The influence of visual speech information on the intelligibility of English consonants produced by non-native speakers

Saya Kawasea)
MARCS Institute, University of Western Sydney, Locked Bag 1797, Penrith, New South Wales, 2751, Australia

Beverly Hannah and Yue Wang
Department of Linguistics, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia, V5A 1S6, Canada

(Received 14 October 2013; revised 11 July 2014; accepted 17 July 2014)

This study examines how visual speech information affects native judgments of the intelligibility of speech sounds produced by non-native (L2) speakers. Native Canadian English perceivers as judges perceived three English phonemic contrasts (/b-v, θ-s, l-/l) produced by native Japanese speakers as well as native Canadian English speakers as controls. These stimuli were presented under audio-visual (AV, with speaker voice and face), audio-only (AO), and visual-only (VO) conditions. The results showed that, across conditions, the overall intelligibility of Japanese productions of the native (Japanese)-like phonemes (/b, s, l/) was significantly higher than the non-Japanese phonemes (/v, θ, j/). In terms of visual effects, the more visually salient non-Japanese phonemes /v, θ/ were perceived as significantly more intelligible when presented in the AV compared to the AO condition, indicating enhanced intelligibility when visual speech information is available. However, the non-Japanese phoneme /j/ was perceived as less intelligible in the AV compared to the AO condition. Further analysis revealed that, unlike the native English productions, the Japanese speakers produced /j/ without visible lip-rounding, indicating that non-native speakers’ incorrect articulatory configurations may decrease the degree of intelligibility. These results suggest that visual speech information may either positively or negatively affect L2 speech intelligibility.

© 2014 Acoustical Society of America. [http://dx.doi.org/10.1121/1.4892770]

PACS number(s): 43.71.Gv, 43.71.Hw, 43.71.Rt [CGC] Pages: 1352–1362

I. INTRODUCTION

The current study investigates how visual speech information in non-native speaker productions affects their intelligibility as native Canadian English perceivers evaluate English consonants produced by Japanese speakers based on audio-only (AO), visual-only (VO), or audio-visual (AV) speech input.

A. Effects of visual information on speech perception

In face-to-face conversation, visual as well as auditory cues are used to process speech input. Indeed, speech perception is an integrated process implementing both auditory and visual information, as revealed by the McGurk effect where the combination of visual /g/ and auditory /b/ cues results in the perception of /d/ upon integration of both modalities (McGurk and MacDonald, 1978). Research has shown that, in perceiving native language (L1) speech sounds, visual speech information revealed by articulatory configurations facilitates segmental perception of consonants and vowels (Summerfield, 1983; Jongman et al., 2003; Davis and Kim, 2004) as well as prosodic perception such as duration (De Jong, 1994; Sweerts and Krahmer, 2008). Such visual benefits are particularly effective in degraded environments such as in the presence of background noise (Sumby and Pollack, 1954; Macleod and Summerfield, 1990; Nielsen, 2004).

Likewise, visual information can provide an additional channel of input to facilitate perception as non-native perceivers perceive challenging, unfamiliar speech sounds in a second language (L2). For instance, when L2 stimuli were presented in an audio-visual (AV) condition compared to an audio-only condition, native English perceivers showed more accurate perception of those non-native sounds, e.g., in French (Reisberg et al., 1987), Korean (Davis and Kim, 1999, 2004), Irish, and Spanish (Erdener and Burnham, 2005). Similar findings have been revealed with native French, Japanese, Korean, and Mandarin learners of English who were better able to identify English speech contrasts presented with the addition of visual input (Werker et al., 1992; Hardison, 1999, 2003; Hazan et al., 2005; Hazan et al., 2006; Wang et al., 2008). Consistently, Spanish-dominant Spanish-Catalan bilinguals distinguished Catalan-specific vowel contrasts with AV input but not in audio-only and visual-only conditions (Navarra and Soto-Paraco, 2007).

Furthermore, research has shown that, for non-native perceivers, L2 stimuli may induce a greater degree of visual reliance as compared to L1 stimuli (De Gelder et al., 1995; Grassegger, 1995; Fuster-Duran, 1996). For instance, native Japanese perceivers showed greater McGurk illusion in the perception of English stimuli compared to Japanese stimuli, whereas the McGurk effect was larger for native English perceivers perceiving Japanese compared to English stimuli.

a)Author to whom correspondence should be addressed. Electronic mail: s.kawase@uws.edu.au
indicating a greater visual reliance in the perception of non-native speech stimuli (Sekiyama and Tohkura, 1993). Likewise, Thai and Japanese natives gave more visually influenced responses when perceiving McGurk stimuli in Mandarin and English than in their native languages (Chen and Hazan, 2007). These results suggest that visual input can serve as an additional channel of information in the perception of unfamiliar L2 stimuli; however, this visual reliance may not always lead to more accurate perception (such as the case of McGurk illusion).

Indeed, while non-native perceivers may be aided by visual speech information in an L2, their perception may also be impeded if they fail to perceive the correct L2 visual cues, rendering visual speech input to be inhibitory (Hardison, 1999; Ortega-Llebaria et al., 2001; Hardison, 2003, 2005a,b; Hazan et al., 2006; Wang et al., 2008). These inefficiencies are generally attributed to the linguistic experience of non-native perceivers, particularly influence from their L1 speech inventories (Hazan et al., 2005; Hazan et al., 2006; Wang et al., 2008, 2009). For example, native French and Mandarin elementary learners of English cannot accurately perceive the visual cues, or efficiently integrate the AV information, of the English interdental fricatives non-existent in French and Mandarin (Werker et al., 1992; Wang et al., 2008). Further research indicates that non-native perceivers’ successful use of L2 visual cues may depend on whether similar cues exist for their L1 categories. For instance, Hazan et al. (2006) showed that Spanish learners were more successful in perceiving visual cues for the English /b-v/ contrast compared to Japanese learners. Although neither Japanese nor Spanish has the voiced labiodental fricative /v/ in their speech inventories, Spanish does have the voiceless labiodental fricative /f/. The results revealed that the Spanish learners may have successfully associated visual cues in their native /f/ with the perception of the non-native /v/ which has the same place of articulation. Associating L1 and L2, the authors proposed three types of visual speech categories in the perception of L2 speech. The first type is when a visual cue is present in both the L1 and L2 (e.g., /b/ in both Japanese and English). The second type is when a visual category occurs in both L1 and L2, but with different acoustic-phonetic realizations (e.g., Spanish does not have the English voiced labiodental /v/ but contains a voiceless labiodental counterpart /f/). The last type is when a visual category or contrast occurs in the L2 but not in the L1 (e.g., the English labiodental fricatives /f/, /v/, or the /f-u/ contrast, which do not exist in Japanese). In the first two cases where visual categories are shared in L1 and L2, non-native perceivers have little difficulty. On the other hand, they would have trouble establishing L2 visual categories in the third case, when there is no equivalent visual category in the L1. These proposals are compatible with the L2 speech learning theories developed on the basis of auditory-acoustic domain (e.g., Speech Learning Model, SLM, Flege, 1995, 2007; Perceptual Assimilation Model, PAM, Best, 1995, and PAM-L2, Best and Tyler, 2007) which posit that category formation in an L2 is most likely if an L2 speech sound has a direct L1 counterpart based on acoustic or gestural articulatory information (or perceived to be phonetically distinct from the closest L1 sound), whereas the L2 sounds that are phonetically similar to but not the same as those in the L1 are often incorrectly assimilated to L1 speech categories. By extension, L2 visual speech learning may also be a matter of establishing visual speech categories under the influence of L1 visual speech cues.

Thus, L2 speech perception may be affected by the interaction of L1 and L2 on both auditory and visual domains. However, the auditory and visual cues of an L2 sound may not necessarily present the same challenge for a learner (Hardison, 1999; Hazan et al., 2006). In fact, L2 learners have been shown to apply different weighting of auditory and visual speech input as a function of their linguistic experience (Sekiyama, 1997; Ortega-Llebaria et al., 2001; Hazan et al., 2006; Wang et al., 2009). For example, for those L2 sounds with less distinctive acoustic cues but more prominent visual cues (e.g., /θ/ in English, Jongman et al., 2000), non-native perceivers tend to increase their visual weighting, leading to increased performance in the perception of these phonemes with visual input as compared to auditory-only input (Wang et al., 2009). Consistently, in the perceptual learning of L2 contrasts, the greatest benefit from visual cues occurs in the identification of the most difficult L2 stimuli, i.e., those contrasts that are least acoustically distinct (Hardison, 2005a). On the other hand, when the auditory perception of L2 sounds is good, non-native perceivers tend to rely less on the visual speech input (Wang et al., 2009). These findings suggest that the perception and acquisition of L2 sounds in the auditory and visual domains may take place in a complementary manner.

A related factor in the effective use of visual cues in speech perception is visual saliency (Hardison, 1999; Hazan et al., 2006). Compared to auditory information, visual information is significantly affected by the place of articulation of speech sounds. Those produced farther back in the vocal tract (e.g., alveolars and velars) appear to be less visually distinctive compared to those with prominent articulatory configurations around the mouth (e.g., labials and dentals). Thus, less visually salient L2 sounds may not produce significant visual benefits in perception, as revealed by AV perceptual training research showing greater learning effects for non-native learners perceiving the visible bilabial L2 speech contrast /b-v/ in comparison with the less visible contrast /f-u/ (e.g., Hazan et al., 2005).

The studies reviewed so far exclusively focus on how visual information affects non-native perceivers’ perception of L2 sounds produced by native speakers (e.g., native Japanese learners of English perceiving English sounds produced by native English speakers). Another type of “non-native perception” could involve non-native speakers with native perceivers (e.g., native English perceivers perceiving English sounds produced by native Japanese speakers). One would assume similar visual effects as non-native productions may involve correct or incorrect visual articulatory information and thus affect the intelligibility of these productions to native perceivers. While this speaker-perceiver pairing may extend the examination of L2 AV perception from “non-native perceivers” to “non-native speakers,” such research is very scarce. Yi et al. (2013) examined English
natives’ AV perception as well as “foreign-accent” rating of English sentences produced by native English (Caucasian-looking) and native Korean (Asian-looking) speakers. They found that visual (facial) speech information facilitated intelligibility of Korean-produced English sentences although the observed visual benefit was smaller than that for the native English-produced ones. Furthermore, Korean speakers were perceived as more accented than the native English speakers when visual (facial) information was presented, showing a strong visual bias, in that native English perceivers associated accentedness with speaker ethnicity as represented in the facial features of the speakers. Thus, this study suggests global (sentential level) positive (yet moderate) facilitative effects of visual information on the intelligibility of non-native productions.

B. Judgments of L2 speech intelligibility

Given the findings of visual effects of overall sentence-level intelligibility on accented speech perception, it is conceivable that additional visual cues in non-native speakers’ productions may influence intelligibility of specific speech sounds in non-native productions. However, with the exception of Yi et al. (2013), studies of L2 speech intelligibility judgments have predominantly focused on auditory-based assessments (e.g., Flege et al., 1995; Munro et al., 1996; Munro and Derwing, 2008). For example, for evaluation of speech sounds produced by L2 learners, native listeners are typically asked to identify and rate L2 productions (e.g., Munro and Derwing, 2008) to evaluate how accurate the productions are (e.g., Larson-Hall, 2006) and assess their degree of “accentedness” based on auditory speech input (e.g., Flege et al., 1995; Munro et al., 1996). These listener-based assessments inevitably result in the issue of inter-rater reliability. For instance, inexperienced native listeners’ judgments are found to differ from those of phonetically trained evaluators (Munro, 2008). Although expert evaluators can more accurately evaluate specific phonetic features, naive listeners’ judgments may enjoy greater external validity in that most conversation is exchanged without formal training in phonetics. However, a further issue thus arises as to how accurate these naive listeners’ judgments are, as they may incorrectly identify a non-native sound production either because it was incorrectly produced or because the judge’s own perception was not correct, particularly when the acoustic features of the sound are not distinctive enough. Indeed, previous research has shown that native listeners of English could not accurately distinguish non-sibilant English fricatives based on auditory input only, although their perception became more accurate when visual speech information was provided (Jongman et al., 2003).

These issues necessitate further research with the inclusion of visual input in non-native speaker intelligibility judgment research. The question here is how visual information in speech segments produced by non-native speakers influences native perceiver evaluations, either in the case of correct visual speech configurations or in the case of incorrect visual configurations. Compared to auditory-only studies, research including both auditory and visual domains can more fully reveal patterns of native and non-native interactions in face-to-face communication.

C. The current study

As reviewed above, visual speech information can increase perceptual accuracy in perceiving both L1 and L2 speech sounds. However, these studies focus on visual benefits in the context of native perceivers perceiving L1 sounds (e.g., Jongman et al., 2003; Summerfield, 1983) or non-native perceivers perceiving L2 sounds (e.g., Hardison, 1999; Hazan et al., 2006; Wang et al., 2009), both produced by native speakers of the target language. It remains to be determined to what extent additional visual cues influence the native perception of speech productions by non-native speakers. While Yi et al. (2013) provides significant insight into the visual effects on the overall intelligibility of non-native sentence productions, no research to date has focused on the segmental level with specific speech contrasts.

Thus, the current study aims to examine how visual segmental speech information in non-native speakers’ productions affects the intelligibility of these sounds as judged by native perceivers. Given that L2 speech learning models (e.g., PAM-L2, Best and Tyler, 2007) make specific predictions of how the perception of L2 sounds may be guided by articulatory information in the L1 speech categories, it can be assumed that the articulatory information in non-native speakers’ productions may affect native perception through visual input in addition to auditory input. Furthermore, it is highly possible that non-native speakers’ articulatory constellations may contribute to different judgments beyond their auditory information in both positive and negative ways depending on whether the articulations are correct and how they are judged. This investigation may thus extend the existing L2 speech findings and theories from the auditory to visual domains, in hopes of advancing our understanding of L2 speech interactions in face-to-face communication.

In the present study, native Canadian English perceivers were asked to identify the initial consonants /b,v,s,θ,l/ in English consonant-vowel (CV) syllables produced by Japanese learners of English based on auditory-only (AO), visual-only (VO), and audio-visual (AV) input. Among the six consonants, the labial and dental phonemes /b,v, θ/ and the rounded /l/ are considered more visually salient than the alveolars /s,l/, as the former involve more visible articulatory movements (Hazan et al., 2006). In terms of the “nativesness” of these sounds, since Japanese does not have /v/ and /θ/, but has /b/ and /s/, Japanese learners of English tend to replace the non-Japanese phonemes with similar counterparts in their L1 (i.e., /b/ with /b/; /θ/ with /s/, Yoshida and Hirasa, 1983). Furthermore, Japanese speakers have difficulty producing and perceiving the /l/ contrast, which does not exist in Japanese (although the Japanese lateral flap /ɾ/ is often taken as a closer match with /l/; e.g., Takagi, 1993). Accordingly, Japanese speakers in the current study are also expected to display less intelligible productions of the non-native phonemes /v, θ, l/ due to the interaction of their L1 and L2 inventories.

On the basis of the previous studies showing an increase in visual cue weighting on the perception of non-native
speech (Sekiyama and Tohkura, 1993; Chen and Massaro, 2004; Chen and Hazan, 2007) and visual facilitative effects among the speech segments shared by both L1 and L2 (Hazan et al., 2006), we hypothesize that native Canadian English judges may perceive the Japanese-produced consonants as more intelligible in the AV condition compared to the AO condition (i.e., positive visual effect) for consonants that exist in both Japanese and English (i.e., /b, s/, as well as /l/ which is similar to /r/ in Japanese). On the other hand, the non-Japanese consonants /v, ð, l/ could be expected to have negative visual effects on intelligibility in the AV and VO conditions when the speakers produce incorrect articulatory configurations. Moreover, we expect these visual effects to interact with visual saliency, with intelligibility being more affected by the visually salient consonants than the less visually salient consonants.

II. METHODS

A. Speakers

Fifteen native speakers of Japanese (eight female, seven male) participated in this study. The age range was from 19 to 31 years old (mean age = 25 years). The Japanese speakers had been living in Vancouver, Canada, as students or workers for an average of 10.2 months (range: 1 week to 3 years), and they were from various cities in Japan. In addition, none of them used English at home or had lived in an English speaking country before arriving in Canada. The Japanese speakers started to learn English as a foreign language during middle school (at the age of 12), which was also their initial exposure to English. According to self-reports, their daily English input was limited (mean daily input of English = 51%; mean daily input of Japanese = 49%). They were thus considered late, intermediate level English learners at the time of testing (cf. Flege et al., 1997).

An additional 15 native speakers of Canadian English (eight female, seven male) participated as a control group. The age range was from 18 to 31 years old (mean age = 22 years). Since this study uses visual information through the speaker’s full face for the identification task, ethnic information from their facial features may crucially affect perceivers judgments. Unlike Yi et al. (2013), which included speaker ethnicity as a factor, the present study focuses on articulatory phonetic effects only. Thus, in order to exclude speaker ethnicity as a variable between the speakers of Japanese and Canadian English, only native English speakers of East-Asian descent (e.g., China, Japan, and Korea) were recruited. All of them were born in Canada with English being their first and dominant language. All the Japanese and English speakers reported that they did not have any speech impairments. They were compensated for their participation.

B. Stimuli

Six English CV syllables, having one of the initial consonants /b, v, ð, l, ð/ followed by the vowel /a/ were used as stimuli. As described above, Japanese learners of English are likely to substitute the non-native phonemes with the similar counterparts in their L1 (i.e., /v/ with /b/; /ð/ with /s/; /l/ with /l/ which is similar to the Japanese /r/). We hypothesize that native Canadian English judges may perceive the Japanese-produced /v/ and /l/ as “native-like (or Japanese-like) phonemes,” and /ð, ð, l/ as “non-native (or non-Japanese) phonemes.” Additionally, /b, v, ð/ (with visible labial and dental articulatory configurations) and /l/ (with visible lip rounding) are considered more visually salient than the alveolar /s, l/ (Hazan et al., 2006). The characteristics of the stimuli are summarized in Table I.

C. Recording procedures

Audio and video recordings were made of the native speakers of Japanese and English individually. Prior to the recording, they were given instructions orally and with a written instruction sheet, and they were asked to read aloud each of the six CV stimuli in citation form at their normal speed and volume. PowerPoint was used to present five repetitions of the six CV stimuli in isolation in a randomized order. The recording of the speaker’s full face was made with a digital camcorder (Canon Vixia HF S30 HD Video camcorder). Because of the low sound quality recorded by the on-camera microphone, concurrent separate audio recordings were made with a Shure KSM 109 condenser microphone using SoundForge 6.4 at a 48 kHz sampling rate. In order to maintain consistency in the stimulus materials, and to avoid any distractions for the perception tasks, all the speakers were seated in front of blue monochrome wallpaper and their faces were centered in the frame. The recordings for each speaker lasted approximately 30 min and were carried out in the recording studio in the Language and Brain Lab at Simon Fraser University (SFU).

D. Stimulus editing

After collecting both audio and video files, three sets of stimuli (AO, VO, and AV) were prepared for the native perceptual evaluation tasks. First, the intensity of the audio recordings was normalized using SoundForge 6.4 to have the same unweighted RMS value. The normalized audio files

<table>
<thead>
<tr>
<th>TABLE I. The stimulus set.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target consonant</td>
</tr>
<tr>
<td>Place of articulation</td>
</tr>
<tr>
<td>Occurrence</td>
</tr>
<tr>
<td>Note</td>
</tr>
</tbody>
</table>
were synchronized with the video files using Final Cut Pro X, after which the audio files collected by the camcorder were deleted. The clip length for each stimulus was approximately 3000 ms containing 1200 ms of a neutral face before and after the stimulus. Thus, each stimulus captured the mouth opening as well as closing. The resolution was 640 pixels × 480 pixels. The video-recorded files with the synchronized opening as well as closing. The resolution was 640 pixels × 3000 ms containing 1200 ms of a neutral face before and after which the audio files collected by the camcorder were synchronized with the video files using Final Cut Pro X, whereas the AO condition and the VO condition were made 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, after editing with Final Cut Pro.

E. Perceivers

Thirty-one native speakers of Canadian English (16 male, 15 female) participated as perceivers (judges) to identify the stimuli produced by the speakers. All were students at SFU, aged between 18 and 42 yr old (mean age = 23.4). None of the perceivers had experience with Japanese learning and they had limited exposure to Japanese-accented English. All the perceivers reported that they had normal hearing, and normal or corrected vision. They were compensated for their participation. Because of file-loading issues in the E-Prime 2.0 stimulus presentation software (Psychology Software Tools), data from two male perceivers were excluded.

F. Perceptual evaluation tasks

The native English perceivers (judges) were tested individually in a sound-treated booth in the Language and Brain Lab at SFU while wearing headphones to listen to the stimuli presented using E-Prime. While the stimuli were CV syllables, the judges were asked to focus on the initial consonants. The stimuli were presented in three modalities, where the perceivers either saw the articulatory movements (speaker face) in the center of the screen and heard the stimulus over the headphones in the AV condition, or only heard the stimulus in the AO condition, or only watched the articulatory movements in the VO condition. All perceivers completed each of the three modality conditions, which were blocked, with the order of modality presentation counter-balanced across perceivers. Each modality contained 180 randomly presented stimuli, including all the six consonants (/b, v, s, ŋ, l, ʃ/) from all the speakers in both speaker groups (15 Japanese and 15 English). Each modality condition lasted approximately 30 min, for a total testing time of 90 min with the three modalities, completed in two visits within the same week. Two modality conditions were completed on the first visit, with a 10-min break between them, and the third modality condition was completed on the second visit. In the identification task, a fixation point was displayed for one second for each trial prior to the target stimulus. Response alternatives were presented on the screen with the corresponding consonant pairs (i.e., /b/ and /v/; /s/ and /ʃ/; / or /l/ and /ʃ/; / or /l/ and /ʃ/; / or /l/ and /ʃ/, where “th” and “v” were used to replace /θ/ and /ʃ/, respectively, since the perceivers were not familiar with the phonetic symbols. The perceivers were to press one of the six buttons on the keyboard labeled with the six response alternatives as shown on the screen. Additionally, they were given an option to press one of ten other alternative buttons if they felt that they perceived an English consonant other than the presented response alternatives. These options were the remaining ten English consonants differing in either place or manner of articulation from the six presented ones. All the consonant options were labeled as letters on the keyboard (e.g., “sh” for /ʃ/). Prior to testing, the perceivers were given instructions as well as a practice session to be familiar with the corresponding sound-label match patterns. They were allowed a maximum response time of 4 s in each trial. All of the participants received monetary compensation.

III. RESULTS

The native English perceivers’ mean correct identification data were analyzed using a three-way repeated measures analysis of variance (ANOVA) with speaker group (Japanese, English), consonant type (/b, v, s, ŋ, l, ʃ/), and modality type (AO, VO, AV) as the within-group factors. Significant main effects were observed for group [F(1, 28) = 1562.5, p < 0.001, partial η² = 0.982], consonant [F(5, 140) = 89.8, p < 0.001, partial η² = 0.762], and modality [F(2, 56) = 55.8, p < 0.001, partial η² = 0.666]. Moreover, significant interactions were found for group × consonant [F(5, 140) = 228.4, p < 0.001, partial η² = 0.891], group × modality [F(2, 56) = 27.2, p < 0.001, partial η² = 0.493], consonant × modality [F(10, 280) = 18.0, p < 0.001, partial η² = 0.392], and group × consonant × modality [F(10, 280) = 13.149, p < 0.001, partial η² = 0.320]. Further analyses were thus conducted to explore these effects.1

A. Consonant intelligibility in each modality

First, sets of two-way (speaker group and consonant) repeated measure ANOVAs were conducted for each modality (AO, VO, and AV) to see how Japanese and English productions were perceived differently for the different consonants in each modality. Figure 1 presents the mean correct identification rates of each target consonant produced by Japanese and English speakers in each modality.

In the AO condition, significant main effects were found for group [F(1, 28) = 1582.5, p < 0.001, partial η² = 0.983] and consonant [F(5, 140) = 74.3, p < 0.001, partial η² = 0.726], along with a significant interaction between group and consonant [F(5, 140) = 86.1, p < 0.001, partial η² = 0.755]. Bonferroni adjusted post hoc tests revealed that the mean correct identification rates differed as a function of speaker group and nativeness of the consonants. As expected, all six consonants produced by the Japanese speakers were significantly less well identified compared to the English productions (p < 0.001). Within the Japanese productions, the native-Japanese-like /b, s, l/ were significantly more intelligible (mean correct identification: /b/ = 91%, /s/ = 93%, /l/ = 88%) compared to the non-native /v, ŋ, ʃ/ (p < 0.001). Among the non-native consonants, /v/ (39%) was significantly less intelligible than /ʃ/ (57%) and /ʃ/ (55%). The judges also showed different identification patterns for the six consonants produced by the English speakers. The perception of /ʃ/ (88%) was significantly less accurate than...
significant difference was observed. which in turn was more accurate than /l/ (p < 0.01). Additionally, /b/ (95%) was significantly less accurate than /l/ (p = 0.043). There were no other significant differences.

In the VO condition, significant main effects of group [F(1, 28) = 296.7, p < 0.001, partial η² = 0.914] and consonant [F(5, 140) = 20.2, p < 0.001, partial η² = 0.419] were observed, along with a significant interaction [F(5, 140) = 124.2, p < 0.001, partial η² = 0.791]. Bonferroni post hoc tests revealed that visual perception differed as a function of speaker group, nativeness of the consonants as well as visual saliency. While /b, s/ did not show significant group differences /b/: JP = 93%, ENG = 93%; /s/: JP = 74%, ENG = 76%, the identification accuracies of non-native /v, θ, ι/ produced by the Japanese speakers (/v: 51%, /θ: 60%, /ι: 35%) were significantly lower than those produced by the native English speakers (/v: 89%, /θ: 96%, /ι: 84%) (p < 0.01). In contrast, the identification of Japanese-produced /l/ (78%) was significantly more accurate than those produced by the English speakers (59%). Within the Japanese productions, the more visually salient /b/ was significantly more intelligible than /l, s/ (p < 0.01), which were then in turn better perceived than the non-native /v, θ/ (p < 0.01) and the least intelligible consonant /ι/ (p < 0.01). For the English productions, as expected, the more visually salient /θ, b, v, ι/ were more intelligible than /s, l/ (p < 0.05).

Among all the consonants, the perception of /θ/ was the most accurate (even though in AO it was the least accurately perceived one), better perceived than /v, s, l/ (p < 0.01). Additionally, /b/ was perceived more accurately than /s, l/, which in turn was more accurate than /θ/ (p < 0.05). No other significant difference was observed.

For the AV condition, significant main effects of group [F(1, 28) = 1105.2, p < 0.001, partial η² = 0.975] and consonant [F(5, 140) = 160.0, p < 0.001, partial η² = 0.851] were also observed, as well as a significant interaction of the two [F(5, 140) = 162.6, p < 0.001, partial η² = 0.853]. Bonferroni post hoc analyses show that the mean correct identification rates of the English productions of /b/ (99%), /v/ (99%), /θ/ (98%), /ι/ (99%), and /l/ (100%) were at ceiling and significantly higher than the Japanese productions (p < 0.05), with the exception of /s/ for which both groups reached ceiling (ENG = 99%, JP = 98%, p = 0.676). For the Japanese group, the native-like /b/ (98%), /s/ (98%), and /l/ (93%) were perceived significantly more accurately than the non-native /θ/ (64%), /v/ (52%), and /ι/ (52%) (p < 0.005), among which the intelligibility of /θ/ was significantly higher than /v, ι/ (p < 0.005), whereas the latter two did not differ (p > 0.05). For the English productions, no statistical difference in consonant intelligibility was found.

Overall, across modalities the English productions were more intelligible than Japanese productions, particularly for the non-Japanese /v, θ, ι/. The intelligibility of the two groups did not differ for /s/ in the AV condition and /b/ in VO due to ceiling, and they both showed relatively poorer intelligibility for /s/ in VO, presumably because /s/ was less visually salient. The Japanese-produced /l/ in the VO condition was more intelligible compared to the English productions. For the Japanese productions, as expected, the Japanese-like /b, s, l/ tended to be more intelligible than the non-native /v, θ, ι/ in all modalities. In addition, modality type appeared to affect intelligibility. For instance, in the AO condition, /v/ was the least accurately identified, at a rate significantly lower than /ι/. In contrast, in VO, /ι/ was the least accurately identified. These results indicate that perceptual accuracy may differ as a function of input modality. Thus, further analyses were performed to examine how modality type affects consonant intelligibility.

B. Effects of modality type

To examine the effects of modality type, separate two-way (consonant and modality) repeated measure ANOVAs...
were conducted for the Japanese and English speaker groups. Figure 2 displays modality comparisons of the mean correct identification rates of each target consonant produced by Japanese and English speakers (data replotted from Fig. 1).

The Japanese group showed significant main effects of modality \(F(2, 56) = 35.5, p < 0.001\), partial \(\eta^2 = 0.559\) and consonant \(F(5, 140) = 179.6, p < 0.001\), partial \(\eta^2 = 0.865\) as well as their interaction \(F(10, 280) = 10.7, p < 0.001\), partial \(\eta^2 = 0.276\). Bonferroni adjusted pairwise comparisons for modality revealed that the overall mean correct identification rates in AV (76%) were significantly higher than AO (71%, \(p < 0.005\)), which was in turn significantly higher than VO (65%, \(p < 0.05\)). Bonferroni post hoc tests also revealed how modality type affected the perception of the different consonants. First, the visually salient consonants /b, v, θ/ patterned similarly in that their intelligibility in the AV condition was significantly higher than in the AO condition (\(p < 0.01\)), with VO being intermediate to these two conditions (i.e., for /b, θ/ it was not significantly different from either condition, \(p > 0.05\); for /v/, both AV and VO were more intelligible than AO (\(p < 0.005\)). In comparison, for the less visually salient /s, l/, the intelligibility in the AV condition was significantly higher than that in the VO condition (\(p < 0.01\)), with AO being intermediate (i.e., for /s/ the intelligibility in AO was significantly lower than in AV but higher than in VO, \(p < 0.01\), but for /l/ AO did not differ from either of the other two conditions, \(p > 0.05\)). However, contrary to these general patterns of AV being the most intelligible condition, the pattern for /l/ was the only exception, where intelligibility significantly decreased from AO to AV, and even further in the VO condition (\(p < 0.05\)).

The analysis with the English controls also produced a significant main effect of modality \(F(2, 56) = 64.6, p < 0.001\), partial \(\eta^2 = 0.698\), with the overall mean consonant identification rate in AV (99%) being significantly higher than in AO (96%, \(p < 0.05\), which was in turn higher than in VO (83%, \(p < 0.005\)). The results also showed a main effect of consonant \(F(5, 140) = 14.3, p < 0.001\), partial \(\eta^2 = 0.337\) along with a significant modality \(\times\) consonant interaction \(F(10, 280) = 25.4, p < 0.001\), partial \(\eta^2 = 0.476\). Bonferroni pairwise modality comparisons for individual consonants showed that, for the visually salient /b, v, θ/, intelligibility in the AV condition was consistently higher than in the AO condition (\(p < 0.01\)), with the VO condition being on par with AV for /b/ but equal with AO for /b, v/ (\(p > 0.05\)). In contrast, the less visually salient but more acoustically salient /s, l/ as well as /l/ showed the expected pattern of higher intelligibility in the AV and AO conditions than in the VO condition (\(p < 0.05\)).

Overall, the integration of AV information appears to improve the intelligibility of Japanese as well as English productions, as perception was mostly more accurate in the AV than AO or VO conditions. Comparing the different consonants across modalities, the perception of the Japanese productions patterned with that of the English group in that AV benefits occurred in the more visually salient consonants /b, v, θ/, while for the less visually salient /s, l/, the productions in VO were significantly less intelligible than in the AO and AV conditions where correct identifications almost reached ceiling. However, unlike the English /l/ productions with equally good intelligibility in both AV and AO, the Japanese-produced /l/ was less intelligible in the AV than in the AO condition, indicating a negative effect of visual input. Thus, the intelligibility of the English consonants produced by native Japanese speakers as well as native English controls was greatly affected by modality type, in that visual information played a significant role in changing perceptual accuracy rates in both facilitative and inhibitory ways.

C. Follow-up analysis: Lip-rounding effect on the perception of /l/

The results reported above indicated a significant negative visual effect with the intelligibility of the Japanese-
produced /\textipa{\textael}/. Thus, further analysis was conducted to unravel what may have triggered this negative effect. Since the articulatory configuration of the English /\textipa{\textael}/ involves lip-rounding which has been shown to be visually salient in AV perception (Traunmüller and Öhrström, 2007), it could be speculated that lip-rounding in the productions of English /\textipa{\textael}/ may facilitate correct perception. However, if the Japanese speakers had assimilated /\textipa{\textael}/ to the Japanese unrounded lateral flap /\textipa{\textael}/ (Miyawaki et al., 1975), it could be assumed that the current results of the low visual intelligibility of /\textipa{\textael}/ may have been due to the lack of lip-rounding in the Japanese productions. In order to test the above hypothesis, an additional two phonetically trained Canadian English perceivers were asked to judge the degree of lip-rounding of /\textipa{\textael}/ produced by the same 30 native Japanese and English speakers in the VO condition. Roundedness was rated on a five-point scale (1: no lip-rounding, 5: full lip-rounding), and the judges were encouraged to use the full scale. The two judges’ data revealed a high inter-rater reliability, with the Cronbach’s Alpha being 0.898. The rating results showed that the English productions were judged to have a significantly higher degree of lip-rounding (Mean = 3.8) than the Japanese productions (2.1) \(t(29) = -5.48, p < 0.001\). A correlation analysis was performed to examine whether visual speech intelligibility (from the identification results) is correlated with the degree of lip-rounding in the production of /\textipa{\textael}/ (from the rating results). As Fig. 3 shows, a strong positive correlation between the identification accuracy in the VO condition and the degree of lip-rounding in the Japanese-produced /\textipa{\textael}/ was found \(r = 0.853, p < 0.001\), indicating that more visible lip-rounding may relate to the higher visual speech intelligibility of /\textipa{\textael}/. Thus, incorrect articulatory configurations by the Japanese speakers, namely, the lack of lip-rounding, decreased the intelligibility of visual speech information, resulting in the inhibitory visual effects in AV speech perception of Japanese-produced /\textipa{\textael}/. However, the English productions did not show such a correlation \(r = 0.142, p > 0.05\), possibly due to the relatively high degree of homogeneity of the native English productions with a near-ceiling effect where both the lip-rounding rating and visual perception accuracy scores were clumped at the higher end of the range.

D. Summary of the results

Overall, visual information affected the intelligibility of speech sounds produced by non-native speakers as perceived by native judges. Particularly, visual facilitative effects were observed among the visually salient consonants (/b, v, \textipa{\textael}/), including the non-native /v, \textipa{\textael}/ for Japanese speakers. For the less visually salient consonants (/s, l/), on the other hand, the native judges relied mostly on auditory input. Moreover, visual input may also result in a negative effect. Japanese-produced /\textipa{\textael}/ showed lower intelligibility with the presence of visual information, which can be attributed to the Japanese speakers’ incorrect articulation (i.e., lack of lip-rounding).

IV. DISCUSSION

A. Facilitative and inhibitory visual effects

Extending the previous findings on the perception of speech produced by native speakers that visual speech information could increase perceptual accuracy of native and non-native sounds (e.g., Reisberg et al., 1987; Werker et al., 1992; Hardison, 1999; Jongman et al., 2003; Hazan et al., 2006; Wang et al., 2009), the current results further show that visual speech information may also affect native judgments of the intelligibility of the speech sounds produced by non-native speakers. This research also extends the previous finding of an overall visual effect on the perception of sentential-level non-native speaker utterances (Yi et al., 2013) to the effects of specific segmental-level speech contrasts.

![Image](attachment://image.png)
As expected, the results support the current hypothesis of positive visual effects on the intelligibility of the native Japanese-like phonemes (i.e., /b, s/) produced by the Japanese speakers. The Japanese-like /l/ revealed a tendency of similar visual effects with the AV condition scoring higher than the AO condition, although this difference was not statistically significant. It should also be noted that the visual intelligibility of these three phonemes all reached the native level as the VO perceptual accuracy for the Japanese productions was on par with that for the English productions. As /b, s, l/ are similar in English and Japanese, the Japanese speakers should not have difficulty producing correct articulatory configurations that would be detectable by native judges. The smaller visual effects and overall lower scores for /l/ compared to /b, s/ are presumably because /l/ is similar to Japanese /l/ to a lesser degree than how similar /b, s/ are to their Japanese counterparts. Most notably, the results show that visual information increased the perceptual intelligibility of the Japanese productions even for the non-Japanese consonants (/v, /), contradicting the hypothesis of inhibitory visual effects for such sounds. Despite the overall poorer intelligibility of these Japanese-produced consonants (compared to the English-produced ones), perceptual accuracy was higher in the AV than the AO condition, suggesting that the native perceivers may have effectively used the visually salient cues when they were produced correctly by the non-native speakers.

Moreover, visual benefit in the intelligibility of the Japanese-produced consonants appears to interact with the auditory domain as a function of auditory and visual saliency. Unlike the visually non-salient /s, l/ which were less intelligible in the VO than the AV condition, for the visually salient /v, /, intelligibility in the VO condition was on par with the AV condition, both of which were higher than the AO condition. As has been claimed previously, greater visual reliance may be required for non-sibilant consonants such as /v, / that have relatively less intelligible auditory cues, as compared to the more auditorily salient consonants such as the sibilant /s/ (Jongman et al., 2003). Thus, the degree of visual weighting is inversely related to the ambiguity of auditory information ( Sekiyama and Tohkura, 1993; Chen and Massaro, 2004; Chen and Hazan, 2007). The current results are consistent with these accounts in that the native English perceivers may have increased visual reliance in accordance with the lower intelligibility in the auditory cues of the Japanese-produced /v, /, whereas they likely relied more on the auditory domain for the visually less distinctive /s, l/. The results, therefore, extend the previous findings from the perception of native productions to non-native productions, indicating how visual benefits complement auditory information in the intelligibility of non-native productions.

On the other hand, the results support the hypothesis of negative visual effects on the intelligibility of non-native productions in that the perception accuracy of the Japanese-produced /l/ was lower in the AV condition compared to the AO condition. Further analysis revealed that the visual intelligibility of the Japanese-produced /l/ was correlated with the perceived degree of lip-rounding, a feature characteristic of native English /l/ production (Traunmüller and Öhrström, 2007). As a significant portion of the Japanese productions of /l/ lacked lip-rounding, incorrect articulatory configurations may have inhibited intelligibility. These results suggest that visual cues in non-native speaker productions could affect speech intelligibility in detrimental as well as facilitative ways. While previous research has shown how visual speech information benefits perception of native speaker productions by native perceivers (Sumby and Pollack, 1954; Macleod and Summerfield, 1990; Nielsen, 2004) as well as by non-native perceivers (Davis and Kim, 1999; Hazan et al., 2006; Wang et al., 2009), research has not yet addressed negative visual effects resulting from incorrect articulation as the productions were all from native speakers in those contexts. The present findings thus add to the AV literature in showing negative visual effects for non-native productions where articulatory configurations were incorrect.

B. Theoretical implications

The above results therefore provide evidence extending the auditory-based L2 speech learning theories (e.g., SLM, Flege, 1995, 2007; PAM-L2, Best and Tyler, 2007) to the visual domain. As previously reviewed, these theories predict that the interaction of L1 and L2 phonetic categories lies in the perceived similarities of the two sound systems, where an L2 sound may be incorrectly assimilated to an L1 category in its neighborhood. While these accounts focus on auditory-acoustic categories, the current results show that the relationship of L1 and L2 speech sounds can be defined by visual categories, in that L1 influences on L2 phonemes are observable by native perceivers through the articulatory movements in non-native speaker productions. For instance, the patterns of the Japanese productions of the English /l/ and /l/ can be explained in terms of the visual categories relating these two sounds and the Japanese /r/. The English /l/ and Japanese /r/ share the same articulatory configuration involving unrounded lips. In English this feature is used contrastively with lip-rounding (for /l/), whereas the contrast does not exist in Japanese. Thus, for /l/, sharing the same visual category with a native phoneme (/r/) may result in the Japanese speakers’ correct articulation (category formation), contributing to its high intelligibility as judged by native English perceivers. This is in contrast to the low intelligibility of /l/ which was predominantly produced with unrounded lips by the Japanese speakers, presumably being assimilated to the Japanese /r/ category. These results indicate that the interaction of L1 and L2 speech categories may occur in the visual domain. However, it should be noted that while the SLM and PAM-L2 predictions are based on learners’ perceived similarities between L1 and L2 sounds, the current patterns are from Japanese speakers’ productions. This raises the subsequent question as to whether the Japanese speakers themselves also perceive the visual information as revealed in their productions. Previous auditory-visual research has shown that Japanese learners of English can more accurately produce the English /l/-/ contrast after they are trained to perceive the visual articulatory information to differentiate these sounds (Hardison, 2003), suggesting the positive relationship between production and perception in L2 speech learning. Further research could include non-native
perception data to investigate the perception-production link in the visual learning of L2 speech.

Moreover, the results indicate that the interaction of L1 and L2 sounds may not present the same challenge in terms of establishing auditory and visual categories. In the current results, the intelligibility of the Japanese-produced /v, θ/ was poor in the audio-only condition but was significantly higher in the AV condition, showing the emergence of new visual categories involving the “(labio- or inter-) dental” place of articulation non-existent in Japanese. These findings may also shed light on the perceptual correlation of auditory and articulatory cues, implying that the intelligibility of speech produced by non-native speakers could potentially be affected by the interaction of auditory and visual information.

Indeed, in expanding the L2 speech learning theories to account for both auditory and visual perception, research needs to take into consideration factors specific to visual speech perception as well as the interaction of the two modalities. In this respect, previous research has noted that the degree of auditory and visual reliance may vary due to visual saliency and demand for visual information, in that visually salient phonemes could produce greater visual effects in AV perception (Hardison, 1999; Hazan et al., 2006; Wang et al., 2009). Consistently, in the current study, the visual facilitative effects were most strongly observed in the visually salient consonants (i.e., /b, v, θ/) in both English and Japanese productions. Furthermore, the results of the Japanese productions show discrepancies between auditory and visual intelligibility. For instance, while the visual intelligibility of the non-Japanese /h/ was significantly lower than the rest of the non-native and native phonemes (due to its incorrect articulation), another non-Japanese phoneme, /v/, was the least auditorily intelligible (presumably due to its weak acoustic saliency as a non-sibilant fricative). Thus, since visual speech intelligibility is not always aligned with auditory intelligibility, it is important to consider the extent to which visual and auditory information is required and integrated for efficient perception of non-native speakers’ productions.

V. CONCLUDING REMARKS

The results of this study reveal that visual speech information plays a significant role in perceived degree of segmental intelligibility of non-native speaker productions. While visual speech information is generally facilitative in the native perception of non-native speech, the effect may also be inhibitory in that incorrect articulatory configurations may lead to decreased visual speech intelligibility. Thus, the current results offer valuable insights into issues in perceiving and assessing non-native speaker productions, complementing the previous research which has primarily been based on auditory judgments (e.g., Flege et al., 1995). Together, these findings suggest that audio-visual speech information is not only integrated in non-native productions but is also observable by native perceivers. Understanding the extent to which multi-modal speech information is used in non-native productions and assessments may facilitate effective cross-linguistic face-to-face communications.

ACKNOWLEDGMENTS

The authors would like to thank Nisha Banga, Daniel Chang, Mathieu Dovan, Keren Hernandez, and Alyssa Lee at Simon Fraser University (SFU) for their assistance in stimulus development and data collection as well as Michael McAuliffe at the University of British Columbia for graphics assistance. We would also like to thank the following colleagues for their valuable comments: Dr. Yukari Hirata (Colgate University), Dr. Murray Munro (SFU), and Dr. Michael Tyler (University of Western Sydney). We would also like to extend our gratitude to all the participants. This research was conducted as S.K.’s master’s thesis at SFU under Y.W.’s supervision, and has been funded by research grants to Y.W. from SFU Vice President Research and the Social Sciences and Humanities Research Council of Canada (SSHRC Standard Grant 410-2006-1034, and SSHRC Insight Grant 435-2012-1641). Portions of this study have been presented at and published in the proceedings of Interspeech 2012.

1 As described in the Methods section, the participants were given the option to type in a response other than the two offered. The results showed that such responses were rare (below 10%) with no consistent misperception patterns for any consonant. Therefore, these responses were not separately analyzed.


