; and better in quiet than in cafe-noise background. A comparison of the native and nonnative performance revealed that Mandarin participants showed (1) poorer identification of the L2 interdental fricatives, (2) a greater degree of reliance on visual information, even when auditory information was available, and (3) a higher percentage of McGurk responses with the incongruent AV speech. These findings indicate that although nonnatives were able to use visual information, they failed to adopt the visual cues that are linguistically characteristic of the L2 sounds, suggesting a language-specific AV processing pattern. However, similarities between the two native groups are also indicative of possible perceptual universals involved. Together they point to an integrated network in speech processing across modalities.

Index terms: audio-visual, speech perception, nonnative.

1. Introduction

Language experience often involves face-to-face interaction with simultaneous perception of a speaker´s voice and facial movements. Previous research has shown that for native (L1) perceivers, speech perception is enhanced when visual information is available [1,2]. This enhancement is especially effective when auditory distinctiveness decreases, such as in a noisy environment [3].

Extension of this research to nonnative perceivers (L2) has shown that L1 experience affects visual perception of L2 speech. Although nonnatives are influenced by visual information when perceiving L2 sounds [4-6], they may also be impeded in correct use of L2 visual cues [7-11], presumably because they have lost sensitivity to visual cues non-existent in their L1 [10,11]. Sensitivity to visual information in L2 speech can nevertheless be enhanced through audio-visual (AV) training, and to some degree, with auditory-only (A) training [11,12]. Related research also shows that nonnative listeners are more affected by noise than native listeners [6]. These findings suggest that an L2 perceiver´s use of AV information may be less stable than that of a native perceiver.

This study explores the extent to which nonnative speakers make use of visual information in L2 speech perception in a quiet and a natural noisy environment.

While many studies focus on stops, nasals [e.g., 6,13], or liquids [10,12], the current study chose to test fricatives [1]. English and Mandarin perceivers are included for their difference in L1 phonetic inventories, where English contains sounds and corresponding visual cues that are non-existent in Mandarin (i.e. interdental fricatives, /θ, δ/, e.g., the initial sounds in “thin” or “this”).

The effects of environment were examined by comparing quiet and natural café-noise backgrounds. Given the acoustic nature of fricatives and their similarity to noise, L2 listeners may be especially affected by the background condition.

Previous research has tested nonnative perception of A,V, and AV congruent conditions [9,10], or mismatched A and V components [14]. The current study extends this research by including all the congruent and incongruent AV modalities (A, V, AVc, and AVi). To tease apart the influence of A and V information in a single stimulus, this study makes use of the classic illusion known as the McGurk Effect [15], by using responses to an incongruent AV stimulus to determine the contribution of each component modality.

Of particular interest was how the Mandarin group would respond in an incongruent AV condition where the A and V components occur in their native Mandarin and nonnative English, and whether the AV fusion would correspond to a percept available in English but not in their native Mandarin. In this case, to what extent does the native Mandarin group remain anchored to the A and/or V components, versus perceive the non-native AV-fused intermediate English percept? In addition, compared to the English natives, are the Mandarin group’s results affected differently by the presence of café-noise?

2. Method

2.1. Materials

Stimuli were developed based on possible English CV syllables having a fricative onset followed by a vowel [e.g., 1,15]. Fricatives differed in voicing (voiceless, voiced) and place of articulation, (labiodental, interdental, alveolar): /f, θ, s, v,δ, z/. Neither the interdental fricatives nor the voiced fricatives occur in Mandarin. The vowel was /a/, /i/ or /u/, all occur in Mandarin.

Audio and video recordings were made of an adult male speaker of Canadian English producing six repetitions of the 18 syllables (6 fricatives x 3 vowels) at a normal speaking rate. Recordings were done in the Phonetics Lab at Simon Fraser University (SFU). For each syllable, one best repetition was selected such that all syllables with the same fricative match in duration. These selected audio stimuli were then normalized for RMS intensity (70dB). Visual stimuli were edited such that for each clip, there are 1s neutral face before and 1s after the stimulus. The frame size was 640x480. Cafeteria noise recorded at SFU was added to a copy of all the stimuli, resulting in a quiet and café-noise (S/N=0dB) version for all the syllables.

The audio and video components of the syllables were used to create stimuli which were audio-video congruent (AVc), audio-video incongruent (AVi), audio-only (A) and video-only (V). Sample AVi stimuli are summarized in Table 1. In the AVi stimuli, the fricative place of articulation was either labiodental or alveolar. The fricative voicing and vowel were always the same for the A and V components of a given AVi syllable. This was the same for stimuli with clear and café-noise backgrounds. This gave 66 stimuli (12 AVi, 18 AVc, 18 A, 18 V) in the clear and in the café-noise background, for a total of 132 stimuli.

Table 1. Sample audio(A) and video (V) components for incongruent AV syllables, and the expected intermediate response (McGurk effect)

Input	A - V	Expected percept
CV	fi – si si - fi	θi

2.2. Participants

Two groups of young adults (mean age = 23) participated in the study: 15 native Canadian English and 20 native Mandarin. Both groups were balanced for sex. The Mandarin participants were nonnative intermediate-level English users, who had studied English since age 12, and had been in Canada for an average of two years. The participants reported having no known hearing loss and normal or corrected vision. All were living in Vancouver, Canada at the time of the study.

2.3. Procedure

An identification task was carried out where each participant was tested on the full set of stimuli in the clear and café-noise backgrounds. Stimuli were blocked by background (quiet, café-noise) and modality (A, V, AV), with AVi and AVc stimuli in the same block. Each block included two repetitions of each stimulus. Stimuli were randomized within a block. The order of background and modality presentation was counter-balanced across participants.

For each trial a stimulus was presented auditorily over headphones, visually on a computer monitor, or both. Response alternatives were presented on the monitor, where, for a given trial, the alternatives were the 6 syllables which matched in voicing with the stimulus, as well as the option to give an alternative response. The participants´ task was to identify the syllable and respond by pressing a key on the computer keyboard for the corresponding syllable displayed on the monitor. Between trials, a fixation point was displayed in the center of the monitor for 1s. Participants had up to 4s to respond.

3. Results

3.1. Input modality

Mean percent correct responses for the A, V and congruent AV conditions are presented in Figure 1. A 4-way mixed analysis of variance (ANOVA) was carried out with L1 (English, Mandarin) as a between-subjects factor, and background (quiet, café-noise), modality (A,V,AV), place of articulation (labial-dental, interdental, alveolar) and voicing (voiceless, voiced) as repeated measures. The dependent variable was perceivers' correct identification for place of articulation regardless of voicing. This calculation was used specifically to accommodate the V-only condition, since facial information does not typically contain cues to voicing.

A significant main effect of L1 was found, with the native (English) perceivers' overall performance being better than that of the non-natives (Mandarin) [$F(1,34)=36.1$, $p<.0001$].

A significant main effect of modality (A,V,AV) was also observed [$F(2,33)=110.8$, $p<.0001$], together with significant interactions of modality x L1 [$F(2,33)=6.3$, $p<.003$], modality x place [$F(2,33)=2.9$, $p<.024$], L1 x place [$F(2,33)=3.4$, $p<.037$], and modality x L1 x place [$F(4,31)=2.9$, $p<.024$]. Post hoc analyses reveal that, compared to the other places of articulation, for the interdental fricatives which are not existent in Mandarin, Mandarin participants did significantly poorer than English participants across all modalities (59% vs. 84% respectively). Moreover, results for the interdental fricatives also show that, whereas English perceivers had lower scores in the V condition (71%) than in A (89%) and AV (93%) conditions, Mandarin perceivers show a progressive increase in the A (50%), V (60%) and AV (67%) conditions, indicating their reliance on visual information for these unfamiliar sounds. Similarly, for labial-dentals which are only voiceless in Mandarin, mean percent correct was not different in the three modalities for the English perceivers (A: 89%, V: 91%, AV: 93%), whereas for the Mandarin participants, AV input was better perceived (90%) than A (84%) or V (82%), consistently indicating the role of visual cues for unfamiliar sounds. Finally, for the alveolars which are familiar to the Mandarin perceivers, the nonnative perceivers' performance was not different from that of the native English participants, both having much poorer scores in the V than in A or AV condition, probably due to the alveolars being least visually accessible among the three types of fricatives. That no difference was observed in the alveolar pattern of responses for the A and AV conditions implies that, in an congruent AV condition, the available A-information overrides the available visual information.

A reliable main effect of background [$F(1,34)=247.5$, $p<.0001$] was also observed, along with an interaction of background x modality x place x L1 [$F(4,31)=10.6$, $p<.0001$]. Post hoc analyses show that, for the English participants, café-noise leads to a lower mean percent correct in the A and AV compared to the quiet condition. Furthermore, whereas the native perception in the V condition was unaffected by noise, the Mandarin participants' perception of the interdentals showed a lower mean score in noise (47%) than in quiet (60%) condition. The interaction also reveals that the English group perceived the interdentals very poorly in the A noise condition (53%) compared to the quiet condition (89%), possibly because,

acoustically, interdentals have a weakly diffuse noise signal which may be easily masked by the background noise. Interestingly, for the same fricatives, the quiet and noise conditions differ very little in the AV condition (93% vs. 87%). In contrast, for the Mandarin group, the presence of noise did not make a great difference in A and AV conditions.

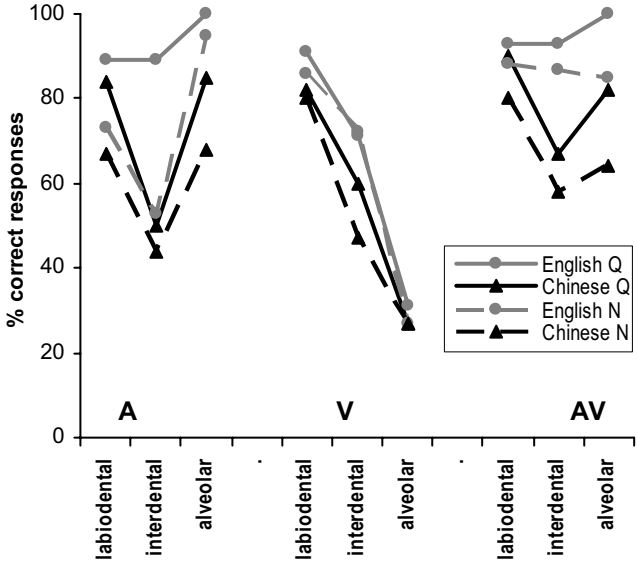

Figure 1. Mean % correct responses with A, V, and AV input in quiet (Q: solid lines) and noise (N: broken lines) backgrounds by English (gray lines) and Mandarin (black lines) participants.

3.2. McGurk effect

Mean percent responses for the incongruent AV condition are presented in Figure 2.

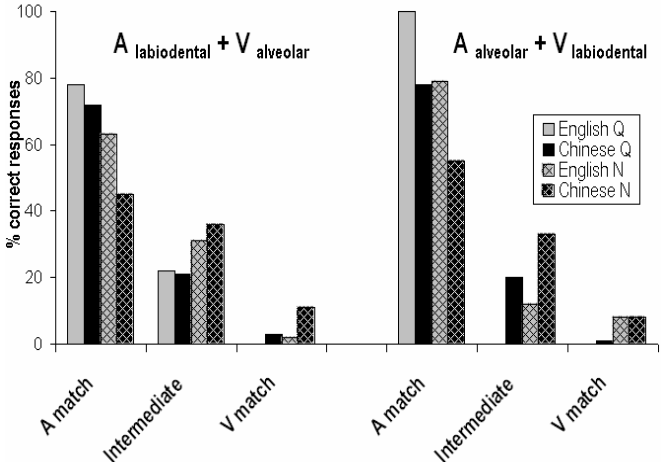

Figure 2. Mean % correct responses for incongruent AV stimuli in quiet (Q: plain bars) and café-noise (N: checked bars) backgrounds by English (gray bars) and Mandarin (black bars) participants. "A match" and "V match" indicate correct responses matching A and V component, respectively. "Intermediate" represents interdental responses matching neither the A nor V component, corresponding to the McGurk effect.

A 4-way mixed ANOVA was carried out based on each of three dependent variables: percent responses matching the A component, matching the V component, and intermediate to A and V components (expected McGurk effect). Here again the measures included responses for place of articulation regardless of voicing. For each ANOVA the between subject factor was L1 (English, Mandarin), and the repeated measures were background (quiet, café-noise), AV place ($A_{labiodental} + V_{aveolar}$, $A_{alveolar} + V_{labiodental}$), and voicing (voiceless, voiced).

Results show that for the incongruent AV stimuli, the Mandarin group had a higher mean percent response (21%) for the intermediate (interdental) fricatives than the English group (11%) [$F(1,34)=9.1$, $p<.0001$], indicating that they more easily fused the incongruent A and V components of the stimuli (McGurk effect) despite the intermediate fricatives being nonnative. Furthermore, a reliable difference was observed for the main effect of AV-place [$F(1,34)=12.3$, $p<.001$], and L1 x AV-place interaction [$F(1,34)=8.2$, $p<.007$]. The English group only demonstrated the McGurk effect when the audio component was labial and video component was alveolar (22%) but not the reverse (0%). In contrast, the Mandarin group showed similar effects for both (21% and 20% respectively). A main effect of L1 was also observed for percent responses matching the A component, with the English group (89%) having a higher mean score than the Mandarin group (75%), indicating that the native group was more accurate in following the auditory component [$F(1,34)=22.9$, $p<.001$].

For the responses matching the V component, no reliable difference was observed between the native and nonnative groups, nor were there any significant interactions.

Comparing the quiet and café-noise conditions, significant differences were observed for both native and nonnative groups. While the mean percent responses matching A was lower in noise than the quiet condition [$F(1,34)=82.4$, $p<.0001$], the likelihood to give a V or fused response increased in café-noise (V match: [$F(1,34)=22.4$, $p<.0001$], intermediate [$F(1,34)=43.3$, $p<.0001$]), showing that visual information is used more in café-noise than in quiet. Finally, voicing had no effect on native or nonnatives responses, in quiet or noise.

4. Discussion and Conclusions

Results show that the native and nonnative groups are both actively using the A component, reflected by the overall better performance in A or AV conditions as compared to the V condition, as well as by the incongruent AV results, where identification of the auditory component in a stimulus overwhelmingly surpassed the visual component. With the overall reliance on auditory information, it is not surprising that both groups are affected by the presence of café-noise. Further, the nonnative group as well as the native group used visual cues when available, but revealed poor performance for the visually presented sounds that are not visually distinct (e.g. alveolars). These results together with previous findings indicate that native and nonnative perception of AV information follows some universal patterns. Indeed, previous research suggests that some aspects of AV speech perception may be neutral across languages, just as primary auditory and visual processing [13].

On the other hand, extensive evidence also points to a language-specific AV processing. Cross-linguistic studies have

shown that native and nonnative speakers weigh the A and V input differently, depending on whether the visual cues are linguistically distinctive in their L1 [8,9]. The current results consistently show that native and nonnative perceivers of AV speech use the A and V components to different degrees. Native perceivers are primarily dependent on the A component, even when visual information is available (i.e. AV congruent stimuli). Nonnative perceivers make use of the available A and V information, and are able to use V information even in the absence of the A component.

Of particular interest are the results of the interdentals non-existent in Mandarin perceivers' L1. Their overall poorer performance in perceiving the interdentals across all input modalities compared to the natives indicates that they have not grasped these nonnative sounds. However, they did reveal a progressive increase in identification from the A, to V and AV conditions in both quiet and noisy backgrounds, relative to the native perceivers' superior A to V responses, suggesting that nonnatives might even surpass the natives in adopting visual information in speech perception. Given the previous finding that perceivers rely more on visual speech information when intelligibility is poor [3], Mandarin perceivers would conceivably resort to the visual information as an additional channel of input in perceiving the difficult nonnative sounds.

This leads to the crucial question as to whether nonnative speakers were simply attending to the visual information, or if they could use these visual cues in a linguistically meaningful manner. Comparing the native and nonnative perceivers' performance in quiet and noise conditions reveals that whereas native speakers can appropriately adopt visual linguistic cues when necessary, nonnatives cannot, even though they use the visual input to a greater degree. The current results show that native English perceivers typically use auditory cues (as shown by the high score in the A-quiet condition, and low in the V-quiet condition), whereas when the signal is masked (shown by the poor score in the A-noise condition), they were able to effectively adopt the visual cues (as shown by the high score in the AV-noise condition). In contrast, although the nonnatives attend more to the V than A input, their performance remained poor in all three (A,V,AV) conditions in both quiet and noise.

Consistent evidence was provided by the incongruent AV data. Similar to the previous results [e.g.,14], Mandarin perceivers had more occurrences of the McGurk effect than the English perceivers, indicating that nonnatives are more vulnerable to this illusion, perhaps due to their unfamiliarity with both the auditory and visual cues to the nonnative sounds.

These results convergingly show that nonnatives may intentionally attend to visual information, but are not able make effective use of the correct visual cues which contain linguistically contrastive information. That nonnatives behave differently than natives may lend support to the language-specific AV processing view, suggesting that visual speech cues may be learned just as the acquisition of auditory speech.

Indeed, the acquisition of nonnative visual cues (termed "visemes") has been assumed to be analogous to that of auditory L2 sound learning [8-11]. In particular, L2 visemes may be classified as "identical", "similar", or "new", depending on whether they have gestural counterparts in the L1, just as proposed in L2 speech learning theories (e.g., Speech Learning Model, [16]; Perceptual Assimilation Model, [17]). Particularly promising is the possibility of bridging the learning patterns across speech input modalities. Moreover, the language-specific as well as universal aspects of AV processing revealed in this and previous research, may also suggest an integrated network in cognitive processing and learning that involves and goes beyond individual modalities and domains.

5. Acknowledgements

We thank Angela Feehan, Elaine Pang, and Kristy Stefanucci at Simon Fraser University (SFU) for their assistance. This project was supported by a research grant from the Social Sciences and Humanities Research Council of Canada (SSHRC) and an SFU Institutional SSHRC grant.

6. References

[1] Jongman, A., Wang, Y., & Kim, B. (2003). Contributions of semantic and facial information to perception of nonsibilant fricatives. *JSLHR* 46, 1367-1377.

[2] Sumby, W. & I. Pollack (1954) Visual contribution to speech intelligibility in noise. *JASA* 26, 212-215.

[3] Erber N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *JSHR* 12, 423-425.

[4] Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading*, pp. 97–113. London: Erlbaum.

[5] Werker, J.F, Frost, P.E, & McGurk, H (1992). La Langue et les levres: Cross-Language Influences on Bimodal Speech Perception. *Canadian J. Psych.* 46, 4, 551-568

[6] Sekiyama, K., Kanno, I., Miura, S., and Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET, *Neuroscience Research*, 47, 277-287.

[7] de Gelder, B. & Vroomen, J. (1992) Auditory and visual speech perception in alphabetic and non-alphabetic Chinese/Dutch bilinguals. In R.J.Harris (Ed.), *Cognitive Processing in Bilinguals*. pp. 413-426. Amsterdam: Elsevier.

[8] Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Percept. & Psych.* 59, 73–80.

[9] Oretga-Llebaria, M.,Faulkner, A., & Hazan, V. (2001) Auditory-visual L2 speech perception: Effects of visual cues and acoustic-phonetic context for Spanish learners of English. *AVSP,* 149-154.

[10] Hazan, V., Sennema, A., Faulkner, A. & Ortega-Llebaria, M. (2006). The use of visual cues in the perception of non-native consonant contrasts. *JASA* 119, 1740-1751.

[11] Hazan, A., A. Sennema, Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication* 47, 360-378.

[12] Hardison, D.M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Appl. Psycholing.* 24, 495-522.

[13] Chen, T. & Massaro, D.W. (2004). Mandarin speech perception by ear and eye follows a universal principle. *Percept. & Psychophys*. 66, 820-836.

[14] Massaro, D.W., Tsuzaki, M., Cohen, M.M., Gesi, A., & Heredia, R. (1993). Bimodal speech perception: An examination across language. *Journal of Phonetics*, 21, 445-478.

[15] McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748.

[16] Flege, J. E. (1995). Second language speech learning: theory, findings, and problems. In W. Strange (ed.) *Speech Perception and Linguistic Experience.* Baltimore: York, pp. 233–273.

[17] Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (ed.) *Speech Perception and Linguistic Experience: Issues in Cross-Language Research.* Baltimore: York, pp. 171–204.