Effects of visual speech information on native listener judgments of L2 consonant intelligibility

Saya Kawase and Yue Wang
Department of Linguistics, Simon Fraser University, Burnaby, B.C., Canada
skawase@sfu.ca

Abstract
The present study examines how visual information in nonnative phonemes affects perceptual accuracy of second language (L2) speech production. Native Canadian English listeners perceived three English phonemic contrasts produced by native speakers of Japanese as well as native speakers of Canadian English as controls, under audiovisual (AV), audio-only (AO), and visual-only (VO) conditions. The phonemes include /θ, θ, l, s/ which are not existent in Japanese (L2 phonemes) as well as /b, s/ that are shared in both Japanese and English consonant inventories. The results showed that the English listeners perceived the Japanese productions of the phonemes /b, v, s, θ/ as significantly more intelligible when presented with the AV condition compared to the AO condition, indicating facilitative effects of visual speech information on their perceptual accuracy of nonnative production. However, the Japanese production of /θs/ was perceived as less intelligible in the AV condition compared to the AO condition, indicating that nonnative speakers’ incorrect articulatory configurations may decrease the degree of intelligibility. These results suggest that listener judgments of L2 productions may be either positively or negatively affected by additional visual speech information.

Index Terms: audiovisual speech, perception of L2 consonants, Japanese learners of English.

1. Introduction
It is widely known that second language learners face difficulties in producing and perceiving L2 sounds due to the interaction of their first language (L1) and second language (L2) inventories. The speech learning model (SLM) [3] and the perceptual assimilation models (PAM) [1, 2] posit an interaction of L1 and L2 speech inventories. With this interaction, new L2 sounds may be assimilated into similar counterparts in their L1. While the SLM posits that L2 learners perceive acoustic features from their input in order to form new phonetic categories, the predications of PAM and PAM-L2 are based on articulatory phonology in which L2 learners are exposed to new articulatory gestures of L2 sounds. L2 learners may thus acquire L2 gestural configurations, yet it is still unknown how both auditory and visual cues are acquired through L2 speech input, and how they influence productions.

Another question that remains is how to measure correct L2 productions. Previous research has examined the accuracy of L2 speech productions, for instance, using acoustic analysis to determine the accuracy of English /θ/ produced by Japanese learners of English [6], and having native listeners evaluate L2 speech intelligibility in perceptual judgment tasks [10]. Whereas these previous studies have mainly focused on the acoustic auditory domain, it is important to consider the role of gestural configurations in speech perception. In face-to-face conversation, not only acoustic information, but also visual speech information (e.g., mouth configurations and movements) is used. As the McGurk effect [9] shows, visual and acoustic cues in speech are integrated in sound identification (i.e., the combination of visual /g/ and auditory /b/ cues results in the perception of /d/). The visual information can facilitate perception of consonants and vowels, which results from cues to place and manner of articulation [14]. In a degraded environment, such as in noise, audiovisual (AV) benefits were also observed for the identification of spondaic words [11]. Visual speech information is effectively used not only in acoustically degraded conditions, but also when speech topics are complicated or when speech is produced by a talker with a strong accent [14]. In the AV perception of nonnative speech, a non-native speaker effect has been reported where listeners increase the visual weighting when listening to non-native productions [7, 8, 13]. Thus, speech perception is a dynamic process where listener judgments may vary on the basis of available modality.

While studies of audiovisual speech perception in the area of L2 speech have been conducted mainly for the perception of nonnative speech [8, 13], it is also important to consider the effects of visual cues to determine the intelligibility of L2 production. Given that visual benefits have been reported to increase perceptual accuracy, having additional visual cues might influence listener judgments of L2 speech intelligibility. In addition, the results of previous studies in which L2 production has been judged by only auditory stimuli may not fully reflect perceptions of L2 speech in face-to-face communication.

2. Present study
The present study examines the effect of visual information on the intelligibility of L2 consonant productions. Canadian English listeners were asked to identify the initial consonants /b, v, s, θ, l, s/ in English CV syllables produced by Japanese learners of English. Since Japanese does not have /θ/ and /s/, but has /b/ and /s/, Japanese learners of English tend to replace the nonexistent phonemes with the similar counterparts existing in their L1 (i.e., /θ/ with /b/; /s/ with /s/ [15]). Furthermore, Japanese speakers tend to assimilate the /l/ and /s/, which do not exist in Japanese, to the Japanese lateral flap /ʃ/, with /l/ being more likely to be categorized as Japanese /s/ [5]. The perception tasks were conducted under the following three conditions: audio-only (AO), visual-only (VO) and audiovisual (AV), to examine how visual information affects the perception of L2 speech. On the basis of previous studies showing an increase in visual cue weighting on the perception of nonnative speech [7, 8, 13], we hypothesized that the listeners may perceive the Japanese production as more intelligible in the AV condition compared to the AO condition. Furthermore, the visual facilitative effects may be less for the
consonants that exist in both Japanese and English phonemic inventories (i.e., /b, v, s/) since auditory input would be intelligible enough for these to be correctly perceived, whereas the non-Japanese consonants /θ, ɹ, 0, l, x/ may require a greater degree of visual weighting.

3. Methods

3.1. Speakers

Fifteen native speakers of Japanese participated in this study. All data were elicited from relatively newly arrived students or workers in Vancouver (mean length of residence=10.2 months). In addition, none of them used English at home or had lived in an English speaking country before arriving in Canada. In general, the Japanese students began to learn English at the age of 13 in a school setting, and they are therefore categorized as late L2 learners. An additional 15 native speakers of Canadian English also participated as a control group. Since this study uses visual information available from the speakers’ full face for the identification task, their perceived ethnicity information may crucially affect the listeners’ judgments. In order to remove the potential influence, only native speakers of English who were of East-Asian descent (e.g., Japan, China, Korea) were recruited.

3.2. Stimuli

Six non-word CV syllables, having the initial consonant /b, v, s, 0, l, x/ followed by a vowel /a/ were used as stimuli. These syllables were presented under the three input types, audiovisual (AV), audio-only (AO) and visual-only (VO). Each condition included a total of 180 stimuli (i.e., 6 consonants x 1 vowel x 15 speakers in each language group x 2 languages).

Table 1: The set of stimuli

<table>
<thead>
<tr>
<th>Stimulation</th>
<th>b/a</th>
<th>v/a</th>
<th>s/a</th>
<th>b/θ</th>
<th>v/θ</th>
<th>s/θ</th>
<th>b/ɹ</th>
<th>v/ɹ</th>
<th>s/ɹ</th>
</tr>
</thead>
<tbody>
<tr>
<td>POA</td>
<td>Bilabial</td>
<td>Labiodental</td>
<td>Alveolar</td>
<td>Bilabial</td>
<td>Labiodental</td>
<td>Alveolar</td>
<td>Bilabial</td>
<td>Labiodental</td>
<td>Alveolar</td>
</tr>
<tr>
<td>Occurrence</td>
<td>ENG</td>
<td>ENG</td>
<td>ENG</td>
<td>ENG</td>
<td>ENG</td>
<td>ENG</td>
<td>ENG</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. POA= place of articulation, ENG= English, JP= Japanese

3.3. Speaking task

Audio and video recordings were made of 15 native speakers of Japanese and English individually. Prior to the recording, they were given instructions orally and with a written instruction sheet. They were encouraged to read aloud the six CV stimuli in citation form, (e.g., ba) at their normal speed and volume. PowerPoint was used to present the randomized stimuli. Overall, the speakers were asked to read the stimulus set with five repetitions, and the best examples were chosen considering noise and other issues (e.g., irregular pause). The recording lasted approximately 30 minutes. The recording of the speaker’s full face was made with a digital camcorder (Canon Vixia HF S30 HD Video camcorder). Because the camcorder has a low quality of sound resolution, a separate audio recording was made with a Shure KSM 109 condenser microphone to SoundForge 6.4 at a 44.1 kHz sampling rate. In order to maintain the consistency in the recording, and to avoid any distractions for the perception tasks, all the speakers were seated in front of monotonous blue wallpaper and their faces were centered in the frame. The recording was carried out in the recording studio in the Language and Brain Lab at Simon Fraser University. The speakers were paid for their participation.

3.4. Stimulus preparation

After collecting both audio and video files, three sets of perception tasks (AV, AO and VO) were prepared. First, the intensity of the audio files was normalized by SoundForge 6.4 to have the same unweighted RMS value. The audio files collected by the camcorder were deleted and the normalized audio files were synchronized with the video-recording files using Final Cut Pro X. Each finished video clip lasted 3000ms. The video-recorded files with the synchronized high-quality audio files were used in the AV condition, whereas the AO condition and the VO condition were made with the extracted audio and video files respectively.

3.5. Listeners

Thirty-one native speakers of Canadian English (sixteen male, fifteen female) participated as listeners to identify the stimuli produced by the speakers. All were students at Simon Fraser University. They were aged between 18 and 42 (M age=23.4). A language background questionnaire revealed that none of the listeners had experience learning Japanese and all had limited or no daily exposure to Japanese-accented English. All the listeners reported that they had normal hearing and normal or corrected vision. Due to file-loading issues in E-prime, data from two male listeners were excluded from analysis.

3.6. Listening task

The listeners were tested individually in a sound-treated room while wearing headphones to listen to the stimuli. While the stimuli were CV syllables, the listeners were asked to focus on the initial consonants and ignore the quality of following vowel /a/. Prior to each session, they received instructions on the screen. They saw the articulatory movements in the center of the screen and heard the stimulus over the headphones in the AV condition, only heard the stimulus in the AO condition, or only watched the mouth movement in the VO condition. E-prime 2.0 was used to run the perception tasks. The order of the modality condition (AV, AO, VO) was randomized. In the identification task, a fixation point was displayed for 1000ms on the display for each trial prior to the target stimuli. Response alternatives were presented on the screen with possible consonants (e.g., /b/ and /v/ for “va”) as well as an option to type what they heard using a keyboard in case the listeners perceived something other than the presented response alternatives. The listeners were allowed a maximum response time of 4 seconds in each trial. Overall, the entire identification task lasted 90 minutes, which was divided across the two visits with 5- to 10-minute breaks within each session.

4. Results

The listeners’ correct identification rates were analyzed using a repeated measure ANOVA with speaker language group (Japanese and English), consonant type (i.e., /b, v, s, 0, l, x/), and modality type (AV, AO, VO) as the within-group factors. Table 2 presents mean correct identification rates for the six stimulus consonants for Japanese and English speakers in each modality. Significant main effects were observed for language
group \[ F(1, 28)=1562.485, p < .001, \text{partial } \eta^2=.982 \], consonant type \[ F(5, 140)=89.764, p < .001, \text{partial } \eta^2=.762 \] and modality type \[ F(2, 56)=55.760, p < .001, \text{partial } \eta^2=.666 \]. Moreover, interactions of language group x consonant \[ F(5, 140)=228.421, p < .001, \text{partial } \eta^2=.891 \], language group x modality type \[ F(2, 56)=27.193, p < .001, \text{consonant x modality type} \[ F(10, 280)=18.015, p < .001, \text{and language group x consonant x modality type} \[ F(10, 280)=13.149, p < .001 \] were found.

Based on these main effects and interactions, further analyses were conducted to explore group differences for language group, consonant, and modality. First, sets of two-way (language group and consonant) repeated measure ANOVAs were conducted for each modality (AO, VO, AV) to see how Japanese produced consonants were perceived differently compared to native English productions in each modality. For brevity, only AO and VO analyses are reported here.

In the AO condition significant main effects were found for speaker language \[ F(1, 28)=1582.521, p < .001, \text{partial } \eta^2=.983 \], and consonant type \[ F(5, 140)=74.316, p < .001, \text{partial } \eta^2=.726 \]. The interaction between speaker language and consonant type was statistically significant \[ F(5, 140)=86.094, p < .001, \text{partial } \eta^2=.755 \]. Bonferroni post hoc tests revealed that while differences in the mean correct identification rates of the six consonants produced by the Japanese and English speakers were all significant \( p < .001 \), the differences varied based on the consonant type. In the perception of the Japanese productions, /b, s, l/ were significantly more intelligible (mean correct identification: /b/ = 91.2%, /s/ = 93.1%, /l/ = 88.4%) compared to /v, θ, ɹ/ \( p < .001 \). Among the low-intelligibility consonants, /θ (39.3%)/ was significantly less intelligible than /θ/ (56.8%), and /θ (55.2%)/. The listeners also showed different perception in the six consonants produced by the Canadian English speakers. The perception of /θ/ (88.4%) was significantly lower than other consonants \( p < .05 \). /v (92.2%)/ was significantly lower than /v, θ, ɹ/ (99.1%), /θ/ (98.6%), and /θ/ (99.5%) \( p < .01 \). There was also a significant difference between /θ/ and /θ/ \( p = .043 \). There were no other significant differences.

In the VO condition significant main effects of speaker language \[ F(1, 28)=296.690, p < .001, \text{partial } \eta^2=.984 \] and consonant type \[ F(5, 140)=20.175, p < .001 \] were observed in the VO condition, along with a significant interaction \[ F(5, 140)=124.172, p < .001, \text{partial } \eta^2=.791 \]. Bonferroni post hoc tests revealed that visual perception differed as a function of speaker productions. While /b, s/ did not show significant differences, the identification of /θ, θ, ɹ/ produced by the native English speakers were significantly better than the ones by the Japanese speakers \( p < .01 \). Furthermore, the identification of /l/ produced by the Japanese productions (77.5%) was significantly more accurate than the one produced by the English speakers (59.3%). Within the Japanese productions, /θ/ was significantly more intelligible than /s, ɹ/ \( p < .01 \), whereas the difference between the latter two consonants was not statistically significant \( p > .05 \). However, the Japanese-produced /l, s/ were more intelligible than their /θ, v/ \( p < .01 \). While /θ, v/ did not show any significant differences \( p > .05 \), these two were more intelligible than /θ/ \( p < .01 \). Among the English productions, /b, v, ɹ, θ/ were more intelligible than /s, l/ \( p < .05 \). Among the four more intelligible consonants, /θ/ was better perceived than /θ/ \( p < .01 \). With regard to the perception of native English productions, the mean correct identification rate of /θ/ was the most accurate (95.6%) even though the performance in the AO condition was the least accurate (88.4%). /θ/ was correctly perceived significantly more frequently than /θ/ (84.1%), s (76.0%) and /l/ (59.3%) \( p < .01 \). /θ/ was correctly perceived more frequently than /s/ \( p < .05 \) and /l/ was significantly less correctly perceived than /θ/. No other difference was observed.

Furthermore, to investigate how modality type affects listener perception of L2 speech, the mean correct identification rates for the six consonants in the three input modalities were submitted to a repeated measure ANOVA with three modality type as input (AV, AO, VO), and six consonant (i.e., /b, v, θ, ɹ, l/ as the within-group factors including only the data on Japanese-produced consonants. The mean scores are displayed in Figure 1. Main effects of modality type \[ F(2, 56)=35.548, p < .001 \] and consonant type \[ F(5, 140)=179.552, p < .001, \text{partial } \eta^2=.865 \] were observed. The interaction of the two factors was also significant \[ F(10, 280)=10.669, p < .001 \]. Bonferroni post hoc tests revealed how visual information facilitated or worsened the perception of various English consonants produced by the Japanese speakers. Along with the higher correct identification rates in the VO conditions compared to the AO condition, higher correct identification rates in the AV condition were observed among /b/ (AV: 97.7%, AO: 91.2%), /v/ (AV: 52.4%, AO: 39.3%), and /θ/ (AV: 63.9%, AO: 56.8%), compared to the AO condition \( p < .01 \). On the other hand, /s/ and /l/ showed a statistically higher correct identification rates in the AO than in the VO (/s/ AO: 93.1%, VO: 74.1%, /l/ AO: 88.4%, VO: 77.5%) \( p < .001 \). While the higher correct identification rates of /s/ was observed in AO condition, having additional visual information in the AV led further increase in the intelligibility (AV: 98.2%, AO: 93.1%) \( p < .01 \).

While positive effects of visual information were observed, the result for /θ/ was not consistent with other consonants. The intelligibility rate in the AV condition (51.6%) was significantly lower than in the AO condition (55.2%) \( p < .05 \). There was also a significant drop between the intelligibility of AV (51.6%) and VO (34.9%) \( p < .001 \). Similarly, significantly lower accuracy in the VO condition was also observed for /θ/ (AV: 98.2%, AO: 93.1%, VO: 74.1%), but relatively intelligible auditory information might be able to compensate for the low intelligibility of the visual information.
5. Discussion and Conclusions

The aim of this study was to investigate how AV speech input modality affects native listeners’ perception of L2 consonants. Native Canadian English listeners assessed English consonants produced by native Japanese speakers in three input modality types (AV, AO, VO). Consistent with the previous findings in the AO condition, the intelligibility of L2 productions varied as a function of L1, with the shared L1-L2 consonants being more intelligible than L2-only consonants. Moreover, the results showed similar L1 influences in the perception of the visual speech. Overall, the VO condition showed lower intelligibility rates in the perception of nonnative phonemes /θ, ʃ/ produced by the Japanese speakers. If the apprehension of articulatory gestures in a native phonological space underlies speech perception and these are inferred from the acoustic signals (c.f., PAM), the listeners might perceive speech as less intelligible in both AO and VO conditions.

Comparing the results across modalities, the findings suggest that additional visual cues were facilitative to the perception of L2 consonants, consistent with the previously found non-native speaker effect with native listeners demonstrating greater visual weighting in the perception of nonnative speech [7, 8, 13]. The comparison among the stimulus consonants showed visual facilitative effects in the perception of both native (i.e., /θ, s/) and nonnative (i.e., /ʃ, ʃ/) phonemes as produced by the Japanese speakers. A new finding compared to previous studies is influence of visual cues in the perception of /ʃ/ in AV perception. While other nonnative phonemes showed a significant increase in the perceptual accuracy with additional visual cues, the listeners perceived /ʃ/ as less intelligible in the AV condition compared to AO condition. This may have resulted from an increase of visual weighting by the native listeners on nonnative phonemes. With this increased visual weighting, if the articulatory configurations of /ʃ/ by the nonnative speakers were incorrect, the perceptual accuracy could be compromised in the AV and VO conditions as compared to the AO condition. Thus, the visual cue weighting could affect L2 speech intelligibility in both facilitative and detrimental ways.

These findings may offer valuable insights into listener-based assessments of L2 production. While visual speech information is generally facilitative in the native perception of nonnative speech, the effect of visual information may also be inhibitory. The enhanced visual weighting on L2 productions could negatively affect native listener judgments when the articulatory configurations of the L2 speech are incorrect. This research thus points to further studies on the nature of visual modality influences on the intelligibility of nonnative speech, as well as the effects of the interactions of auditory and visual input modalities.

6. Acknowledgements

The authors would like to thank Beverly Wu, Mathieu Doven, Nisha Banga and Keren Hernandez at Simon Fraser University (SFU) for their assistance in stimulus development and data collection. We would also like to extend our gratitude to all the participants. This research has been funded by a research grant from SFU Vice President Research.

7. References