Effects of Visual Speech Information on Native Listener Judgments of L2 Speech Consonants

by

Saya Kawase

B.A., Waseda University, 2010

Thesis Submitted in Partial Fulfillment of
the Requirements for the Degree of
Master of Arts

In the
Department of Linguistics
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY

Summer 2012

Approval

Name: **Saya Kawase**

Degree: **Master of Arts**

Title of Thesis: ***Effects of visual speech information on native listener judgments of L2 speech consonants***

Examining Committee:

Chair: **Dr. Chung-hye Han**
Associate Professor of Linguistics, Simon Fraser University

Dr. Yue Wang
Senior Supervisor
Associate Professor of Linguistics, Simon Fraser University

Dr. Murray Munro
Supervisor
Professor of Linguistics, Simon Fraser University

Dr. Yukari Hirata
External Examiner
Associate Professor of Japanese, Chair of East Asian Language & Literature, Colgate University, USA

Date Defended/Approved: August 2, 2012

Partial Copyright Licence

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics,

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University;

or has conducted the research

c. as a co-investigator, collaborator or research assistant in a research project approved in advance,

or

d. as a member of a course approved in advance for minimal risk human research, by the Office of Research Ethics.

A copy of the approval letter has been filed at the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

update Spring 2010

Abstract

Research on the intelligibility of non-native (L2) speech productions has focused on native listener judgments of auditorily presented L2 productions. However, little research has explored how visual information in L2 speech productions affects native listeners' perception. In the present study, native Canadian English perceivers were asked to identify six English phonemes /b, v, s, θ, l, ɹ/ produced by native speakers of Japanese and native speakers of Canadian English as controls under three input modalities: (1) audiovisual (AV), with simultaneous presentation of speaker voice and facial/mouth movements, (2) audio-only (AO), with speaker voice only, and (3) visual-only (VO), with speaker face only. The results show facilitative effects of visual speech information on the intelligibility of non-native productions as well as deteriorative effects due to lack of visible lip-rounding in the Japanese-produced /ɹ/. These results suggest visual speech information may either positively or negatively affect the intelligibility of L2 productions.

Keywords: audiovisual speech perception of L2 consonants; Japanese learners of English.

To my family

Acknowledgements

I am one of the most fortunate students in that I have had numerous opportunities to conduct and assist research in the Language and Brain Lab (LABlab) at Simon Fraser University (SFU). First of all, I would like to thank my supervisor, Dr. Yue Wang, who has given me considerable support. Through the two years I've spent as a LABer, my brain and mind were trained. I never happened to have an easy time, but I enjoyed it even though the tough times. Thank you so much, Yue!

I would also like to thank Beverly Hannah, Lindsay Walker, Daniel Chang, Alyssa Lee, Keren Hernandez, Nisha Banga, and Mathieu Dovan at the LABlab for their assistance and mental support. Starting from Summer 2011, this project was completed after four semesters. It was a long journey, and yet your kindness saw this project through to the end. Thank you!

Besides lab activities, I was lucky and grateful to have opportunities to conduct research in classes supervised by Dr. Murray Munro. It was an honour and a lot of fun to work with a researcher who has made significant contributions in the area of second language speech learning. I won't forget the time we had chatting when we came back from Ling 813 class with Kazuya. Thank you, Murray!

I also appreciate research guidance and support from my external committee member, Dr. Yukari Hirata. Thank you for taking your time to review my paper, and flying all the way from New York for my defense.

As my research interests are second language speech perception and production, human participants for testing are always necessary. I would like to thank all of the participants (approximately 160 during the past two years) who have contributed to further understanding of second language speech processing. In order to achieve this number of participants, a recruitment campaigns were conducted by forwarding emails and announcing in classes. I would also like to thank Dr. Cliff Burgess, Rita and Tracy who have given me support for recruitment.

My time at SFU has been partially funded by Dr. Yue Wang as a research assistant, as well as the Dean of Graduate Studies and the Department of Linguistics at Simon Fraser University. I very much appreciate your financial support.

Lastly, but not at least, I would like to thank my beloved family. I've learned work ethics from my dad who has been working in Southeast Asia as an architect. Special thanks go to Mom for giving me mental support all the time to survive in Canada. I also thank my big brother who has given me lots of guidance since I was born.

A portion of this research will be presented at Interspeech 2012 in Portland.

Table of Contents

Approval ii
Partial Copyright Licence iii
Abstract iv
Dedication v
Acknowledgements vi
Table of Contents viii
List of Tables x
List of Figures xi

1. Introduction 1

2. Literature reviews 3
2.1. Audiovisual speech perception 3
2.1.1. Mechanism of audiovisual speech perception 3
2.1.2. Benefits of visual speech information in L1 4
2.1.3. Benefits of visual speech information in L2 5
2.1.4. Non-native speaker effect by native perceivers 7
2.2. Auditory and visual categorical development 8
2.2.1. Cross-linguistic differences 8
2.3. Second language speech learning 9
2.3.1. Theory of L2 speech acquisition 10
2.3.2. Factors affecting L2 speech acquisition 13
2.3.3. L2 Visual speech learning 15
2.3.4. Factors affecting L2 visual speech learning 15
2.4. Assessments of L2 productions 17
2.4.1. Auditory-based judgments 17
2.4.2. Factors affecting native listener judgments 18
2.5. Summary 19

3. Present study 20

4. Methodology 21
4.1. Speakers 21
4.2. Stimuli 22
4.3. Speaking tasks 22
4.4. Stimulus preparation 23
4.5. Perceivers 23
4.6. Perception task 24
4.6.1. Identification task 24
4.6.2. Goodness rating task 24

5. Results 26
5.1. Identification data 26
5.1.1. Identification across groups 27

5.1.1.1. AO condition 27
5.1.1.2. VO condition 29
5.1.1.3. AV condition 30
5.1.1.4. Summary of cross-linguistic comparison 33
5.1.2. Identification across modalities 33
5.1.2.1. Japanese speakers 33
5.1.2.2. English speakers 36
5.1.2.3. Summary 38
5.2. Visual effect size 38
5.3. Follow-up analysis: Lip-rounding effect on the perception of /ɹ/ 41
5.4. Patterns of perceptual confusion 43
5.5. Results of goodness rating task 47
5.5.1. AO condition 48
5.5.2. VO condition 49
5.5.3. AV condition 49
5.5.4. Cross-modality comparisons: Japanese-produced consonants 50
5.5.5. Summary of goodness-rating data 50

6. Discussion 53
6.1. General discussion 53
6.2. Cross-linguistic differences in the perception of audio and visual modalities 56
6.2.1. AO results 56
6.2.2. VO condition 57
6.2.3. AV condition 59
6.3. Effects of modality type on native judgments of L2 productions 59
6.4. Theoretical implications 61

7. Conclusion 64

References 67

Appendices 74
Appendix A: Speaker Background 75
Appendix B: Perception tasks 78

List of Tables

Table 1 The set of stimuli.....22

Table 2 Mean correct identification rates for six consonants produced by the Japanese and English speakers in AO, VO, and AV conditions. Standard deviations are given in parenthesis.....27

Table 3 Confusion matrix for consonant identification of Japanese productions with mean identification rates (%). Stars denote statistically significant differences.....46

Table 4 Mean goodness rating scores for six consonants produced by the Japanese and English speakers in AO, VO, and AV conditions. Standard deviations are given (Standard deviation in parenthesis).....48

List of Figures

Figure 1. Mean correct identification rates of each consonant given by English (ENG) and Japanese speakers (JP) in AO. Stars indicate statistically significant differences. 28

Figure 2. Mean correct identification rates of each consonant given by English (ENG) and Japanese speakers (JP) in VO. Stars indicate statistically significant differences. 30

Figure 3. Mean correct identification rates of each consonant given by English (ENG) and Japanese speakers (JP) in AV. Stars indicate statistically significant differences. 32

Figure 4. Mean correct identification rates of each consonant given by Japanese speakers in the three conditions (AV, AO and VO). 35

Figure 5. Mean correct identification rates of each consonant given by English speakers in the three conditions (AV, AO and VO). 37

Figure 6. Mean visual effect (VE) of /b, v/ (top) and /l, ɹ/ (bottom) given by English (ENG) and Japanese (JP) speakers. The brackets enclose +/- one standard error. 40

Figure 7. Correlation between mean visual intelligibility rates and degree of lip-rounding given by English (ENG) and Japanese (JP) speakers (1: no lip rounding, 5: full lip rounding). 42

Figure 8. Bar graphs showing mean rating scores of each consonant given by Japanese speakers in the three conditions (AV, AO and VO). The brackets enclose +/- one standard error..... 52

1. Introduction

Difficulties in producing second language (L2) speech sounds have been reported in previous studies. While numerous studies have addressed talker demographics to examine variance of L2 productions such as age of acquisition (e.g., Flege, Munro, & MacKay, 1995; Munro, Flege, & MacKay, 1996) and length of residence in an English speaking country (e.g., Larson-Hall, 2006), it is still essential to consider possible variances from listener factors which might cause issues such as listener-based assessment of L2 production. In order to assess the production at both segmental and supra-segmental levels, native listener judgements are commonly used to determine the successful production of phonemes such as /ɹ/ by Japanese learners of English (e.g., Larson-Hall, 2006) and judge the intelligibility of an utterance (e.g., Munro & Derwing, 1995). However, other studies have also revealed a talker-listener interaction in the perception of L2 productions with regards to perceptual flexibility from listener experience factors (e.g.Bent & Bradlow, 2003; Clarke & Garrett, 2004; Isaacs & Trofimovich, 2011; Pinet, Iverson & Huckvale, 2011; Sumner, 2011).

While variance in the perception of L2 productions from listener-related factors is inevitable, there is a possible factor which may affect L2 speech judgements: visual speech information. In face-to-face conversation, we use visual cues to process speech input, and it is widely known that speech perception is an integrated process implementing both auditory and visual information (McGurk & MacDonald, 1976). Visual speech information was reported as an effect of altering perceptual accuracy, as previous studies have revealed positive visual effects in both first language (L1) and L2 speech perception (e.g., Chen and Hazan, 2007 & 2009; Davis & Kim, 1999 & 2004; de Gelder, Bertelson, Vroomen & Chen, 1995; Fuster-Duran, 1996; Hazan, Sennema, Faulkner, Ortega-Llebaria, Iba & Chung, 2006; Hazan & Li, 2008; Hirata & Kelly, 2010; MacLeod & Summerfield, 1990; Nielsen, 2004; Sekiyama & Tohkura, 1993; Sumby & Pollack, 1954; Reisberg, McLean & Goldfield., 1987; Summerfield, 1983 & 1992; Wang, Behne & Jiang, 2009), and yet these studies were conducted to examine visual benefits

in the context of native perceivers perceiving L1 sounds or the naïve and experienced L2 perceivers perceiving L2 sounds. Thus, it remains to be determined to what extent additional visual cues influence the native perception of L2 productions.

The aim of this thesis is thus to investigate the effects of visual information on native judgements of L2 phonemes produced by Japanese learners of English. Native Canadian English listeners were presented with three English phonemic contrasts produced by native speakers of Japanese as well as native speakers of Canadian English as controls. The phonemes were /v, θ, l, ɹ/, which are not present in Japanese as well as /b, s/, which are shared by both Japanese and English consonant inventories. These stimuli were presented under audiovisual (AV), audio-only (AO), and visual-only (VO) conditions in order to address how listeners' use of modality affects perception. The findings may offer valuable insights into listener-based assessments of L2 production as well as understanding of speech integration processing on the perception of L2 speech.

2. Literature reviews

2.1. Audiovisual speech perception

2.1.1. Mechanism of audiovisual speech perception

It has been widely examined how visual information plays a significant role in speech perception. The McGurk effect (McGurk & Macdonald, 1976) is commonly used to demonstrate the process of auditory and visual signal integration by which perceivers perceive /da/ when visual information indicates /ga/ and auditory information indicates /ba/. As the binding of audio and visual signals occurs without perceiver intention, it has been claimed that audiovisual agreement is an automatic process occurring prior to attentional selection (McGurk & Macdonald, 1976; Massaro, 1987; Soto-Faraco Navarra & Alsius, 2004; Colin, Radeau, Soquet, Demolin, Colin, Deltenre, 2002; Colin, Radeau, Soquet, Deltenre, 2004). Research supported the view as both explicit (Massaro,1987) and implicit instructions (Soto-Faraco Navarra & Alsius, 2004) to pay attention to audio, visual, or both modalities did not affect the perceivers' responses to McGurk effect stimuli. On the other hand, some researchers challenged the attention-free account (Navarra, Alsius, Soto-Faraco & Spence, 2010a; Alsius, Navarra, Campbell & Soto-Faraco, 2005). Alsius et al. (2005) examined the effects of attention on the responses to the McGurk effect with a high-demand task in which perceivers were asked to respond to rapidly presented auditory and visual stimuli. It was found that visually influenced responses were significantly decreased when the task was delivered with concurrent auditory or visual tasks provided as a distraction. With an aurally or visually degraded condition, the McGurk effect was also used to examine whether perception would be affected by distraction. A stronger McGurk illusion was observed when the participants were instructed to look at a talker's face rather than to look at a distracter positioned on the face (Tippana, Anderson & Sams, 2004). Furthermore, it has also been shown that selective attention in one modality could prevent distractions from another modality

(Duncan, Sander & Ward, 1997). Thus, perceiver attention may play a significant role in audiovisual speech perception.

Furthermore, temporal alignments of auditory and visual information have been investigated to understand the binding of auditory and visual cues. Research found that visual information needs to precede auditory information in order to be perceived as a synchronized stimulus. In behavioral studies, participants were asked to identify whether the presented stimuli were synchronized or not, with the results showing that the subjective impression of audio and visual synchronization required the precedence of visual information over auditory information (Soto-Faraco & Alsius, 2007; Navarra, Alsius, Velasco, Soto-Faraco & Spence, 2010b). Other neurological studies showed the visual anticipatory effect (van Wassenhove, Grand & Poeppel, 2005; Stekelenburg & Vroomen, 2007; Navarra, Alsius, Soto-Faraco & Spence, 2010a). In fact, the processing of anticipatory visual speech information seems to shorten the latencies of N1 and P2 auditory evoked potential, indicating faster processing of auditory signals (van Wassenhove, Grand & Poeppel, 2005). While the visual information facilitates the neural processing of auditory information, the magnitude of the change in latencies is affected by the saliency of visual speech information. A further study also demonstrates sped-up processing, but only when the anticipatory visual speech exists (Stekelenburg & Vroomen, 2007). Navarra, Alsius, Soto-Faraco and Spence, (2010a) cited unpublished research by Vatakis and Spence suggesting that "the temporal aspects of the perception of audiovisual speech are modulated by the visibility of place of articulation (in consonants) and roundedness (in vowels), but not by the (less) visible manner of articulation or height" (85). In sum, visual speech information seems to precede auditory information, and auditory processing is facilitated by visual information only when the salient visual anticipatory effect exists.

In the following sections, how visual information is effectively used in both native (L1) and second language (L2) speech perception is detailed.

2.1.2. Benefits of visual speech information in L1

Visual information can help our perception in both regular and degraded conditions. Under ideal listening conditions, normal hearing populations can perceive

more accurately when visual cues were included in addition to auditory signals (e.g., Davis & Kim, 2004; Reisberg et al., 1987). In particular, visual information can facilitate the perception of consonants and vowels, resulting from place and manner of articulation cues (Summerfield, 1983). Summerfield (1992) claims that familiarity of words and the presence of clear onsets are necessary to increase the benefits of lip-reading. Visual information facilitates not only categorical perceptions, but also prosodic processing. For instance, more dynamic jaw movements were observed when speech amplitude increased (Eward, Beckman & Fletcher, 1991). As for duration of speech, articulatory configurations need to be maintained to lengthen the durations (de Jong, 1995). While visual cues are provided across facial regions, prosodic cues were carried greater in the upper face compared to the lower face (Swerts & Krahmer, 2008). While the present study examined the visual benefits in categorical perception, visual information seems to influence our perception on both segmental and prosodic levels.

Visual benefits in degraded environments, such as in noise, was reported (e.g., MacLeod & Summerfield, 1990; Nielsen, 2004; Sumby & Pollack, 1954). Additional visual cues enabled participants to tolerate additional noise although this can be influenced by a type of noise such as a level of four to six dB signal-to-noise (SNR) (MacLeod & Summerfield, 1990), and by approximately 16 dB (Sumby & Pollack, 1954). Accordingly, visual facilitative effects were reported in native language (L1) processing. The next section will reveal more visual benefits for the case of L2 perception.

2.1.3. Benefits of visual speech information in L2

Researchers have shown the benefits of visual information on L2 speech perception in various languages. In the perception of Korean, naïve English perceivers showed better detection of Korean speech when the stimuli were presented in the audio-visual (AV) condition compared to the audio-only (AO) condition (Davis & Kim, 1999). In this study, the English L1 participants were asked to determine whether a presented short Korean phrase included a target Korean sound. The results showed that more accurate detection was found when the following phrases were presented in AV compared to AO, indicating visual benefits for naïve L2 participants. Visual benefits were also observed among the L2 learners. The study examined the use of visual cues in the perception of non-native phonemic contrasts by Japanese, Spanish, and Korean

learners of English (Hazan, Sennema, Faulkner, Ortega-Llebaria, Iba & Chung, 2006). In their first experiment, effects of visual cues on /v-b-p/ phonemic contrasts were examined with Japanese and Spanish learners of English. Neither Japanese nor Spanish has the labiodental fricative /v/ in their inventory, though Spanish has the voiceless labiodental fricative, /f/. The target consonants /p, b, v/ were embedded in CV, VCV, or VC structures and were presented in audio only (AO), visual only (VO), and audiovisual (AV) conditions. The results showed while the Spanish participants showed better perceptual accuracy compared to the Japanese participants in the three conditions, both groups showed that the accuracy rate in AV was significantly higher than AO, and that both the AV and the AO were significantly higher than VO. In addition, for the Spanish group, visual identification did not show a significant decline compared to the auditory identification. However, this was not the case for the Japanese group. The authors suggested that the Spanish learners may have successfully associated visual cues in their native /f/ for the perception of non-native /v/.

In the second experiment, a less visually salient pair /l-ɹ/ was used to examine the perception of non-native phonemic contrasts by the Japanese and Korean learners of English. In Japanese, both /l/ and /ɹ/ are not existent, but lateral flap /ɾ/ which occurs only in word initial positions is. On the other hand, Korean has three types of representations [l, n, ɾ] for a liquid. The results found that the Korean learners showed significantly better identifications than the Japanese learners, especially in AO and AV. In addition, a significantly lower identification rate was observed in VO, though significant AV benefits observed in the /v-b-p/ contrast were not observed in the results of /l-ɹ/. Accordingly, the authors concluded that the learners' L1 and visual saliency may have affected the observed AV benefits.

Likewise, effects of learners' linguistic experience were investigated in terms of AV benefits of L2 speech perception. Wang, Behne and Jiang (2009) examined the influence of native language phonetic systems in AV perception by speakers of Korean, Mandarin and English. The results showed that both Korean and Mandarin perceivers made effective use of visual information to perceive difficult non-native consonants. Another interesting finding is that the Korean participants showed greater difficulty in perceiving /θ/ in VO although they could reach native-like perception in AO. The authors

suggest that the Korean participants may lack accurate use of visual cues in the L2 consonants due to reliance on highly-accurate acoustic cues.

While L2 audiovisual perceptions on segmental levels demonstrate AV benefits, visual cues in prosodic information also yield effects of AV training. Hirata and Kelly (2010) investigated effects of AV training on the perception of Japanese vowel length contrasts. Two types of AV training with regard to visual cues were given to the naïve English participants. Japanese has five short vowels /i, e, a, o, u/ as well as the corresponding long vowels /i ː , e ː , a ː , o ː , u ː /, whereas the durational contrasts are not existent in English, resulting in difficulties perceiving the vowel length distinction (Hirata, Whitehurst & Cullings, 2007). Part of the training includes auditory stimuli with visual speech information particularly focusing on lip-movements from a talker's face, and hand gestures corresponding to the vowel length contrasts were added to the other AV training. After the training, significant improvements of the phonemic contrasts were demonstrated especially among the trainees who focused on cues from the integration of audio and lip-movements. The hand-gesture group did not significantly outperform the auditory-trainee control group after training.

2.1.4. Non-native speaker effect by native perceivers

Besides visual benefits in the perceptions of L1 and L2, facilitative visual effects were shown in the native perception of L2 production, which is our particular focus in this study. Research has shown that non-native stimuli induce further visual reliance in audiovisual speech perception (Dutch & Cantonese in de Gelder, Bertelson, Vroomen & Chen, 1995; German & Spanish in Fuster-Duran, 1996; Australian and Hungarian in Grassegger, 1995; Japanese in Sekiyama & Tohkura, 1993). For instance, Japanese participants showed greater McGurk illusion effects on the perception of English stimuli compared to Japanese stimuli, indicating that Japanese participants showed a greater visual reliance on non-native speech perception. This phenomenon is known as the "foreign language effect" (Sekiyama and Tokuhira 1993) or "non-native speaker effect". In Chen and Hazan (2007), Thai and Japanese perceivers gave more visually influenced responses when they perceived Mandarin and English McGurk stimuli. The speakers of English also showed the non-native speaker effect as the English participants showed a greater visual reliance on Mandarin tokens compared to English tokens (Hazan & Li,

2008). They suggested that participants may have a strategy to use less auditory information for non-native talkers when neither the auditory and visual information is perfectly intact. They also claimed that this lessened auditory reliance occurred when the syllables produced by non-native talkers existed in the perceivers' L1 phonemic inventory. Chen and Hazan (2009) also investigated the non-native speaker effect by comparing the Mandarin and English of adults and children. By looking at the results showing a high correlation between visual effects and mouth-reading performance, it was also shown that reliance on visual cues by the adults was significantly higher than for the children as adults had a better mouth-reading ability compared to children.

Thus, research revealed that visual information yielded benefits of enhancing perceptual accuracy as well as inducing further visual reliance in the perceptions of L1 and L2. The AV benefits were also observed in both segmental and prosodic perceptions. Furthermore, the native perceivers demonstrated the "non-native speaker effect" in which visual reliance in non-native speech was enhanced.

Research of cross-linguistic differences in reliance of auditory and visual cues will be reviewed in Section 2.2. Later sections will cover research examining L2 auditory and visual categorical developments.

2.2. Auditory and visual categorical development

2.2.1. Cross-linguistic differences

In cross-linguistic comparisons, some perceiver groups, such as English, Finnish and French demonstrated McGurk illusions (Massaro, 1998) while other groups, such as Japanese and Mandarin showed a weaker effect (Sekiyama, 1997; Sekiyama and Tokuhira, 1993). This cross-linguistic difference raised a question in terms of perceivers' developmental factors. Sekiyama and Burham (2008) examined the developmental onset of inter-language differences in audio and visual weightings between speakers of Japanese and English. While Japanese speakers demonstrated less reliance on visual information, it was not clear whether the inter-language differences in the weightings were nature or nurture. In their first experiment, adult speakers of Japanese and English were compared to see their visual and auditory reliance in the McGurk illusion as well as

their reaction time for the stimulus responses. The findings also supported less visual reliance by the speakers of Japanese compared to the speakers of English in AV although the cross-linguistic differences in response accuracy was not observed in unimodal conditions (i.e., AO and VO). In terms of group differences in reaction time, the English perceivers showed shorter reaction times than the Japanese perceivers in VO whereas the Japanese perceivers showed faster responses in AO. Within each speaker group, Japanese speakers demonstrated faster responses in AO compared to AV whereas the English perceivers showed faster responses in AV compared to AO. This increased reliance on and faster responses to the visual information could be supported by the visual anticipatory effect as detailed in the previous section, indicating that the adult Japanese perceivers may demonstrate less visual anticipatory effect whereas the English perceivers may efficiently process auditory and visual information as the visual anticipatory effect proposes. In their second experiment, they examined cross-linguistic developmental effects by comparing three age groups: age 6, 8 and 11. While the English children also demonstrated higher visual reliance compared to the Japanese children, the increase in the degree of visual influence was observed between age 6 and 8 only by the English perceivers in both response accuracy and reaction time data. In terms of the auditory reliance, the age 6 group of Japanese children demonstrated superiority of auditory reliance, but not at the latter ages. Thus, they conclude that the earlier availability and the use of visual information may contribute to better audiovisual integration by the English perceivers whereas the earlier availability and the use of auditory information may contribute to relatively less efficient audiovisual integration by the Japanese perceivers.

2.3. Second language speech learning

The present study examines the native judgments of L2 consonants produced by Japanese learners of English. The following sections summarize theories of L2 speech acquisition as well as relevant claims of L2 visual speech learning.

2.3.1. Theory of L2 speech acquisition

With regard to theories of second language learning, it is widely known that second language learners face difficulties due to the interaction of their first language (L1) and second language (L2) inventories. The speech learning model (SLM: Flege, 1995; 2007) and the perceptual assimilation models (PAM: Best, 1995; PAM-L2: Best & Tyler, 2007) are the two major models for cross-linguistic categorical perception. PAM (Best, 1995) posits that L2 sounds tend to be perceived according to the similarities and dissimilarities to the closest gestural constellations in native phonological space. The phonetic input and the perceivers' native phonological space are defined by the spacial layout of the vocal tract (i.e., articulatory gestures). Thus, similarities of native and non-native segments are assessed by the similarities in the gestural constellations, and predict perceptual assimilation of the non-native phonemes to the native categories. The following cases were introduced regarding patterns of perceptual assimilation of non-native phonemes. The first case is when a non-native phoneme is assimilated to a native category, which results in either a good, acceptable or deviant examplar. The second case is when a non-native phoneme is assimilated as an uncategorized speech sound, namely occurring in a native phonological space but not being categorized in any particular native category. The third case is when a non-native sound is not assimilated to a native category. Within the cases, five types of categorizations are described. Two-category assimilation (TC type) is when each non-native segment is assimilated to a different native category. Category-goodness difference type (CG type) is when both non-native sounds are assimilated to the same native category, but differ regarding the degree of discrepancy from the native segment. When the degree of discrepancy is equal, this belongs to single-category assimilation (SC type). In addition, when both non-native sounds fall outside of any native category within the phonetic space, this refers to UU type (both uncategorizable). When one non-native sound falls outside of a native category whereas the other assimilated to a native category, this belongs to UC type (uncategorized versus cateogorized). The last case is NA type (nonassimilable) in which both non-native sounds are perceived as non-speech sounds.

The speech learning model (SLM: Flege, 1995; 2007) also accounts for how issues in L2 phonemic contrasts occur. One of the assumptions of speech learning in this theory is that phonetic categories of both L1 and L2 exist in a common phonological

space. The L1 and L2 sounds are accordingly perceptually related to one another (Hypothesis 1). When new L2 sounds are acquired, new phonetic categories can be established by discerning phonetic differences from the closest L1 (Hypothesis 2). The phonetic differences are more likely to be discerned when the phonetic dissimilarities is bigger (Hypothesis 3). Furthermore, it tends to fail to form a new L2 category by an equivalent classification (Hypothesis 5). As for relations of speech perception and production, it is hypothesized that production of L2 sounds correspondents the phonetic representation as accurate perceptual targets guide to L2 learning, (Hypothesis 7).

While both SLM and PAM account for L2 speech learning and possible issues in non-native speech acquisition, significant differences in the two theories exist. Firstly, their target of L2 learners differs. SLM targets experienced L2 listeners, namely bilingual speakers studying L2 for many years in an English speaking country. Conversely, the PAM targets naïve listeners of L2 sounds. Furthermore, the SLM posits that L2 learners perceive acoustic features from their input in order to form new sound categories whereas perception in the PAM (Best 1995) and PAM-L2 (Best & Tyler, 2007) are on the basis of articulatory phonology in which L2 learners are exposed to new articulatory gestures of L2 sounds which leads to acquisition. Flege (2003) pointed out the differences in the SLM and PAM regarding how L1 and L2 sounds interact. PAM posits that L2 speech sounds are perceived according to *the similarities and discrepancies* from native instances. On the other hand, SLM takes into account phonetic distances between L1 and L2 sounds. The SLM posits that new phonetic categories of L2 sounds are more likely to be established for L2 sounds that are perceived to be distant from the closest L1 sound. The speech learning process also differs. SLM explains that the speech learning process occurs with passive reception of meaningless acoustic features whereas the speech learning in PAM occurs by actively seeking meaningful distal events (i.e., articulatory gestural information).

These theories were examined with various L2 categorical acquisitions. As the present research targets speech productions of Japanese learners, categorical assimilations by Japanese learners of English are mainly described in this section. One of the most common difficulties in non-native contrasts is the English /l-ɹ/ distinction by adult Japanese speakers, which has been extensively examined by numerous scholars.

Adult Japanese learners of English have difficulty in distinguishing /l/ and /ɹ/ (Miyawaki, Jenkins, Strange, Liberman, Verbrugge & Fujimura, 1975; MacKain, Best, & Strange, 1981; Mochizuki, 1981) since Japanese speakers tend to assimilate the /l/ and /ɹ/, which do not exist in Japanese, to the Japanese lateral flap /ɾ/. According to PAM, both English /l/ and /ɹ/ are perceived as 'poor' examplars of Japanese /ɾ/ (Best & Strange, 1992), being consistent with following perception studies in which Japanese listeners rated English /l/ and /ɹ/ as equally as far from Japanese /ɾ/ (Iverson, Kuhl, Akahane-Yamada, Diesch, Tohkura, Kettermann & Siebert, 2003). While PAM predicts equal difficulties in the acquisition of both English /l/ and /ɹ/ by Japanese speakers, research has found that adult Japanese listeners were more likely to categorize /l/ as Japanese /ɾ/ (Takagi, 1993, Aoyama, Flege, Guion, Akahane-Yamada & Yamada, 2004).

On the other hand, Flege's (1995) SLM predicts that native Japanese adults are more like to establish English /ɹ/ compared to /l/, and will be more likely to produce English /ɹ/ accurately as Japanese lateral flap is perceptually closer to English /l/. Flege, Takagi and Mann (1996) found that /ɹ/ was more correctly identified than /l/ by both experienced and inexperienced Japanese listeners of English, which is congruent with the SLM hypotheses. Aoyama et al. (2004) found that both of their child and adult Japanese speakers produced significantly better /l/ compared to /ɹ/. These findings oppose the SLM prediction, suggesting an easier categorical formation of English /ɹ/. However their study also found that the Japanese children showed greater improvement for /ɹ/ in both production and perception over approximately a year's length of residence in the US, which in turn supports the SLM hypotheses.

In the current study, L2 phonemes produced by adult Japanese speakers will be examined. Including SLM and PAM, which mainly argue theories of speech perception, there is an on-going question regarding the relationship between speech perception and production. One account is a unitary system of speech perception and production suggesting that phonetic contrasts L2 learners produce by the vocal tract are identical to acoustic contrasts in their perceptual analyses (e.g., Liberman and Mattingly, 1989; Pisoni, 1997; Stevens, 1972). This claim is congruent with the SLM (1995) predicting that correct production of L2 sounds represents successful perceptual phonetic

categorical formation. Another approach takes into account direct realism suggesting that objects of speech perception are the articulatory gestures and there is a direct link between perception and production which PAM posits. On the other hand, other researchers argue that the there is an indirect relationship between speech perception and production. Namely, acoustic signals are objects of speech perception and not mediated by articulatory representations (Stevens and Blumstein, 1981).

Accordingly, empirical research has been conducted to examine the relationship between speech perception and production. Research suggests that auditory-articulatory unitary relations are not always found. Hattori and Iverson (2009) found that categorical assimilation of /l- ɹ/ were unpredictable between perception and production, with /l/ being more likely to assimilate Japanese /ɾ/. Other studies also found a similar asymmetry such as the Japanese speakers performing better at perception compared to production (e.g., Goto, 1971).

In sum, while theories of speech perception extend to discuss theories of speech production, some limitations have been found. Firstly, the relationship between speech perception and production is not clear. While theories of L2 speech learning (i.e., SLM and PAM) are extensively tested especially in speech perception, further studies are required to understand to what extent speech perception is related to production. It is important to address this relationship as it would show how speech learning in L2 learners' production occurs. Furthermore, the interaction of L1 and L2 sounds is complex even only in speech perception, and non-native contrasts are not extensively examined with other non-native contrasts. While this section mainly introduced issues in English /l-ɹ/ contrasts by Japanese speakers, a small number of studies examined the categorical perception and production beyond a pair of contrast especially in consonants (e.g., Guion et al., 2000).

2.3.2. Factors affecting L2 speech acquisition

L2 auditory speech acquisition is affected by various other factors beyond the effect of L1. For instance, age was examined with regard to the age of acquisition (e.g., Flege, Munro, & MacKay, 1995; Munro, Flege, & MacKay, 1996), and age of learning (e.g., Aoyama et al., 2004). Aoyama et al. (2004) showed that while adult Japanese

learners showed an initial advantage in categorical discrimination of /l-ɹ/ compared to the child Japanese participants (average age of 10.5 years), their results approximately a year later showed significant improvement among the children compared to the adults. Other research has also replicated the initial advantage by adult L2 listeners. Research has also examined the child-adult differences with early-arrival Korean listeners of English vowel perceptions and productions (Baker, Trofimovich, Flege, Mack & Halter, 2008). The Korean adult and child participants were asked to classify the English vowels with the Korean vowels, and the children mapped the English vowels less frequently onto their respective Korean vowels compared to the adults. The results of the identification task replicated an initial advantage by adults showing that the Korean adults also displayed better identification accuracy compared to the Korean children. However, in a production task the Korean children outperformed adults on productions of non-native English vowels /I, ʊ/.

Length of residence in an English speaking country is also another main factor, and has been extensively examined with regard to the native-like attainment for those who had extensive LOR (e.g., Flege, Takagi, & Mann, 1996, Larson-Hall, 2006; Saito & Brajot, in press). Although the findings were not completely consistent, longer LOR resulted in development of L2 speech acquisition. Saito and Brajot (in press) examined the effects of LOR and AOA on the development of acoustic cues in the production of /ɹ/ by speakers of Japanese. Native Japanese speakers who differed in their LOR and AOA were asked to produce /ɹ/ elicited from word- and sentence-reading tasks as well as picture description tasks aimed at collecting their spontaneous speech. The results demonstrated attainment of a native-like range of English /ɹ/ with regards to minor acoustic cues, namely F1, F2, and the transition duration, within approximately one year. A higher correlation of F3 and LOR was found, particularly in spontaneous speech, indicating that longer LOR results in the acquisition of a major acoustic cue.

The experience of L2 learning itself also affects acquisition (Guion, Flege, Akahane-Yamada & Pruitt, 2000), and training studies demonstrated influences such as an effect on learners' L1. Morrison's longitudinal study (2002) examined the effect of linguistic experience on learning tense and lax vowel distinctions by Japanese and Spanish learners of English. While both learner groups do not possess the quality

difference in their vowel inventories, Japanese has the quality difference in vowels such as /i-iː/ and /u-uː/. After six months of English exposure in an English speaking country, the Spanish learners were more successfully able to acquire the quality difference compared to the Japanese learners indicating the consistent existence of wrong acoustic cues influenced by a L1.

2.3.3. L2 Visual speech learning

While L2 visual learning theories have not been established, Hazan et al. (2006) introduced three types of visual speech categories in the perception of L2 speech. The first case is when a visual cue is present in both the L1 and L2. For instance, alveolar fricative /s/ exists in both Mandarin and English. The second case is when a visual category occurs in the L2 but not in the L1 such as labiodental fricatives /f, v/ which do not exist in Japanese. The last case is when a visual category occurs in both L1 and L2, but is used for a different phonetic distinction. In Spanish, voiced labiodental fricative /v/ does not exist although the voiceless labiodental fricative /f/ occurs.

While theories such as SLM and PAM consider auditory L2 learning, the question remains as to how learners process visual cues and learn them for new L2 sounds. Given that L2 learners acquire new L2 sounds through exposure to new articulatory gestures according to PAM and PAM-L2, it is highly possible that L2 learners may also acquire visual articulatory cues as well as auditory cues in their perception. Nonetheless, further research is required.

2.3.4. Factors affecting L2 visual speech learning

As L2 auditory speech learning is affected by various factors such as first language (e.g., Flege, Munro, & MacKay, 1995; Munro, Flege, & MacKay, 1996), age of acquisition, and length of residence in an English speaking country (e.g., Larson-Hall, 2006), L2 visual speech learning is also affected by such factors. Research has shown that a longer LOR (length of residence) in an English-speaking country alters learners' visual categories (Wang, Behne & Jiang, 2008), and a certain correlation of auditory and visual categorical development may exist in L2 speech acquisition. Wang et al. (2008) examined the effects of L2 learners' length of residence (LOR) in Canada on audiovisual

speech perception. Native English as a control as well as two Mandarin groups were compared: a group of participants who had lived in Canada an average of 10 years (called the long LOR group) and another group of participants who had lived in Canada an average of 2 years (called the short LOR group). Significant group differences were observed in the non-native consonant /θ/ in which the long LOR group achieved native-like performance in AO perception and AV+ perception. Namely, correct identification rates for native English and the long LOR Mandarin perceivers were higher than those of the short LOR Mandarin perceivers. However, there were no group differences in VO. In AV incongruent perception (A /f/ + V /s/; A /s/ + V /f/), auditory-reliant responses were observed among the native English and the long LOR groups and visually-reliant responses were observed among the short LOR group. This study suggests that developments of visual and auditory categories may not be aligned as the results showed Mandarin learners' LOR affected only auditory perception but not their visual perception. In addition, a higher visual reliance in AV was found only among the short LOR Mandarin perceivers although presented auditory stimuli included native phonemes (i.e., /f, s/). In addition, Mandarin participants would not necessarily be required to rely on visual cues to respond, as the stimuli were not degraded.

Furthermore, visual saliency in non-native phonemes also affects learning as shown in L2 categorical perception. Hazan, Sennema, Iba & Faulkner (2005) investigated whether Japanese learners of English can be trained with the use of auditory and visual information for English /v/-/b/-p/ and /l/-/ɹ/ phonemic contrasts, comparing training with auditory information only to audiovisual information. The results of this study indicated that AV training was more effective on the contrast between English bilabial and labiodental consonants (/b/-/v/) whereas AO training was more effective on the contrast of /l/-/ɹ/, showing that visual information was effectively used for a visually salient phonemic contrast. Thus, while perception among speakers of Japanese is relatively less reliant on visual information and integration of auditory and visual cues, they are still able to use the cues to learn non-native phonemes.

2.4. Assessments of L2 productions

While the previous sections review research showing visual benefits in speech perception in L1 and L2, native listeners are often recruited as judges to determine the accuracy of L2 productions. The next sections reveal possible variance in listener-based judgments and raise the question of the auditory-based assessment considering the effective use of visual information in speech perception.

2.4.1. Auditory-based judgments

Auditory-based assessment has been conducted in order to measure L2 productions. In a review cited from *Phonology and Second Language Acquisition*, Munro (2007) introduced three ways of measuring L2 speech: "(1) responses from unsophisticated listeners' of the assessment, (2) impressionistic analyses from expert evaluators, and (3) acoustic phonetic analyses" (199-200). The first type refers to impressionistic assessment in which phonetically untrained or naïve listeners are asked to identify or rate speech samples. The second type indicates when phonetically-trained or expert listeners judge specific phonetic features with their expert knowledge, such as counting segmental errors. The third approach requires acoustic measurement with software that contains a formant measurement tool such as Praat or Wavesurfer. While there was a significant correlation between the assessment of native ESL instructors and linguistic analyses in both segmental and prosodic features (Anderson-Hsieh, Johnson & Koehler,1992), Munro (2007) emphasizes the importance of untrained listener assessment to maximize the external validity in which most of our cross-linguistic conversation is exchanged without meta-linguistic knowledge. In fact, listener-based assessment has been widely conducted. For measurements of L2 English productions, native English listeners were asked to identify L2 productions such as English vowels (e.g., Munro & Dewing, 2008). Furthermore, rating tasks have been frequently used to measure how good L2 productions were (e.g., Larson-Hall, 2006) and how accented L2 productions were (e.g., Flege, Munro, & MacKay, 1995; Munro, Flege, & MacKay, 1996).

2.4.2. Factors affecting native listener judgments

While research using native listener judgments is prolific, as we saw in the previous section, a significant listener variance in L2 assessment may exist as Munro (2007) also pointed out in the review. For instance, familiarity of non-native speech was shown to facilitate the native listeners' comprehensibility (e.g., Gass & Varonis, 1984). The listener music background also affected the L2 speech judgments in terms of accentedness (Isaacs & Trofimovich, 2011). Shared talker and listener experiences have been found to influence perception. For instance, while L1 English listeners performed better in listening to native speech, L2 listeners showed high comprehension accuracy in the speech of high-proficiency L2 talkers from the same L1 and different L1 as equal to the native talkers (Bent & Bradlow, 2003). Since the L2 listeners perceived accented speech accurately for the talkers who have both shared and different phonological systems, it may be less likely that listeners benefit by only having a shared phonological system. Another research has shown that common phonological systems between the speakers and listeners did not appear to benefit the perception of L2 speech, and yet the phonetic details in the productions of the talkers and the listeners showed a correlation in their accurate perception (Pinet, Iverson & Huckvale, 2011). Even short, intensive exposure may alter the perception system. Perceptual adaptation studies demonstrate robust perceptual adaptation to foreign-accented speech (Clarke & Garrett, 2004) as well as a categorical shift by trained listeners to accommodate VOT variations (Sumner, 2011).

While listener-related factors were mainly considered, it is also possible that their judgment can be affected beyond their experience factors. As the previous section revealed, visual information plays a significant role in speech perception, and visual speech information may indeed affect native listener judgments. No study has yet to address the potential contribution of visual influences on native listener assessment, and yet, it is important to consider the effects of visual information in order to understand how people understand L2 speech in face-to-face conversation.

2.5. Summary

Listener-based assessment of L2 production has been extensively conducted. While research has contributed to our understanding of how people perceive L2 speech, it is important to consider the effects of visual information on listener-based judgments. Indeed, this consideration may help understand how people perceive L2 productions in face-to-face conversations. Furthermore, it is highly possible that L2 speakers' articulatory constellations may contribute to different judgements beyond their auditory information in both positive and negative ways. In particular, when there is a gap between auditory representation and visual representation of the target phonemes, how will the judges be affected? These questions are important to address since incorrect judgments of L2 productions may affect assessment of L2 speech as well as speech training for L2 learners.

3. Present study

The present study examines the effect of visual information on the intelligibility of L2 consonant production. Canadian English listeners were asked to identify the initial consonants /b, v, s, θ, l, ɹ/ in English CV syllables produced by Japanese learners of English. Since Japanese does not have /v/ and /θ/, but has /b/ and /s/, Japanese learners of English tend to replace the nonexistent phonemes with similar counterparts existing in their L1 (i.e., /v/ with /b/; /θ/ with /s/) (Yoshida & Hirasaka, 1983). In the context of PAM (1995), the /b/-/v/ and /s/-/θ/ contrasts are examples of the "Uncatogorized versus Catogorized" (UC) type. Furthermore, Japanese speakers tend to assimilate the /l/ and /ɹ/, which do not exist in Japanese. This PAM category type is "both Uncategorizable (UU)" type. However Japanese has a lateral flap /ɾ/, with /l/ being more likely to be categorized as Japanese /ɾ/ (e.g., Takagi, 1993). Accordingly, Japanese speakers in our study may also display difficulties in producing non-native phonemes /v, θ, l, ɹ/ due to the interaction of L1 and L2 inventories.

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 non-native speech (Hazan et al., 2010, Hazan et al., 2008, Sekiyama & Tohkura, 1993) and visual facilitative effects among the non-native consonants occurring in both L1 and L2 (Hazan et al., 2006), we hypothesize that native listeners 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/). On the other hand, the non-Japanese consonants /v, θ, l, ɹ/ could be expected to show negative visual effects on intelligibility rates in the AV conditions due to assimilations to L1 counterparts for /v, θ/ and /l, ɹ/ when the speakers produce incorrect articulatory configurations.

4. Methodology

4.1. Speakers

Fifteen native speakers of Japanese (eight female, seven male) participated in this study. The age range was from 19 to 31 (Mean age=25). The Japanese speakers had been living in Vancouver, Canada, as students or workers for an average of 10.2 months, and were from various cities in Japan. None of them had lived in an English speaking country before arriving in Canada. We selected relatively new arrivals as LOR may have affected productions (c.f., Flege, Takagi, & Mann, 1996, Larson-Hall, 2006; Saito & Brajot, in press). In addition, none of them used English at home or had lived in an English speaking country before arriving in Canada. The speakers of Japanese started to learn English as a foreign language during middle school (at the age of 12), when they were initially exposed to English. According to self-reports, their daily English input varied (mean daily input of English=51%; mean daily input of Japanese=49%), and yet none of them had reached an advanced level at the time of testing.

An additional 15 native speakers of Canadian English (eight female, seven male) also participated as a control group. The age range was from 18 to 31 (Mean age=22). Since this study uses visual information through the speakers' full face for the identification task, ethnic information from their facial features may crucially affect perceiver judgments. Thus, in order to exclude speaker' ethnicity factors between the speakers of Japanese and Canadian English, only English speakers of East-Asian descent (e.g., Japan, China, Korea) were recruited. All of them were born in Canada and English was their first language. All the speakers reported that they did not have any speech impairments. They were compensated for their participation. The details of speaker language backgrounds appear in Appendix B.

4.2. Stimuli

Six English CV syllables, having the initial consonant /b, v, s, θ, l, ɹ/ followed by the vowel /a/ were used as stimuli. Since Japanese does not have /v/ and /θ/, but has /b/ and /s/, Japanese learners of English are likely to substitute the nonexistent phonemes with the similar counterparts existing in their L1 (i.e., /v/ with /b/; /θ/ with /s/) (Yoshida & Hirasaka, 1983). In addition, Japanese speakers tend to assimilate the /l/ and /ɹ/, which do not exist in Japanese, to the Japanese lateral flap /ɾ/. Studies found that /l/ was more likely to be categorized as Japanese /ɾ/ (Takagi, 1993). Hereby, /b/ and /s/ are referred to as 'native phonemes', and /v/, /θ/, /l/, and /ɹ/ referred to as 'non-native phonemes.'

Table 1 ***The set of stimuli***

	/b/	/v/	/s/	/θ/	/l/	/ɹ/
POA	Bilabial	Labiodental	Alveolar	Interdental	Alveolar	Alveolar
Stimuli	/ba/	/va/	/sa/	/θa/	/la/	/ɹa/
Occurrence	ENG JP	ENG	ENG JP	ENG	ENG	ENG

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

4.3. Speaking tasks

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. 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. The speakers were asked to read the stimulus set with five repetitions, and the best examples were chosen considering noise or 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 low quality sound resolution, a separate audio recording was made with a Shure KSM 109 condenser microphone to 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 monotonous 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.

4.4. Stimulus preparation

After collecting both audio and video files separately, three sets of perception tasks (AV, AO and VO) were prepared. First, the intensity of the audio files collected with a Shure KSM 109 condenser microphone was normalized using SoundForge 6.4 to have the same unweighted RMS value. The normalized audio files were synchronized with the video-recording files using Final Cut Pro X, then the audio files collected by the camcorder were deleted. The frame length was 3000ms containing 1.2 seconds of a neutral face before and after the stimulus. Thus, each stimulus captured the mouth opening as well as closing. The resolution was 640x480 pixels. 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.

4.5. Perceivers

Thirty-one native speakers of Canadian English (sixteen male, fifteen female) participated as perceivers 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 with Japanese learning and all of them had a limited amount of daily exposure to Japanese-accented English. All the listeners reported that they had normal hearing and normal or corrected vision. They were compensated for their participation. Due to file-loading issues in E-prime (Psychology Software Tools), data from two male listeners were excluded.

4.6. Perception task

The perceivers were tested individually in a sound-treated room in the Language and Brain Lab and asked to wear 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 the following vowel /a/. Prior to each session, they received instructions on the screen. They looked at the articulatory movements on 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 (Psychology Software Tools) was used to run the perception tasks. The order of the modality condition (AV, AO, VO) was randomized across the participants. Overall, the whole experiment lasted two and a half hours, which divided across the three visits with 5- to 10-minute breaks within each session.

The following sections explain the details of the two tasks, an identification task (Section 4.6.1.) and a goodness rating task (Section 4.6.2.).

4.6.1. Identification task

In the identification task, the perceivers were asked to identify presented stimulus consonants by pressing a button on a keyboard. 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 perceivers perceived something other than the presented response alternatives. The perceivers were allowed a maximum of 4 seconds in each trial to indicate their response.

4.6.2. Goodness rating task

After completing the identification tasks in AO, VO and AV, the perceivers proceeded to take the goodness rating task, in which they were asked to rate each speaker's pronunciation of the initial consonant of the stimulus syllable on a scale of one to five (i.e., 1: Poor, 5: Excellent). They were encouraged to use the full 5-point scale, focusing their rating exclusively on the pronunciation of the initial consonant. The intended initial consonant appeared for 1000ms on the screen with fixation cross,

followed by the stimulus in order to avoid mismatch between judgments and the actual stimuli. A five point scale with the stimulus consonant was then presented as a response screen, which lasted a maximum of four seconds. The stimuli for the goodness rating tasks were delivered using a randomized order in each of the three conditions (AV, AO, and VO). Again, the participants completed the goodness rating tasks in AV, AO, and VO.

5. Results

5.1. Identification data

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

Further analyses were conducted to explore these effects. 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.

Table 2 ***Mean correct identification rates for six consonants produced by the Japanese and English speakers in AO, VO, and AV conditions. Standard deviations are given in parenthesis.***

	AO		VO		AV	
	Japanese	English	Japanese	English	Japanese	English
/b/	91.2 (10.6)	94.9 (7.3)	93.3 (9.3)	92.6 (9.7)	97.7 (3.7)	99.1 (2.4)
/v/	39.3 (13.0)	92.2 (7.8)	51.3 (13.9)	89.0 (18.0)	52.4 (9.4)	98.6 (2.8)
/s/	93.1 (8.1)	99.1 (2.4)	74.1 (24.6)	76.0 (24.2)	98.2 (4.7)	98.6 (3.7)
/θ/	56.8 (14.4)	88.4 (11.6)	59.7 (14.8)	95.6 (7.4)	63.9 (11.2)	98.4 (3.4)
/l/	88.4 (16.0)	98.6 (2.8)	77.5 (15.5)	59.3 (18.0)	92.6 (11.9)	98.8. (3.1)
/ɹ/	55.2 (12.1)	99.5 (1.7)	34.9 (13.9)	84.1 (15.4)	51.6 (11.3)	99.5 (1.7)

5.1.1. Identification across groups

5.1.1.1. AO condition

The mean correct identification rates in AO condition were submitted to a repeated measures ANOVA with speaker language group (Japanese and English) and consonant type (i.e., /b, v, s, θ, l, ɹ/) as within-group factors. Figure 1 presents mean correct identification rates of each consonant given by English and Japanese speakers in the AO condition. Significant main effects were found for the speaker language [F(1, 28)=1582.521, p< .001, partial η^2=.983] (Mean JP=70.7%, Eng =95.5%) and consonant types [F(5, 140)=74.316, p<.001, partial η^2=.726] (Mean intelligibility rates of six consonants are as follows: /b/= 93.1%, /v/=65.7%, /s/= 96.1%, /θ/=72.6%, /l/= 93.5%, /ɹ/= 77.3%). The interaction between speaker language and consonant type was also statistically significant [F(5, 140)=86.094, p<.001, partial η^2 = .755]. Bonferroni post hoc tests revealed that while overall differences in the mean correct identification rates between the productions of 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 less intelligible consonants, /v/ (39.3%) was significantly less intelligible than /θ/ (56.8%) and /ɹ/ (55.2%) (p<.005). On the other hand, the listeners also showed differing perception in the six consonants produced by the Canadian

English speakers. The perception of /θ/ (88.4%) was significantly lower than all the other consonants (p<.05). That of /v/ (92.2%) was significantly lower than /l/ (98.6 %), /s/ (99.1%), and /ɹ/ (99.5%) (p<.01). There was also a significant difference between /ɹ/ and /b/ (p=.043). There were no other significant differences.

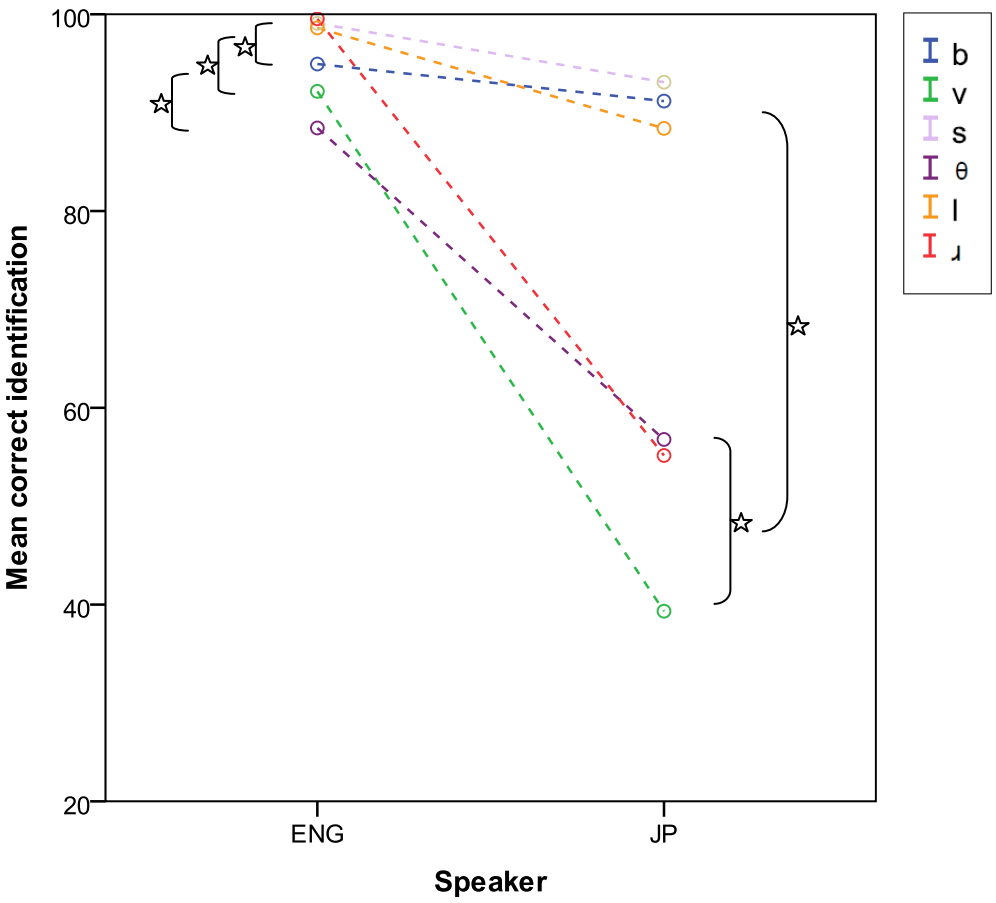


Figure 1. ***Mean correct identification rates of each consonant given by English (ENG) and Japanese speakers (JP) in AO. Stars indicate statistically significant differences.***

5.1.1.2. VO condition

In the VO condition the mean intelligibility rates were submitted to a repeated measures ANOVA with speaker language group (Japanese and English), and consonant type (i.e., /b, v, s, θ, l, ɹ/) as within-group factors. Figure 2 is a line graph representing the summary. The main effects of speaker language [$F(1, 28)=296.690$, $p<.001$, partial $\eta^2=.914$] (Mean JP= 65.1%, ENG= 82.8%) and consonant type [$F(5, 140)=20.175$, $p<.001$] were observed (/b/= 93.0%, /v/=70.1%, /s/= 75.1%, /θ/=77.6%, /l/= 68.4%, /ɹ/= 59.5%) in the VO condition, along with a significant interaction [$F(5, 140)=124.172$, $p<.001$, partial $\eta^2=.791$]. Bonferroni post hoc tests revealed that visual perception differed as a function of speaker language based on the consonant type. While /b/ ($p=.692$) and /s/ ($p=.365$) did not show a significant difference between the speakers of Japanese and English, the perception of /v, θ, ɹ/ produced by the native English speakers was significantly better than those of Japanese speakers ($p<.001$). On the other hand, the perception of /l/ produced by the Japanese speakers (77.5%) was significantly more accurate than those produced by the English speakers (59.3%) ($p<.001$). Within the Japanese productions, /b/ was significantly more intelligible than any other consonant ($p<.01$). Furthermore, the Japanese-produced /l/ (77.5%) and /s/ (74.1%) were more intelligible than /θ/ (59.7%), /v/ (51.3%), and /ɹ/(34.9) ($p < .01$). While /θ, v/ did not show any significant differences ($p > .05$), these two were more intelligible than /ɹ/ ($p<.01$). /ɹ/ was significantly perceived as least intelligible compared to any other Japanese productions ($p<.001$). Among the English productions, /θ/ (95.6%), /b/ (92.6%), and /v/ (89.0%) were more intelligible than /ɹ/ (84.1%), /s/ (76.0%) and /l/ (59.3%)($p < .05$). In addition, /ɹ/ was significantly better perceived than /l/ ($p<.001$). No other difference was observed.

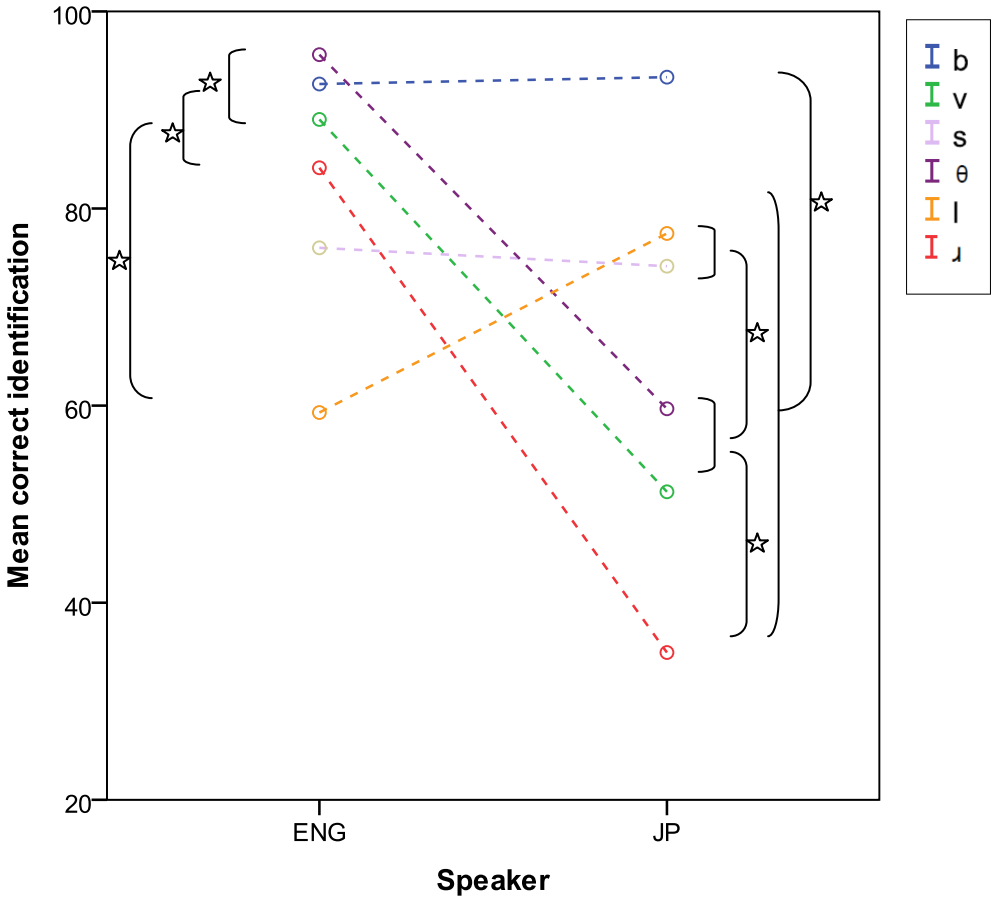


Figure 2. ***Mean correct identification rates of each consonant given by English (ENG) and Japanese speakers (JP) in VO. Stars indicate statistically significant differences.***

5.1.1.3. AV condition

In order to examine how the native English perceivers perceived the English consonants produced by Japanese and Canadian English speakers in multimodal perception (i.e., AV condition), the averaged intelligibility scores were submitted to repeated measures ANOVAs with speaker language group (Japanese and English), and consonant type (i.e., /b, v, s, θ, l, ɹ /) as the within-group factors. Main effect of speaker language group was observed [$F(1, 28)=1105.171$, $p < .001$, partial $\eta^2=.975$] (Mean JP= 76.1%, ENG= 98.8%). Main effect of consonant type was also observed [$F(5,$

140)=160.003, p<.001, partial η^2=.851]. Overall mean intelligibility rates are as follows: /b/= 98.4%, /v/= 75.5%, /s/= 98.4 %, /θ/= 81.2%, /l/= 95.7%, and /ɹ/= 75.6%. A significant interaction of speaker language and consonant type was observed [F(5, 140)=162.581, p<.001, partial η^2=.853], thus post-hoc tests were run to reveal the interaction. English productions were perceived as more intelligible such as /b, v, θ, l, ɹ/ (p<.05), but not in the perception of /s/ (p=.676). With regard to within-language group comparisons, Bonferroni post hoc tests showed that while differences in the mean intelligibility scores of the six consonants produced by the English speakers were not all significant (p>.05), Japanese-produced consonants behaved differently as we saw in AO and VO condition. /s/ (98.2%), /b/ (97.7%) and /l/ (92.6%) were perceived significantly more accurately than /θ/ (63.9%), /v/ (52.4%), and /ɹ/ (51.6%) (p<.005). Among the least accurate consonants, the intelligibility rate of /θ/ was significantly higher than /v/ and /ɹ/ (p<.005), and yet the rest of the two did not show a significant difference (p >.05).

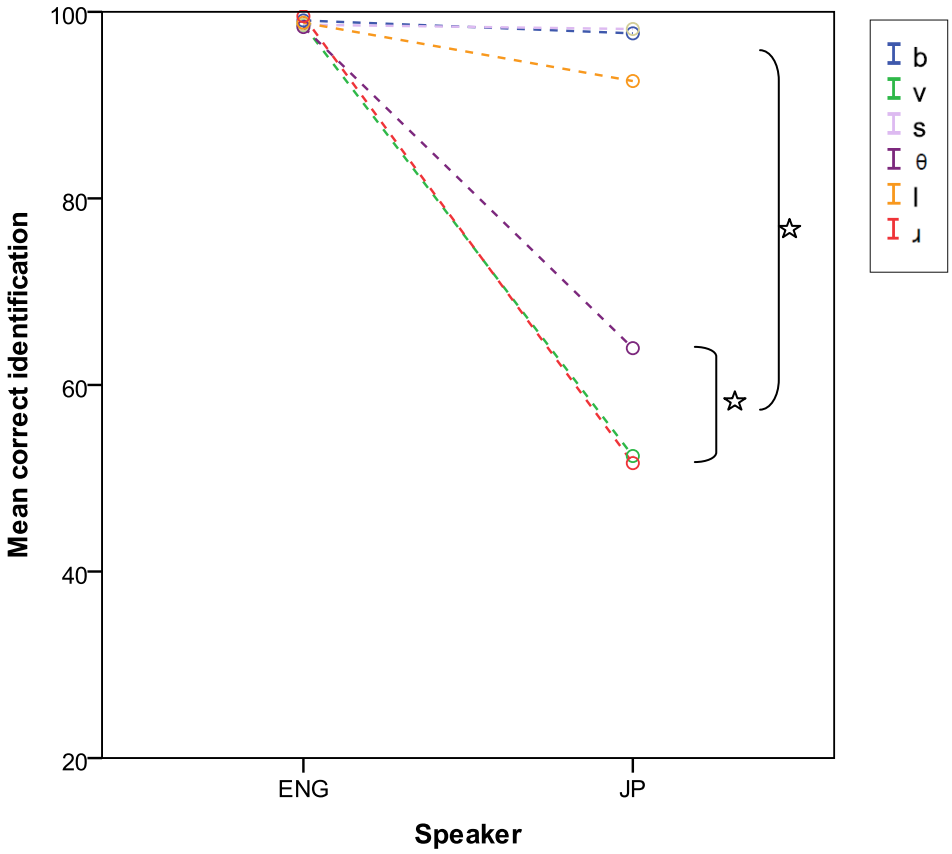


Figure 3. ***Mean correct identification rates of each consonant given by English (ENG) and Japanese speakers (JP) in AV. Stars indicate statistically significant differences.***

5.1.1.4. Summary of cross-linguistic comparison

In this section, consonant identification rates in each modality are shown in order to display how perception was affected by the speaker language and the consonant type. While the AO condition showed all consonants were perceived as more intelligible when produced by the native English speakers, perception of /s/ in the AV condition as well as /b/ and /s/ in the VO condition did not differ in the identification accuracy rates between the two speaker groups. Furthermore, /l/ in the VO condition was perceived as more intelligible among the Japanese productions compared to the English productions. In terms of within-group comparisons, /s, b, l/ in Japanese tended to be perceived more intelligible than /v, θ, ɹ/ in all conditions (AV, AO, VO). Among the English productions, no significant differences in the comparisons of the six consonants were observed in the AV condition, and yet some significant differences were observed in AO and VO conditions, indicating that perceptual accuracy may differ even among the native productions especially in single modality conditions. The details of how modality type affects perception will be revealed in the next section (Section 5.2.).

5.1.2. Identification across modalities

In order to examine how modality type used in perception affects identification scores, separate two-way (consonant and modality) repeated measures ANOVAs were conducted for both Japanese and English speaker groups.

5.1.2.1. Japanese speakers

The mean consonant identification rates of Japanese-produced consonants are displayed in Table 2 and Figure 4. Main effects of modality type [$F(2, 56)=35.548$, $p<.001$] and consonant type [$F(5, 140)=179.552$, $p<.001$, partial $\eta^2=.865$] were observed. Bonferroni adjusted pairwise comparisons revealed that overall mean consonant identification rates in AV (76.1%) were significantly higher than AO (70.7%) ($p < .005$), and the AO was significantly higher than VO (65.1%) ($p < .05$). The mean results of consonants across modalities were as follows: /b/ = 94.1%, /s/ = 88.5%, /l/ = 86.2%, /θ/ = 60.1%, /v/ = 47.7% and /ɹ/ = 47.2%. The interaction of the two factors was also

significant [$F(10, 280)=10.669$, $p<.001$]. Bonferroni post hoc tests revealed how modality type facilitated or worsened the perception of various English consonants produced by the Japanese speakers. The consonant identification rate in AO was significantly higher than VO in the perception of /s/ and /ɹ/ ($p<.005$). On the other hand, the rate in VO was significantly higher than AO in the perception of /v/ ($p<.005$). The rest of the consonants /b, θ, l / did not show significant differences between AO and VO condition ($p>.05$), indicating that the perceptions of both native (i.e., /b/) and non-native (i.e., /θ, l/) consonants via visual information did not differ from the ones via auditory information.

In the comparisons between AV and VO, first, the identification accuracy in /b/ and /θ/ did not show significant differences ($p>.05$). The perception of /v/ also did not show a difference between AV and VO ($p>.05$), although that of /v/ in VO was significantly higher than AO ($p<.005$). However, the perception of /s, l, ɹ/ was significantly more intelligible in AV compared to VO ($p<.001$), indicating that having both auditory and visual information facilitated the perception of both native (i.e., /s/) and non-native (i.e., /l, ɹ/) phonemes.

Lastly, the comparisons between AV and AO showed a visual facilitative effect for the native /b, s/ and non-native /v, θ/ as the perception in the AV condition was significantly more intelligible than the AO conditions ($p<.01$). While positive effects of visual information on the non-native phonemes were observed, the result of /ɹ/ was not consistent with other consonants. The accuracy rate in the AV condition (51.6%) was significantly lower than the AO condition (55.2%) ($p < .05$) (along with a significant drop between the AV (51.6%) and VO (34.9%) ($p < .001$)). /l/ did not show a significant difference in the identification rates between AV and AO.

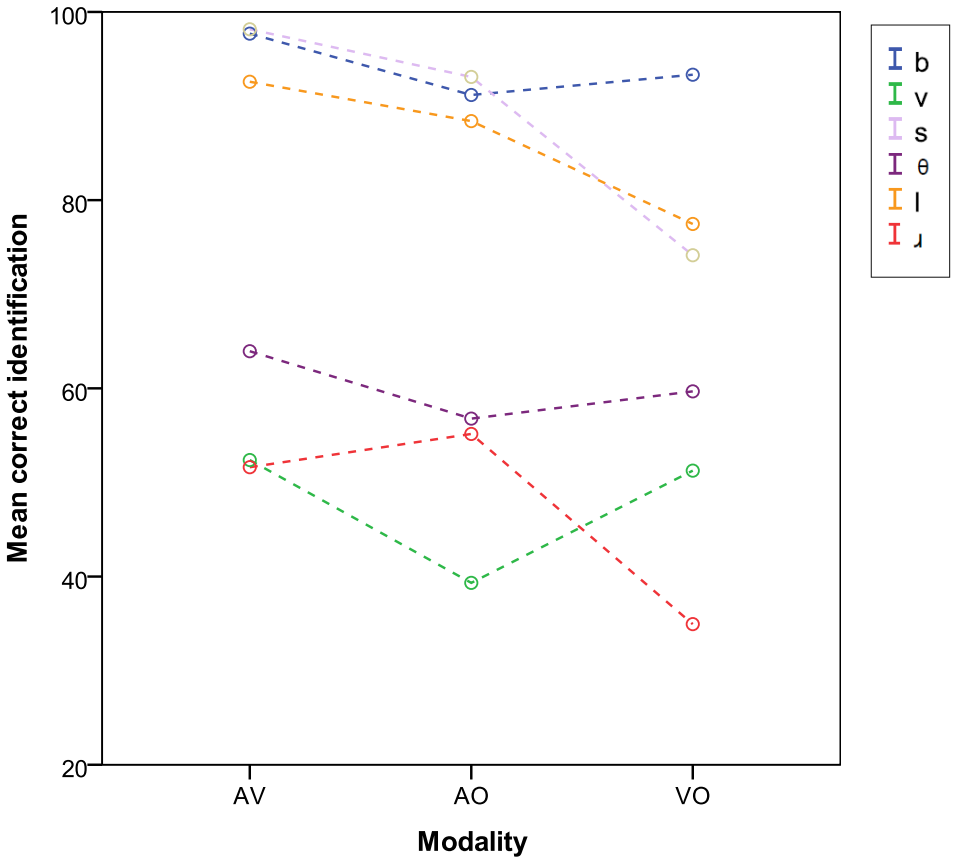


Figure 4. ***Mean correct identification rates of each consonant given by Japanese speakers in the three conditions (AV, AO and VO).***

5.1.2.2. English speakers

This section reports how modality type affected the perception of native English productions. The mean consonant identification rates of English-produced consonants are displayed in Table 2 and Figure 5. Consistent with the cross-modality comparisons with the Japanese speakers, mean correct identification rates including English productions were analysed with a repeated measures ANOVA with consonant type (b, v, s, θ, l, ɹ) and modality type (AV, AO, VO) as the within-group factors. The results show main effects of modality type [$F(2, 56) = 64.626$, $p<.001$] and consonant type [$F(5, 140)=14.261$, $p<.001$]. Bonferroni adjusted pairwise comparisons revealed that overall mean consonant identification rates in AV (98.8%) were significantly higher than AO (95.5%) ($p<.05$) and as well, AO is significantly higher than VO (82.8%) ($p < .005$). Mean intelligibility rates of consonants across the three modality types were as follows: /b/ = 95.6%, /v/ = 93.3%, /s/ = 91.2%, /θ/ = 94.1%, /l/ = 85.6% and /ɹ/ = 94.4%.

The interaction of the two factors was also significant [$F(10, 280)=25.404$, $p<.001$]. First, for AO and VO, pair-wise comparisons with Bonferroni adjustments revealed that AO was significantly better perceived among /s, l, ɹ / compared to VO ($p<.05$). On the other hand, VO was significantly better perceived than AO in the perception of /θ/ ($p<.005$). There were no such differences in /b/ and /v/ ($p>.05$). In addition, comparisons between AV and VO revealed that AV was significantly higher than VO in /b, v, s, l, ɹ / ($p < .05$), but not in /θ/ ($p>.05$). Finally, in the comparisons between AO and AV, the accuracy rates of /b, v, θ/ in AV were significantly higher than AO ($p < .01$), but no significant differences were found in /s, l, ɹ / ($p>.05$).

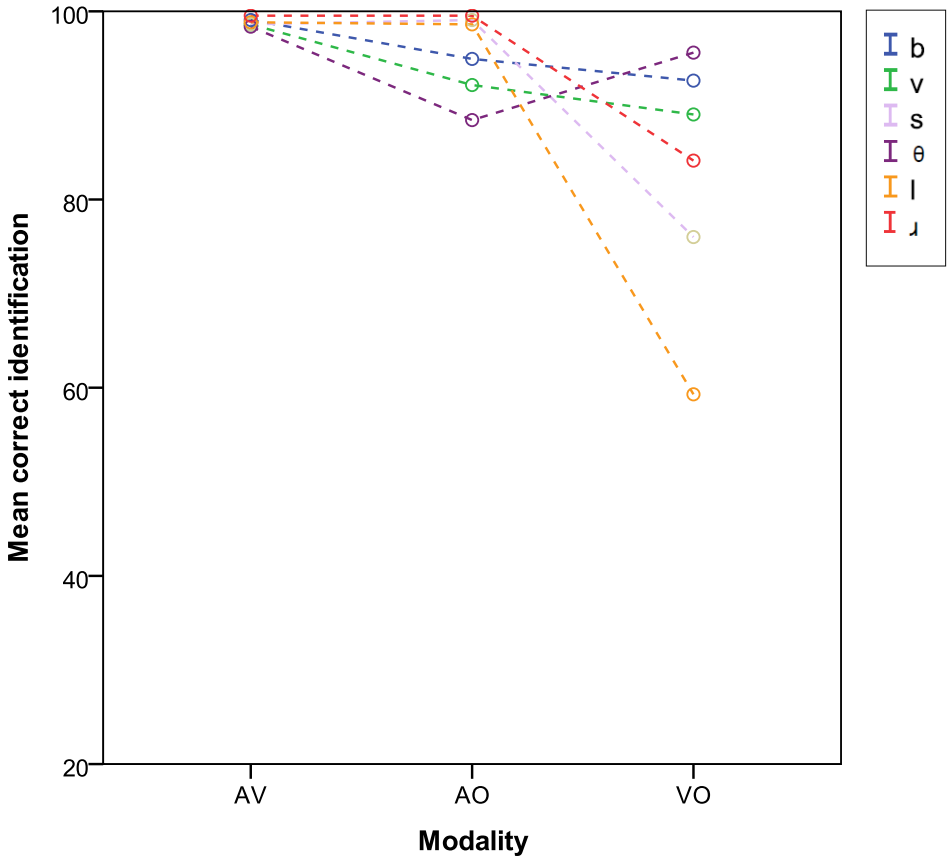


Figure 5. ***Mean correct identification rates of each consonant given by English speakers in the three conditions (AV, AO and VO).***

5.1.2.3. Summary

Results of modality comparisons in each speaker group showed how modality type affected the perception of native and non-native productions. With regard to the Japanese productions, both native (/b/ and /s/) and non-native (/v/ and /θ/) were perceived better in the AV conditions than the AO conditions. However, the perception of /ɹ/ showed different patterns, with the AO condition being more intelligible than the AV condition, along with a significant inferiority of visual intelligibility compared to the audio intelligibility. For /l/, the identification rate of AV was significantly higher than VO, but not in AO. As for the English productions, those which did not show a significant auditory superiority in comparisons between AO and VO (i.e., /b, v, θ/) revealed significant visual benefits in the comparisons between AO and AV. Overall, the perception of both native (English) and non-native (Japanese) productions was greatly affected by the modality type, in that visual information indeed played a significant role in changing perceptual accuracy rates in both facilitative and inhibitory ways. While the current section revealed the effect of visual information in the separate analyses of cross-modality comparisons of Japanese and English productions, it was unknown how the two language groups differ with regard to the effects of visual information. In the following section, further analyses will be reported to examine the differences in the effects of visual information in detail.

5.2. Visual effect size

This section addresses the effect of visual saliency on perception of L2 production with data from /b, v, l, ɹ/ in order to maintain consistency with previous research showing visual saliency effects in audiovisual perception (i.e., Hazan et al., 2006). In previous analyses, two speaker groups were analyzed separately in repeated measures ANOVAs with consonant type (b, v, s, θ, l, ɹ) and modality type (i.e., AV, AO, VO) as the within-group factors. In order to see how two language groups differed with regard to the effect of visual salience, a repeated measures ANOVA was run to evaluate the within-subject effects of visual saliency (High, Low), intelligibility level (High and Low), and language group (Japanese and English). A factor of intelligibility level was added as previously shown that /b/ and /l/ were perceived more intelligibility than /v/ and

/ɹ/ in AO and AV. In this analysis, the dependent variable was visual effect (VE), which is also called AV benefits, and is widely used in AV perception studies to measure a degree of possible improvements from the additional visual cues (e.g., Grant & Seitz, 1998; Hazan et al., 2006; Sommers, Tye-Murray & Spehar, 2005). VE was calculated relative to the scores given by AO, namely VE= (AV-AO)/ (100-AO). Figure 6 displays mean visual effect (VE) of /b, v/ (top) and /l, ɹ/ (bottom) given by English and Japanese speakers. There was a significant main effect for saliency [$F(1, 28)=22.415$, $p<.001$], but neither intelligibility level [$F(1, 28)=3.588$, $p=.069$], nor language group [$F(1, 28)=3.871$, $p=.059$] showed significant main effects. Pairwise comparisons with Bonferroni adjustments show that VE for a highly-salient pair (/b, v/) was significantly higher (VE=0.389) than a less salient pair (/l, ɹ/) (VE=0.081). Other main effects did not reach significance (intelligibility: [$F(1, 28)=3.588$, $p=069$]; language group: [$F(1, 28)=3.871$, $p=.059$]). However, there was a significant interaction between intelligibility and language group [$F(1, 28)=9.933$, $p<.005$], and Bonferroni post hoc tests revealed that while there are significant differences between Japanese and English productions, this only occurred among the less intelligible phonemes (i.e., /v, ɹ/) ($p<.001$), and not in the highly intelligible phonemes (i.e., /b, l/). No other significant interactions were found. Thus, although Japanese-produced phonemes were affected by the high and low intelligible phonemes, both language groups showed a significant VE difference on the basis of visual salience on the productions.

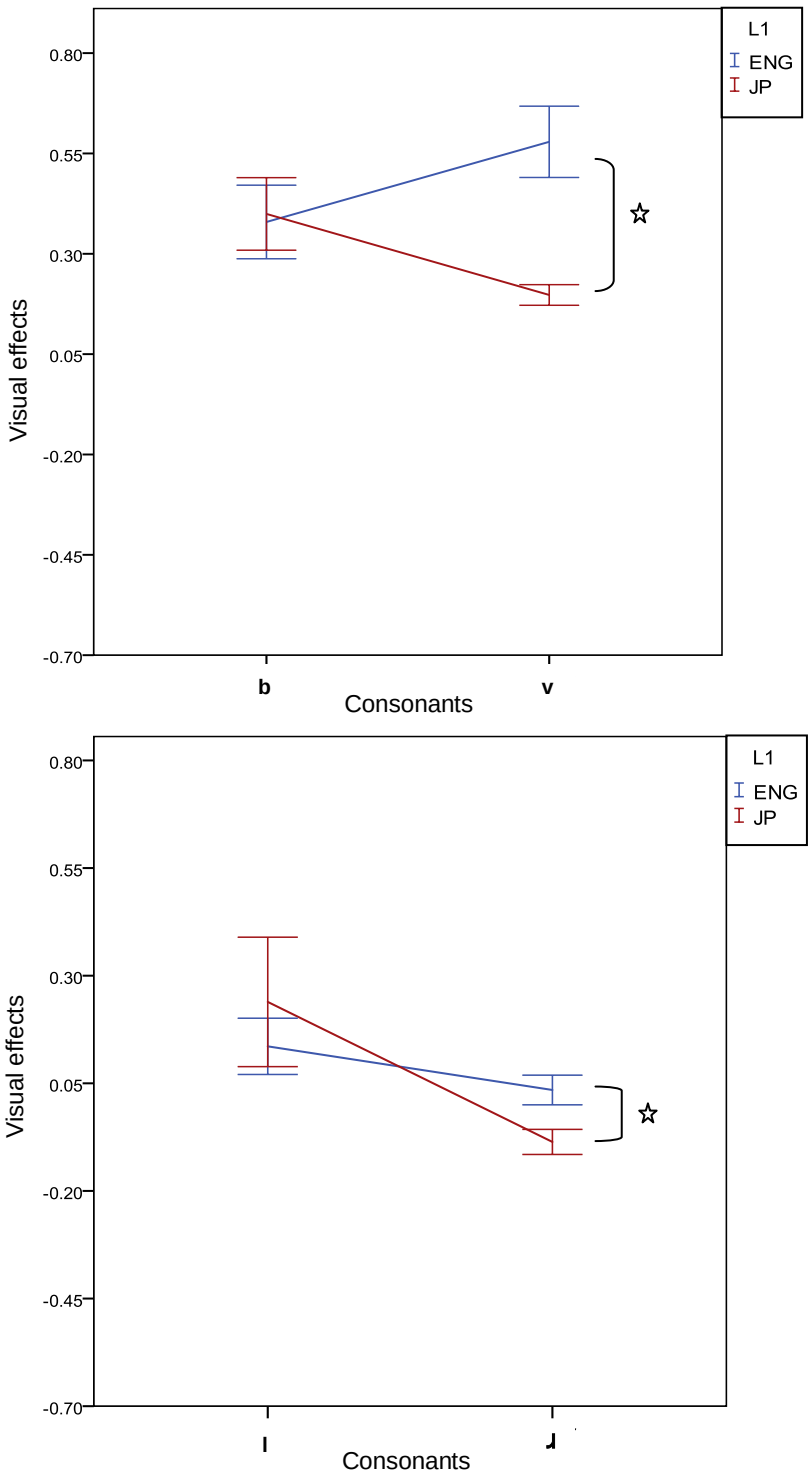


Figure 6. ***Mean visual effect (VE) of /b, v/ (top) and /l, ɹ/ (bottom) given by English (ENG) and Japanese (JP) speakers. The brackets enclose +/- one standard error.***

5.3. Follow-up analysis: Lip-rounding effect on the perception of /ɹ/

Interestingly, a significant negative visual effect was found in the Japanese-produced /ɹ/ (See Section 5.1.2.1) whereas native English speaker productions did not show this effect. While tongue movements are barely visible, lip-rounding of /ɹ/ may serve to increase visual intelligibility. Lip movements have been considered in audiovisual studies (e.g., Traunműller & Őhrstrőm, 2006); however mostly the effects were examined in the context of vowel perception. In fact, lip-rounding is a factor in the production of English /ɹ/ (Ladefoged, 1993), and yet the Japanese speakers may have assimilated /ɹ/ into the native Japanese lateral flap /ɾ/ (Miyazawa, 1993), which is unrounded. As such, it can be assumed that the current results of significant declines in the visual intelligibility of /ɹ/ may have been due to lack of lip-rounding in the Japanese productions. In order to test this proposal, an additional two Canadian English speakers were asked to judge the degree of lip rounding in the production of /ɹ/ using a five point scale (1: no lip rounding, 5: full lip rounding). The same thirty speakers' productions in both Japanese and English speaker groups were presented. The judges were encouraged to use the full five-point scale. High inter-rater reliability was reported with the data from the two judges (The Cronbach's Alpha was .898). A correlation analysis was performed to examine whether visual speech perception accuracy is correlated with the degree of speaker's lip rounding (from the rating results) in the production of /ɹ/. As Figure 7 shows, a high correlation level between the identification accuracy in the VO condition and the degree of lip rounding was found ($r = .793$, $p < .001$), indicating that more visible lip-rounding led the higher visual speech intelligibility of /ɹ/. Accordingly, incorrect articulatory configurations, namely the lack of lip-rounding, decreased the intelligibility of visual speech information, which may have resulted in the inhibitory visual effects in audiovisual speech perception.

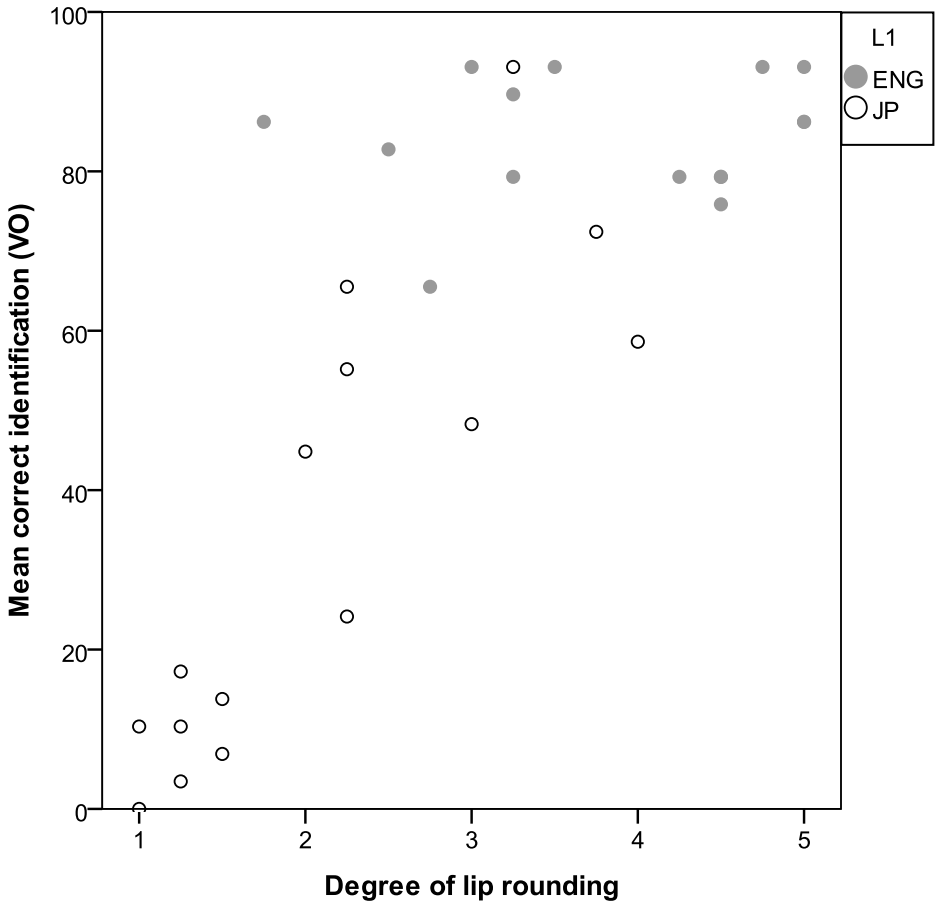


Figure 7. ***Correlation between mean visual intelligibility rates and degree of lip-rounding given by English (ENG) and Japanese (JP) speakers (1: no lip rounding, 5: full lip rounding).***

5.4. Patterns of perceptual confusion

The perceivers' response patterns for the three modality conditions (AV, AO, VO) were analyzed for the six Japanese-produced consonants, as shown in Table 3. Section 5.1.2.1. shows that lower rates of consonant identification /v, θ, ɹ/ were observed. The decline of correct identification rates may be due to assimilations to the L1 phonological counterparts. Japanese inventory does not have /v/ and /θ/, but has /b/ and /s/. Thus, Japanese learners of English tend to replace the non-Japanese phonemes with similar counterparts in their L1 (i.e., /v/ with /b/; /θ/ with /s/) (Yoshida & Hirasaka, 1983). As for /l- ɹ/, Japanese speakers tend to assimilate the /l/ and /ɹ/, which do not exist in Japanese, to the Japanese lateral flap /ɾ/, with /l/ being more likely to be categorized as Japanese /ɾ/ (Miyazawa, 1993). Thus, it was assumed that the the lower rates of consonant identification in / v, θ, ɹ/ may be due to the assimilation to /b, s, l/ respectively.

To compare differences in the degree of perceptual confusions across the three input modality types, the mean consonant identification scores were submitted to a repeated measures ANOVA with response pattern (perceived as correctly or assimilated onto the counterparts (e.g., /b/ for /v/)) and modality type (AV, AO, VO) as the within-group factors, and the post hoc analyses being Bonferroni adjusted. Separate analyses were conducted for each pair (i.e., /b-v/, /s-θ/ or /l-ɹ/) according to the possible assimilations. The analyses reveal how perceivers' response patterns were affected by different modality types.

First, perceptual confusions of consonants shared in both Japanese and English inventories /b, s/ are presented. On the perception of /b/, a main effect of response pattern was found [$F(1, 28)=4325.174$, $p<.001$]. The post hoc analyses showed that, across modalities, /b/ was significantly more likely to be perceived as /b/ than /v/ ($p<.001$). No other difference was observed. For /s/, main effects were found in both response pattern [$F(1, 28)=902.979$, $p<.001$] and modality type [$F(2, 56)=10.712$, $p<.001$]. In addition, the interaction of response pattern and modality type was found [$F(2, 56)=18.077$, $p<.001$]. Post hoc tests revealed that, across modalities, /s/ was consistently more frequently perceived as /s/ than as /θ/ ($p<.001$). However, the rate of /s/ being incorrectly perceived as /θ/ increased in the VO condition compared to the AV

and the AO conditions (p<.05). Thus, native-like consonants did not show perceptual assimilation, despite that perceptual confusion increased in VO condition on the perception of /s/.

On the other hand, misperception of non-native consonants /v, θ, ɹ/ was more biased toward the native counterparts /b, s, l/ than /v, θ, ɹ/ respectively. In the perception of /v/ in which the Japanese speakers tend to substitute the native /b/, no main effects of response pattern and modality type were found [response pattern: F(1, 28)=0.561, p=.460], modality type: F(2, 56)=2.203, p=.120]. However, a significant interaction of the two factors were observed [F(2, 56)=28.385, p<.001]. Pairwise comparisons with Bonferroni adjustments showed that while /v/ was perceived more frequently as /v/ than /b/ in AV and VO (p<.001), /v/ was perceived as /b/ more often rather than perceived as /v/ in AO (p<.001), indicating an assimilation of /v/ to /b/ in the AO condition. Furthermore, the increased bias was observed between AV and AO (AV: 43.0%< AO: 55.6%, p<.001) as well as VO and AO (VO: 39.5%< AO: 55.6%, p<.001). Thus, a significant increase in the bias toward the native counterparts was observed in AO and the bias was significantly decreased when the visual information was presented.

Likewise, /θ/ was more frequently perceived as /s/. Main effects were found in both response pattern [F(1, 28)=93.774, p<.001] and modality type [F(2, 56)=6.559, p<.005]. In addition, the interaction of response pattern and modality type was found [F(2, 56)=4.655, p<.05]. Pairwise comparisons with Bonferroni adjustments revealed that while /θ/ was significantly more often misperceived as /θ/ than the native counterpart /s/ (p<.001), the degree of the bias varied depending on the modality type. While the native English perceivers more frequently perceived /θ/ as /s/ in the three modality types, the bias significantly increased in VO compared to AO (VO: 25.3%< AO: 35.2%, p<.01). No significant differences were found between AV and AO (AV: 31.0%, AO: 35.2%, p=.177) as well as between AV and VO (AV: 31.0%, VO: 25.3%, p=.131). Thus, significant reduction of the perceptual bias in /θ-s/ was found in the visual intelligibility.

Lastly, perceptual confusion of /l/ and /ɹ/ is reported. /l/ is not existent in Japanese, but tends to be judged as similar to Japanese lateral flap /ɾ/ compared to /ɹ/. Main effects were found in both response pattern [F(1, 28)=459.234, p<.001] and modality [F(2, 56)=8.874, p<.001]. The interaction was also found [response pattern x

modality: $F(2, 56)=8.874$, $p<.001$]. Post hoc tests showed that misperception of /l/ was barely affected by different modality types. The rates of perceptual bias toward /ɹ/ were significantly lower in any of the conditions (AV:4.8%, AO:6.4%, VO:9.2%) compared to the rates of perceptual accuracy (i.e., perceived correctly as /l/ in AV:92.6%, AO: 88.4%, VO=77.5%) ($p<.001$), indicating few misperceptions to /ɹ/ regardless of modality types.

As for the perception of /ɹ/, significant perceptual confusions were observed. A main effect of modality was found [$F(2, 56)=23.892$, $p<.001$], but a main effect of response pattern did not appear [$F(1, 28)=3.262$, $p=.082$]. However, an interaction was found [response pattern x modality: $F(2, 56)=13.441$, $p<.001$]. Pairwise comparisons with Bonferroni adjustments revealed that /ɹ/ was perceived more frequently as correct compared to the bias toward /l/ in AO ($p<.001$). However, /ɹ/ was perceived as /l/ significantly more often than /ɹ/ in VO ($p<.05$). In AV, /ɹ/ was not perceived as /l/ more frequently as /ɹ/ ($p=.055$). With regard to the degree of perceptual assimilation of /ɹ/, the rate of perceiving /ɹ/ as /l/ was significantly higher than AO (AV: 43.7%> AO: 39.3%, $p<.05$), indicating that additional visual information triggered further bias to the perception of /l/. No other significance was found.

In sum, the patterns of perceptual bias were affected by the different modality types. In particular, effects of visual information on the perceptual confusions were found among non-native consonants. In /v/ and /θ/, the perceptual confusions to the native counterparts were observed, and yet the degree of confusion was significantly decreased in the VO conditions compared to the AO conditions. Furthermore, a significant decrease was also observed in the AV condition compared to the AO condition in the perception of /v/. On the other hand, visual information also played a role in increasing the bias. In the perception of /ɹ/, the perceptual assimilation to /l/ was increased in AV compared to AO. Thus, visual information affected the degree of perceptual confusion in both positive and negative ways.

Table 3 ***Confusion matrix for consonant identification of Japanese productions with mean identification rates (%). Stars denote statistically significant differences.***

Presented (column)\ Perceived (row)		b	v	s	θ	l	ɹ
b	AV	97.7	1.8				
	AO	91.2	4.8				
	VO	93.3	4.1				
v	AV	43	52.4				
	AO	55.6	39.3				
	VO	39.5	51.3				
s	AV			98.2	1.6		
	AO			93.1	3.9		
	VO			74.1	15.9		
θ	AV			31	64.1		
	AO			35.2	56.8		
	VO			25.3	59.3		
l	AV					92.6	4.8
	AO					88.4	6.4
	VO					77.5	9.2
ɹ	AV					43.7	51.6
	AO					39.3	55.2
	VO					43	34.9

Note. The “other” responses are not included.

5.5. Results of goodness rating task

In this section, the results of goodness rating tasks are presented. This reveals the degree to which modality type affects perception in terms of how good the presented stimuli are. The mean goodness rating scores were analyzed using a repeated measure ANOVA with speaker language group (Japanese and English), consonant type (i.e., /b, v, s, θ, l, ɹ/), and modality type (AV, AO, VO) as the within-group factors. Table 4 represents mean goodness rating scores for six stimulus consonants for Japanese and English speakers in each modality. Significant main effects were found in language group [$F(1, 28)=694.098$, $p < .001$, partial $\eta^2=.961$], consonant type [$F(5, 140)=117.767$, $p<.001$, partial $\eta^2=.808$] as well as modality type [$F(2, 56)=17.521$, $p<.001$]. Pairwise comparisons with Bonferroni adjustments show that the overall native English productions were perceived as better compared to the Japanese productions (Mean JP=3.33, Eng=4.50). AV was the highest rated following AO and VO, but there was no significant difference between AV and AO. Moreover, interactions of language group x consonant [$F(5, 140)=212.771$, $p<.001$, partial $\eta^2=.884$], language group x modality type [$F(2, 56)=46.228$, $p<.001$, partial $\eta^2=.623$], consonant x modality type [$F(10, 280)=28.554$, $p<.001$], and language group x consonant x modality type [$F(10, 280)=6.730$, $p<.001$] were found.

Firstly, 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 in the goodness ratings compared to native English productions in each modality.

Table 4 ***Mean goodness rating scores for six consonants produced by the Japanese and English speakers in AO, VO, and AV conditions. Standard deviations are given (Standard deviation in parenthesis).***

	AO		VO		AV	
	Japanese	English	Japanese	English	Japanese	English
/b/	4.0 (0.61)	4.7 (0.28)	4.1 (0.61)	4.4 (0.50)	4.1 (0.57)	4.8 (0.24)
/v/	2.4 (0.45)	4.5 (0.37)	3.0 (0.39)	4.2 (0.51)	2.7 (0.39)	4.6 (0.30)
/s/	4.0 (0.60)	4.7 (0.29)	3.7 (0.65)	4.1(0.61)	4.1 (0.64)	4.7 (0.28)
/θ/	2.8 (0.55)	4.2 (0.58)	3.2 (0.41)	4.5 (0.33)	2.8 (0.57)	4.5 (0.36)
/l/	3.9 (0.61)	4.7 (0.33)	3.6 (0.51)	3.8 (0.58)	3.8 (0.62)	4.7 (0.23)
/ɹ/	2.7 (0.39)	4.8 (0.27)	2.4 (0.39)	4.2 (0.48)	2.6 (0.37)	4.7 (0.29)

5.5.1. AO condition

In order to investigate how the listeners judged the goodness of Japanese-produced consonants with auditory information, the mean rating scores for the six consonants were submitted to a repeated measure ANOVA with speaker language type (Japanese and English), and the six consonants /b, v, s, θ, l, ɹ/ as the within-group factors. Main effects were found for speaker language type [$F(1, 28)=533.341$, $p<.001$, partial $\eta^2=.950$] and consonant type [$F(5, 140)=114.260$, $p<.001$, partial $\eta^2=.803$]. Overall mean rating scores for the Japanese-produced consonants was 3.29, and that for English-produced consonants was 4.60. The rating scores of six consonants produced by the English talkers were all significantly higher than those by Japanese talkers ($p<.001$). A significant interaction was also found [$F(5, 140)=82.762$, $p<.001$]. Bonferroni post hoc tests showed significant rating differences were found within each language group. As for the Japanese-produced consonants, /b/ (4.01), /s/ (4.00) and /l/ (3.86) were perceived as better compared to /θ/ (2.75), /ɹ/ (2.72) and /v/ (2.40) ($p<.001$). The rating score of /v/ was significantly lower /θ/ and /ɹ/ ($p<.005$). No other statistical differences were found. As for the English productions, the rating scores of /ɹ/ (4.79), /l/ (4.72), /s/ (4.71), and /b/ (4.70) were significantly higher than /v/ (4.47) and /θ/ (4.21) ($p<.05$). In addition, there was a statistical difference between the rating of /v/ and /θ/.

5.5.2. VO condition

With regard to the VO condition, main effects were found for speaker language type [$F(1, 28)=343.434$, $p<.001$, partial $\eta^2=.925$] and consonant type [$F(5, 140)=38.212$, $p<.001$]. Overall mean rating scores for the Japanese-produced consonants was 3.34, and for English-produced consonants was 4.20. As in the AO condition, all of the six consonants were rated as better in English productions than the Japanese productions. ($p<.01$). A significant interaction was also found [$F(5, 140)=123.112$, $p<.001$]. Within the Japanese productions, Bonferroni post hoc tests revealed that /b/ (4.14) was rated higher than any other Japanese-produced consonants ($p<.01$). /s/ (3.69) and /l/ (3.59) were rated significantly higher than /θ/ (3.23), /v/ (2.98), and /ɹ/ (2.41) ($p<.01$). A significant difference was found between /v/ and /ɹ/ as well ($p<.001$).

While the English-produced consonants were perceived as relatively better, some variability was found. /θ/ (4.53) was perceived as the best among the consonants, but the difference from /b/ (4.42) did not reach a significant level ($p > .05$). English-produced /θ/ and /b/ were rated significantly higher than /s/ (4.05) and /l/ (3.81) ($p<.05$). The rating score of /θ/ was also significantly higher than /v/ (4.21) and /ɹ/ (4.20) ($p<.01$). No other statistical differences were found.

5.5.3. AV condition

Similarly to the previous sections, the mean rating scores were submitted to a repeated measure ANOVA with speaker language type (Japanese and English), and six consonants (i.e., /b, v, s, θ, l, ɹ/) as the within-group factors. Main effects were found in speaker language type [$F(1, 28)=501.261$, $p<.001$, partial $\eta^2=.947$] as well as consonant type [$F(5, 140)=91.503$, $p<.001$, partial $\eta^2=.776$]. The overall mean rating score in each language group was 3.34 for the Japanese-produced consonants and 4.68 for the English-produced consonants, and all six consonants were perceived better as English productions compared to Japanese production ($p<.001$). A significant interaction of the two factors was also found [$F(5, 140)=136.525$, $p<.001$, partial $\eta^2=.830$]. Bonferroni post hoc tests revealed that Japanese-produced /b/ (4.12), /s/ (4.06) and /l/ (3.78) were perceived significantly better than /θ/ (2.84), /v/ (2.71), and /ɹ/ (2.58) ($p<.001$). There was also a statistical difference between /b/ and /l/ ($p<.001$).

As for the rating scores from data of English-produced consonants, a significant difference was only found between /b/ (4.75), /l/ (4.74), /s/ (4.73) versus /ɹ/ (4.72), and /θ/ (4.53) (p<.001).

5.5.4. Cross-modality comparisons: Japanese-produced consonants

In order to investigate whether modality type affects perception regarding how good the presented stimuli are, cross-modality comparisons were conducted. Figure 8 summarizes the mean rating scores of each consonant produced by Japanese speakers across the three modality types (AV, AO and VO). The mean rating scores for the six consonants in the three input modalities were submitted to a repeated measure ANOVA with three modality types as input (AV, AO, VO), and six consonants (i.e., /b, v, s, θ, l, ɹ/) as the within-group factors including the data of Japanese-produced consonants. This ANOVA yielded main effects for consonant type [F(5, 140)=197.427, p<.001, partial η^2=.876], but not modality type [F(2, 56)=.793, p=.457]. However, a significant interaction between the two was found [F(10, 280)=18.371, p<.001]. Bonferroni post hoc tests showed that most of the consonants except the perception of /b/ yielded significant differences on the basis of available modality types. In the comparisons of the rating scores between AO and VO, /s, l, ɹ/ were rated higher in AO compared to VO (p<.05). On the other hand, VO were better rated compared to AO on the perception of /v, θ/ (p<.001). There was no such difference in /b/ (p<.05). Furthermore, the comparisons between AV and VO revealed that /s, ɹ/ were rated better in AV compared to VO (p<.05). However, the superiority of VO rating scores among /v, θ/ were also observed compared to AV (p<.01). No other significant difference was observed.

5.5.5. Summary of goodness-rating data

In the goodness rating data, overall English productions were always perceived as better compared to the Japanese productions regardless of modality type. Similar to the identification data, highly intelligible consonants /b, s, l/ were rated higher than low intelligible consonants /v, θ, ɹ/ in the all modality types. English-produced /θ/ was rated relatively lower in AO and AV, but was rated as high in VO. In cross-modality comparisons with Japanese-produced rating data, only /v/ showed a significant increase

in AV compared to AO while the identification data show the visual benefits in AV among /b, v, s, θ/. Auditory superiorities over the visual information (AO>VO) were observed in /s, l ɹ/, and VO ratings in /s, ɹ/ was also lower than AV (AV>VO). Furthermore, visual superiorities over the auditory information (VO>AO) were observed in /v, θ/, and the AO rating of /v/ was also significantly lower than AV.

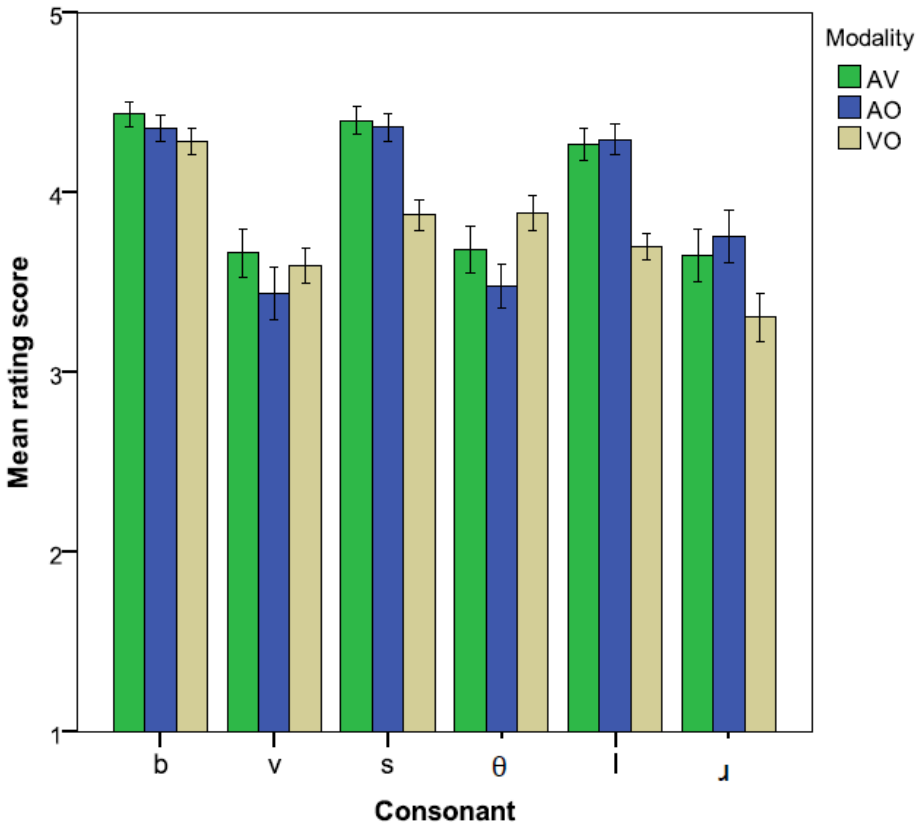


Figure 8. ***Bar graphs showing mean rating scores of each consonant given by Japanese speakers in the three conditions (AV, AO and VO). The brackets enclose +/- one standard error.***

6. Discussion

6.1. General discussion

The aim of this study is to investigate how the AV speech input modality affects native perceivers' perception of L2 consonants. Native Canadian English perceivers assessed English consonants produced by native Japanese speakers in three input modality types (AV, AO, VO). Previous research showed listener variance on native judgements of L2 production as a function of listener experience factors (Clarke & Garrett, 2004; Gass & Varonis, 1984; Isaacs & Trofimovich, 2011; Pinet, Iverson & Huckvale, 2011; Sumner, 2011). In this study, a modality factor of possible variance in native judgements was considered to examine whether additional visual cues affected their judgements of L2 speech intelligibilities. In particular, shared visual cues by L1 and L2 speech were more effectively acquired (Hazan et al., 2005 & 2006), thus perceivers may perceive Japanese productions as more intelligible when presented with both audio and visual information (AV) compared to audio-only information (AO) for consonants that exist in both Japanese and English (i.e., /b, s/). However, the non-Japanese consonants /v, θ, l, ɹ/ could be expected to show negative visual effects on the intelligibility rates in the AV conditions because of assimilations to L1 counterparts for /v, θ/ and /l, ɹ/. As Japanese does not have /v/ and /θ/, 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., /v/ with /b/; /θ/ with /s/) (Yoshida & Hirasaka, 1983) as well as to assimilate the /l/ and /ɹ/, which do not exist in Japanese, to the Japanese lateral flap /ɾ/ (Takagi, 1993). Thus, visual benefits on the non-native phonemes may be lost by their different articulatory configurations affected by their L1. Furthermore, research showed an increase in visual cue weighting on the perception of non-native speech (e.g., Hazan et al., 2008, Sekiyama & Tokuhira, 1993), so the perceivers' additional visual reliance may also induce further visual effects in both facilitative and inhibitory ways.

The results in this study revealed that visual information played a significant role in affecting the native judgements of L2 productions. While we expected the AV benefits mainly in the native phonemes /b, s/, the English perceivers perceived the Japanese productions of the both native and non-native phonemes /b, v, s, θ/ as significantly more intelligible in AV compared to AO. The facilitative visual effects may be due to their enhanced visual weighting on L2 productions as previously shown (Chen & Hazan, 2007 & 2009; de Gelder, Bertelson, Vroomen & Chen, 1995; Fuster-Duran, 1996; Grassegger, 1995; Hazan & Li, 2008; Sekiyama & Tohkura, 1993). This phenomenon has been demonstrated as native perceivers increased their reliance on visual cues when perceiving L2 sounds in the context of McGurk stimuli.

The details of how additional visual cues in AV perception increase L2 consonant intelligibilities will be discussed. As for Japanese-produced native phonemes /b, s/, the audio (AO) intelligibilities reached the native-like identification accuracy. The visual (VO) intelligibility in the Japanese-produced /b/ was also high and was not significantly different from the audio one. By combining the audio and visual information, a further increase was found in the audiovisual (AV) intelligibility. This may be due to further reliance on intelligible VO cues as in the 'non-native speaker effect' (e.g., Hazan et al., 2008, Sekiyama & Tokuhira, 1993). On the other hand, VO intelligibility in /s/ was significantly lower than the intelligible AO cues. This result is consistent with previous research showing less VO intelligibility in alveolar consonants compared to more fronted consonants (c.f., Wang et al., 2009) due to a less visual saliency as the articulatory configurations of /s/ are located at the further back of vocal tract. Although the significant decline in the VO intelligibility was found in /s/, additional visual cues in the AV condition led to a further increase in the AV intelligibility in /s/. This may be due to fewer dependencies on visual cues as Wang et al. (2009) suggest that robust auditory cues enable the perceivers to use less visual cues.

Not only native phonemes, but non-native phonemes /v, θ/ also showed facilitative visual effects in AV perception. Both sounds had significantly lower audio and visual intelligibilities compared to native phonemes /b, s/, and yet the differences between the audio and visual intelligibilities were found within the non-native phonemes. For Japanese-produced /v/, the VO intelligibility was significantly higher than the AO intelligibility. The relatively more accurate VO cues may enable native judges to use

visual cues more effectively by presumably additional reliance on visual cues (i.e., non-native speaker effect). Japanese-produced /θ/ also showed relatively higher intelligibilities in visual cues over the audio cues. The difference did not reach a statistical significance in the identification scores, and yet the native judges perceived the visual intelligibility of /θ/ as significantly better compared to the audio intelligibility according to the goodness rating data. This may imply that the perceivers' representation of /θ/ matched the Japanese-produced /θ/ more in the visual cues compared to the audio cues, which help the use of visual cues. Thus, the positive visual effect in the AV perception of Japanese-produced /θ/ may have been found by the additional visual cues.

However, the present research revealed that the additional visual cues could negatively affect native judgments when articulatory configurations of the L2 speech were incorrect. While the audio intelligibility of Japanese-produced /ɹ/ was as low as other non-native phonemes /v, θ/, the difference is that the visual intelligibility was significantly lower whereas other non-native phonemes showed lower audio intelligibilities. Presumably, if perceivers showed more visual reliance on the perception of Japanese-produced /ɹ/, it is reasonable to understand significant inhibitory visual effects in the AV intelligibility. As we expected that the inhibitory effects were likely to occur due to incorrect articulatory configurations by the Japanese speakers, further analysis was conducted to examine whether the low visual intelligibility of /ɹ/ was due to an incorrect articulatory configuration, namely lack of lip-rounding by the Japanese speakers. While incorrect articulatory configurations may have resulted from various factors, we assumed that Japanese speakers may have assimilated English /ɹ/ onto Japanese lateral flap /ɾ/ (e.g., Takagi, 1993), which is unrounded. The results revealed that a significant amount of Japanese productions of /ɹ/ lacked lip-rounding, supporting that non-native speakers' incorrect articulatory configurations decreased visual intelligibility as well as positive visual effects on audiovisual speech perception. Thus, while visual speech information is generally facilitative in the native perception of non-native speech, the effect of visual information may also be inhibitory when visual speech intelligibility is significantly degraded.

While native and non-native productions showed significant visual benefits in the results of identification and goodness rating tasks, the degree of visual benefits relative to the quality of audio information remains unknown. The effects of visual saliency were significant, which is consistent with previous research (e.g., Navarra et al., 2007; Hazan et al., 2006), showing that visual benefits existed especially with the visually salient sounds. However, the general intelligibility of Japanese productions interacted with the visual effects. Less visually salient pairs /l- ɹ/ in the non-native productions showed significantly fewer visual effects compared to the native productions. Thus, visual saliency played a significant role in determining the visual benefit, indicating that less intelligibility among the non-native phonemes led to a further decline along with less visual saliency.

Section 6.2 discusses the results of each modality condition (AO, VO, AV) to demonstrate the cross-linguistic differences in the perceptions of audio and visual modalities. In the following sections, the results of modality-type comparisons are discussed in detail in terms of how native judgments of L2 production are affected by additional visual information.

6.2. Cross-linguistic differences in the perception of audio and visual modalities

6.2.1. AO results

In order to address how native listener perception is affected by speaker language and consonant type in auditory-based judgments, Japanese-produced consonants were judged by native English listeners in identification and goodness rating tasks. Overall results showed that all of the Japanese produced consonants were perceived as less intelligible compared to productions of native English speaking control subjects. Within the Japanese-produced consonants, significant differences were observed between the native /b, s/ and the non-native phonemes /v, θ/, which is consistent with previous findings (Takagi, 1993; Yoshida & Hirasaka, 1983) showing assimilations of /b-v/ and /s-θ/.

Accordingly, the interaction of first and second language inventories may affect the accuracy of L2 productions as theories of L2 speech acquisition posit (SLM: Flege, 1995; PAM: Best, 1995; PAM-L2: Best & Tyler, 2007) leading to difficulties of producing non-native phonemes compared to native phonemes as intelligibilities of Japanese-produced /v, θ, ɹ/ were significantly lower than the other native phonemes /b, s/.

On the other hand, Japanese-produced non-native /l/ was significantly more intelligible than the other non-native phonemes /v, θ, ɹ/. This higher intelligibility among the non-native phonemes may have resulted from different assimilation patterns. While /v/ and /θ/ are considered to be assimilated to the native counterparts /b, s/ as the current data in the confusion matrix of this experiment demonstrates, both /l/ and /ɹ/ may have been assimilated into the Japanese flap /ɾ/ (Miyazawa, 1993). Although inconsistent findings were found regarding which segment (/l/ or /ɹ/) is easier to produce (/l/: Aoyama et al., 2004, /ɹ/: Hattori & Iverson, 2009), higher intelligibility of /l/ over /ɹ/ was found, which is consistent with Aoyama et al. (2004), where the authors suggested that substitution of Japanese /ɾ/ in the productions of /l/ may lead to the higher intelligibility of /l/.

6.2.2. VO condition

This section discusses how the native perceivers assessed the visual speech intelligibility of the Japanese and the English productions. Similar to the AO results, the VO results showed L1 influences in visual speech intelligibilities. Overall, the English productions were more intelligible than the Japanese productions, particularly among the non-native phonemes for the speakers of Japanese. The identification results show that the English productions were significantly more intelligible among the non-native stimuli /θ, v, ɹ/. 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.

On the other hand, /l/ was significantly better perceived among the Japanese productions compared to the English productions. This is an interesting result since /l/ is a non-native phoneme. No previous studies have addressed the visual intelligibility of

Japanese-produced /l/, but the current study revealed high visual intelligibility. In other words, as long as L2 productions are composed with correct articulatory configurations, negative visual effects can be avoided even in a non-native phoneme. As the above discussion of the AO condition shows, the higher intelligibility of /l/ may be due to Japanese-produced /l/ being mapped onto Japanese /ɾ/, indicating that the substitution of native /ɾ/ may indicate the higher intelligibility of /l/ (Aoyama et al., 2004). With regard to a theory accounting for visual categorical formation, Hazan et al. (2006) introduced three types of visual speech categories in the perception of L2 speech. One of the three cases is when a visual cue is present in both the L1 and L2. The second case is when a visual category occurs in the L2 but not in L1. The last case is when a visual category occurs in both L1 and L2, but is used for a different phonetic distinction. Although none of these match the Japanese productions of /l/, being closer to a native phoneme compared to another similar phoneme /ɹ/ may be a contributing factor to higher intelligibility.

Furthermore, significant group differences between the speakers of Japanese and English were observed in native phonemes /b, s/. Hazan et al. (2006) claimed that shared phonological inventories in L1 and L2 (referred to as native phonemes in this study) could be beneficial in visual categorical development, even though this model is on the basis of L2 perception. The current study also observed the benefits of visual intelligibility in native phonemes but in the production of L2.

Moreover, a significant decline of L2 intelligibility in VO was found in a non-native phoneme. The Japanese-produced /ɹ/ was significantly less intelligible than not only the native phonemes but also the other non-native phonemes. Follow-up analysis was conducted to examine how visual intelligibility in Japanese-produced /ɹ/ could be related to a factor of visual information, namely lip-rounding. A high correlation of visual speech intelligibility with the degree of lip-rounding was observed, indicating that the visual intelligibility of /ɹ/ listener judgements declined when lip-rounding was not sufficiently pronounced. In particular, a significant portion of the Japanese productions of /ɹ/ lacked lip-rounding altogether. Thus, incorrect articulatory configurations of the Japanese speakers may have decreased intelligibility.

6.2.3. AV condition

Similar to the VO condition, Japanese-produced consonants were overall judged significantly less intelligible than the English productions in the goodness rating task. However, the results of the identification task demonstrated different group superiorities across the consonants. English productions were judged to be significantly more intelligible among the perceptions of /v, θ, l, ɹ/ as well as marginally more intelligible in the perception of /b/ compared to the Japanese productions. There was no significant group difference in the perception of /s/. Within the Japanese productions, native phonemes /s, b/ as well as a non-native phoneme /l/ were perceived significantly better than /v, ɹ/, so the higher intelligibilities among the native phonemes were consistent with AO and VO data, suggesting that shared phonemic inventories may facilitate the perception of L2 speech (Hazan et al., 2006). In addition, the lower intelligibilities of Japanese produced /v, θ, l, ɹ/ may be due to assimilation between the native /b, s/ and the non-native phonemes /v, θ/ respectively as well as /l-ɹ/, which is consistent with previous findings (Aoyama et al., 2004; Takagi, 1993; Yoshida & Hirasaka, 1983). As no other studies to date have examined how AV information affects the intelligibility of non-native productions (i.e., L1 perceivers perceiving L1 sounds produced by L2 speakers), consistency with previous studies cannot be discussed. Further visual benefits will be discussed in the next section, taking into the account the results of cross-modality comparisons.

6.3. Effects of modality type on native judgments of L2 productions

Comparing the results across modalities, the findings suggest that additional visual cues are facilitative to the perception of L2 consonants. This is consistent with the previously found non-native speaker effect in which native perceivers demonstrated a greater visual weighting in the perception of L2 productions (Chen & Hazan, 2007 & 2009; de Gelder, Bertelson, Vroomen & Chen, 1995; Fuster-Duran, 1996; Grassegger, 1995; Hazan & Li, 2008; Sekiyama & Tohkura, 1993). The comparisons among the

stimulus consonants showed visual facilitative effects in the perception of both native (i.e., /b, s/) and non-native (i.e., /v, θ/) phonemes as produced by the Japanese speakers. One of our new findings compared to previous studies is the influence of visual cues in perception of /ɹ/ in the AV condition. Interestingly, while other non-native phonemes showed a significant increase in perceptual accuracy with additional visual cues, the perceivers perceived /ɹ/ as less intelligible in the AV condition compared to AO condition. Further analysis showed that the visual intelligibility of /ɹ/ is correlated with the perceived degree of lip-rounding. As a significant portion of the Japanese productions of /ɹ/ lacked lip-rounding, incorrect articulatory configurations of the Japanese speakers may have decreased the AV benefits. Thus, the visual cue weighting could affect L2 speech intelligibility in both facilitative and detrimental ways. While other AV research showed AV benefits in L1 perceivers perceiving L1 sounds in regular and degraded conditions (e.g., Davis & Kim, 2004; Reisberg et al., 1987; MacLeod & Summerfield, 1990; Nielsen, 2004; Sumby & Pollack, 1954) as well as L2 perceivers perceiving L2 sounds produced by L1 speakers (Davis & Kim, 1999; Hazan et al.,2006; Wang et al., 2009), none of this research showed AV positive and negative effects in the context of native judgments of L2 productions.

Results of the confusion matrix also revealed further details of how visual information in non-native consonants affects native judgments of L2 speech production. As speakers of Japanese tend to cause assimilations of /b-v/, /s-θ/ (Yoshida & Hirasaka, 1983) as well as /l-ɹ/ (Sekiyama & Tohkura, 1993), the perceptual confusions could be affected by the additional visual information. The results of the perceptual confusions suggested that decreased and increased identification accuracy may be inversely correlated with perceptual confusion rates. Namely, perceptual confusions were decreased in /v/ when additional visual cues were presented. On the other hand, the perceptual confusion of /ɹ/ to /l/ increased when the visual cues were presented as the saliency of visual cues was relatively lower than that of the auditory cues. Thus, the degree of perceptual confusion was affected in both positive and negative ways in the native judgements. This finding is also interesting as previous AV research included very few analyses of perceptual confusion (Wang et al., 2009), but included only the discrimination scores (e.g., Hazan et al., 2006).

6.4. Theoretical implications

It is also noteworthy to mention theoretical implications on the models of second language speech learning regarding the acquisition of the auditory and visual cues. The speech learning model (SLM: Flege, 1995; 2007) and the perceptual assimilation model (PAM: Best, 1995; PAM-L2: Best & Tyler, 2007) are the two major models for cross-linguistic categorical acquisition considering. Both theories posit that both L1 and L2 sounds exist in the same phonetic space, and the interaction may cause assimilation of L2 sounds according to the similarities and dissimilarities in the articulatory constellation (i.e., PAM) or in phonetic distances (i.e., SLM) from the closest L1 sound. While these models predict difficulty in acquiring new L2 sounds, only auditory cues have been focused on in the study of L2 production. The current study may be able to shed light on potential visual correlates, along with auditory cues, with implications for the theories of L2 speech learning.

Firstly, the present study found symmetrical decreases in the intelligibilities of Japanese-produced non-native phonemes (i.e., /v, θ/) in both the AO and the VO conditions compared to the native counterparts. These findings suggest visual speech acquisition may also be affected by the interaction of L1 and L2 sounds in a similar way to auditory acquisition. It is reasonable to assume this, considering a direct realism based model such as PAM in which L2 speech learning occurs by perceiving the gestural constellation of L2 sounds directly and assessing the similarities to, and discrepancies from, the closest native constellation. Thus, visual speech learning may occur simultaneously when L2 listeners acquire auditory L2 phonetic contrasts.

While similar degradation was observed in both auditory and visual intelligibilities among non-native phonemes compared to the native phonemes, further modality effects were found. For instance, the visual intelligibility of Japanese-produced /ɹ/ was significantly lower than the rest of non-native phonemes and native phonemes. On the other hand, Japanese-produced /v/ was the least intelligible, auditorily. As both SLM and PAM only consider auditory cues, modality influences were not taken into account. Thus, our findings of cross-modality comparisons imply potential modality effects on the consonant intelligibilities of L2 production.

The model of L2 AV perceptual learning introduced by Hazan et al. (2006) also provide significant insights into understanding L2 visual speech acquisition. By comparing the relationship between L1 and L2 phonologies regarding their viseme inventories, the authors introduced three types of visual speech categories in the perception of L2 speech. The first case is when a visual cue is present in both the L1 and L2. The second case is when visual category occurs in the L2 but not in L1. The last case is when a visual category occurs in both L1 and L2, but is used for a different phonetic distinction. The model points out the significant discrepancies regarding audio and visual speech acquisition. For instance, the number of categories is different. Visual speech categories (i.e., visemes) are significantly lower than the number of auditory-based phonemes as some phonemic distinctions are shared in the same visual category. In fact, Spanish learners were shown to be able to successfully acquire the voiced labiodental fricative /v/ visually as the visual category is used in their native voiceless labiodental fricative /f/. However, they have difficulty in hearing the voicing distinction of the labiodental fricative, which does not exist in Spanish. Thus, non-native contrasts differing within the native visual category tend to be less problematic compared to the auditory categorical formation.

Furthermore, L2 learners' awareness may differ in the processing aspect of phoneme learning. Hazan et al. (2006) pointed out that "(L2 learners) will likely possess poor knowledge of specific phoneme category in the L2 and therefore, may have to learn to associate a particular visual cue" (p. 1741). Thus, L2 visual speech learning requires less effort for categorical perception within a visual class, but may require more explicit instruction to guide speakers to successful production.

While Hazan et al. (2006) is the only theory considering audiovisual speech perception, discrepancies between audio and visual speech intelligibilities were not taken into account. Given that intelligibilities of these two modalities are not aligned as we found in the current study, it may be important to extend the model.

In fact, Pisoni (1997) notes the importance of taking into consideration "multimodal relations between the auditory and visual correlates of speech" (p. 28). As no previous studies have examined visual speech acquisition especially in the context of

speech production, this would be an interesting direction for further investigation in future studies.

7. Conclusion

The aim of the present research is to examine the effects of visual information on native judgments of L2 productions. The results in this study reveal that visual information plays a significant role in native judgements of L2 productions. While visual speech information is generally facilitative in the native perception of non-native speech, the effect may also be inhibitory. In particular, incorrect articulatory configurations of /ɹ/ may lead to the decline of visual speech intelligibility.

Overall, the results of the present study add to our current understanding of visual facilitative effects on speech perception. No other studies to date have considered how AV information affects the intelligibility of non-native productions (i.e., L1 perceivers perceiving L1 sounds produced by L2 speakers), and have only examined either L1 perceivers perceiving L1 sounds by L1 speakers in regular (e.g., Davis & Kim, 2004; Reisberg et al., 1987), degraded conditions (MacLeod & Summerfield, 1990; Nielsen, 2004; Sumby & Pollack, 1954), or L2 perceivers perceiving L2 sounds produced by L1 speakers (Davis & Kim, 1999; Hazan et al.,2006; Wang et al., 2009). Thus, the current study gives a new understanding of visual effects on speech perception in the context of native perceivers' judgments of L2 productions.

These findings offer valuable insights into listener-based assessments of L2 production. While previous research examining non-native productions has been based on auditory-based judgements (e.g., Flege, Munro, & MacKay, 1995; Larson-Hall, 2006; Munro & Dewing, 2005; Munro, Flege, & MacKay, 1996), the current research extends to audiovisual perception. It is important to consider the effects of visual information as the results of the present study show that visual information affects native judgements in both positive and negative ways. Thus, audiovisual judgements of L2 productions are necessary to understand how cross-linguistic communications are successfully delivered in face-to-face conversation.

While the present research provides meaningful insight into the effects of visual information on listener-based L2 judgment, there are several limitations that should be considered in future research. Firstly, elicitation techniques may need to be expanded for future research. Variations due to different task types to elicit L2 productions have been shown (e.g.,Flege, Takagi & Mann, 1996; Hardison, 2003; Saito & Brajot, in press). The spontaneous productions of English /l-ɹ/ have been judged significantly less intelligible compared to production in definition and reading tasks (Flege, Takagi & Mann, 1996). Thus, while the current study asked speakers to read CV syllables, L2 spontaneous productions need to be examined in future studies because other possible variance may appear.

Moreover, further research is required beyond segment phonemic contrasts such as prosodic (e.g., Hirata and Kelly, 2010) and sentence-level (e.g., Irwin, Pilling & Thomas, 2011) information. Irwin et al. (2011) examined the effect of regional accent on speech-reading ability with single-clause sentences. Although overall mean rates of correct keywords are less than 10%, intelligibility of the Nottingham talkers were higher than the Glasgow talkers overall. Thus, visual information in the accented-sentences marginally played a role to alter intelligibility regarding the type of regional accent. Further research is needed to address the visual benefits beyond segmental perception.

Variability can be found beyond different elicitation techniques as well as different stimulus sets. While it has been primarily listener effects that have been examined in previous research, influences from talker-listener interactions cannot be ignored. The current research included a sufficient sample size of speakers, 15 speakers in each language group, whereas sample sizes for previous audiovisual speech tended to be much smaller (e.g., two talkers (Chen & Hazan, 2007 & 2009)). At the same time, talker variability may influence native listener judgments as has been reported in other audiovisual studies such as in clear speech (e.g., Gagné, Masterson, Munhall, Bilida & Querengesser, 1994). Thus, further research is required to address the issue of variability.

While these factors still need to be addressed in future research, the current findings have contributed to the understanding of multimodal speech perception by showing how the visual modality influences the intelligibility of non-native speech, as well

as the effects of the interactions between auditory and visual input modalities in non-native speech perception.

References

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology, 15*(9), 839-843. doi:10.1016/j.cub.2005.03.046

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning, 42*(4), 529-555. doi:10.1111/j.1467-1770.1992.tb01043.x

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics, 32*(2), 233-250. doi:10.1016/S0095-4470(03)00036-6

Baker, W., Trofimovich, P., Flege, J. E., Mack, M., & Halter, R. (2008). Child-adult differences in second-language phonological learning: The role of cross-language similarity. *Language and Speech, 51*(Pt 4), 317-342. doi:10.1177/0023830908099068

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America, 114*(3), 1600-1610. doi:10.1121/1.1603234

Best, C. T. (1995). A direct realist view of cross-language speech perception. In *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171-204). Baltimore: York Press.

Best, C. T., & Tyler, M. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O. Bohn & M. J. Munro (Eds.), *Language Experience in Second Language Speech Learning : In Honor of James Emil Flege* (pp. 13-34). Philadelphia: John Benjamins Publishing Company.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707-729. doi:10.1016/j.cognition.2007.04.005

Chen, Y., & Hazan, V. (2007). Language effects on the degree of visual influence in audiovisual speech perception. *Proceedings of the 16th International Congress of Phonetic Sciences*. (pp.2177-2180). Saarbrueken, Germany:

Chen, Y., & Hazan, V. (2009). Developmental factor and the nonnative speaker effect in auditory-visual speech perception. *Journal of the Acoustical Society of America* 126(2), 858-865 doi:10.1121/1.3158823.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America, 116*(6), 3647-3658.

Colin, C., Radeau, M., Soquet, A., & Deltenre, P. (2004). Generalization of the generation of an MMN by illusory McGurk percepts: Voiceless consonants. *Clinical Neurophysiology, 115*(9), 1989-2000. doi:10.1016/j.clinph.2004.03.027doi:10.1121/1.1815131.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: A phonetic representation within short-term memory. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology, 113*(4), 495-506. doi:10.1016/S1388-2457(02)00024-X

Davis, C & Kim, J. (1999). Perception of clearly presented foreign language sounds: The Effects of visible speech, *Proceedings of Auditory-Visual Speech Processing 1999*. Santa Cruz, USA.

Davis, C., & Kim, J. (2004). Audio-visual interactions with intact clearly audible speech. *The Quarterly Journal of Experimental Psychology.A, Human Experimental Psychology, 57*(6), 1103-1121. doi:10.1080/02724980343000701

de Jong, K. J. (1994). The correlation of P-center adjustments with articulatory and acoustic events. *Perception & Psychophysics, 56*(4), 447-460. doi:10.3758/BF03206736

de Gelder, B., Bertelson, P., Vroomen, J., & Chen, H.C. (1995). Interlanguage differences in the McGurk effect for Dutch and Cantonese listeners. *Proceedings of the Fourth European Conference on Speech Communication and Technology*, pp. 1699–1702.

Duncan, J., Martens, S., & Ward, R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature, 387*(6635), 808-810. doi:10.1038/42947

Edwards, J., Beckman, M., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *The Journal of the Acoustical Society of America, 89*(1), 369-382. doi:10.1121/1.400674

Flege, J. E. (1995). Second language speech learning theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233– 277). Baltimore: York Press.

Flege, J. E. (2003) Assessing constraints on second-language segmentalproduction and perception. In Schiller, N. & Meyer, A. (eds) *Phonetics and phonology in Language Comprehension and Production*. pp. 319-355.

Flege, J. E. (2007). Language contact in bilingualism: Phonetic system interactions. In J. Cole & J. I. Hualde (Eds.), *Laboratory Phonology 9*. Berlin: Mouton de Gruyter.

Flege, J. E., Munro, M. J., & Mackay, I. R. A. (1995). Effects of age of second-language learning on the production of English consonants. *Speech Communication, 16*(1), 1-26. doi:10.1016/0167-6393(94)00044-B

Flege, J. E., Takagi, N., & Mann, V. (1995). Japanese adults can learn to produce English /r/ and /l/ accurately. *Language and Speech, 38 (1)*, 25-55.

Flege, J. E., Takagi, N., & Mann, V. (1996). Lexical familiarity and English-language experience affect japanese adults' perception of /ɹ/ and /l/. *Journal of the Acoustical Society of America, 99*(2), 1161-1173.

Fuster-Duran, A. (1996). Perception of conflicting audio- visual speech: an examination across Spanish and German. In D.G.. Stork & M.E. Hennecke (Eds.), Speechreading by humans and machines (pp. 135–143). New York: Springer-Verlag.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34*(1), 65-89.

Goto, H. (1971). Auditory perception by normal japanese adults of the sounds “L” and “R. *Neuropsychologia, 9*(3), 317-323. doi: 10.1016/0028-3932(71)90027-3

Grassegger, H. (1995). McGurk effect in German and Hungarian listeners. *Proceedings of the International Congress of Phonetic Sciences*, (pp. 210–213). Stockholm, Sweden.

Guion, S. G., Flege, J. E., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *The Journal of the Acoustical Society of America, 107*(5 Pt 1), 2711-2724. doi:10.1121/1.428657

Hattori, K., & Iverson, P. (2009). English /r/-/l/ category assimilation by Japanese adults: Individual differences and the link to identification accuracy. *The Journal of the Acoustical Society of America, 125*(1), 469. doi: 10.1121/1.3021295

Hazan, V., Li, E. (2008). The effect of auditory and visual degradation on audiovisual perception of native and non-native speakers. *Proceedings of Interspeech 2008*. (pp.1191-1194). Brisbane, Australia.

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America, 119*(3), 1740-1751. doi:10.1121/1.2166611

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual perception and production of training on the consonants by Japanese learners of English. *Speech Communication, 47*(3), 360-378. doi:10.1016/j.specom.2005.04.007

Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research : JSLHR, 53*(2), 298-310. doi:10.1044/1092-4388(2009/08-0243)

Hirata, Y., Whitehurst, E., & Cullings, E. (2007). Training native English speakers to identify Japanese vowel length contrast with sentences at varied speaking rates. *The Journal of the Acoustical Society of America, 121*(6), 3837-3845. doi:10.1121/1.2734401

Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics, 32*(1), 113-140. doi:10.1017/S0142716410000317

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87*(1), B47-B57. doi: 10.1016/S0010-0277(02)00198-1

Ladefoged, P. (2006). *A course in phonetics* (5th ed.). Boston, MA: Thomson, Wadsworth.

Larson-Hall, J. (2006). What does more time buy you? another look at the effects of long-term residence on production accuracy of English / [inverted r] / and / l / by Japanese speakers. *Language and Speech, 49*(4), 521-548.

Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science, 243*(4890), 489-494.

MacKain, K., Best, C., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics, 2*(4), 369-90.

MacLeod, A. and Summerfield, Q. (1990). A procedure for measuring auditory and audiovisual speech reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology, 24,* 29-43.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, N.J: Erlbaum Associates.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748. doi:10.1038/264746a0

Miyawaki, K., Jenkins, J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics, 18*(5), 331-340. doi:10.3758/BF03211209

Mochizuki, M. (1981). The identification of /r/ and /l/ in natural and synthesized. *Journal of Phonetics, 9*, 283-303.

Morrison, G. (2002). Perception of English /i/ and /l/ by Japanese and Spanish listeners: Longitudinal results. *Proceedings of the Northwest Linguistics Conference*, 29.

Munro, M. J. (2008). Foreign accent and speech intelligibility. In Hansen Edwards, J. G. & Zampini, M. L. (Eds.). *Phonology and Second Language Acquisition* (pp. 193-218). Amsterdam: John Benjamins.

Munro, M. J., & Derwing, T. M. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language Learning, 58*(3), 479-502. doi:10.1111/j.1467-9922.2008.00448.x

Munro, M. J, Flege, J. E., & Mackay, I. R. A. (1996). The effects of age of second language learning on the production of English vowels. *Applied Psycholinguistics, 17*(3), 313-334.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49*(s1), 285-310. doi:10.1111/0023-8333.49.s1.8

Navarra, J., Alsius, A., Velasco, I., Soto-Faraco, S., & Spence, C. (2010). Perception of audiovisual speech synchrony for native and non-native language. *Brain Research, 1323*, 84-93. doi:10.1016/j.brainres.2010.01.059

Navarra, J., Alsius, A., Velasco, I., Soto-Faraco, S., & Spence, C. (2010). Is attention involved in the audiovisual integration of speech? *Information Fusion, 11*, 4-11.

Nielsen, K. (2004). Segmental differences in the visual contribution to speech intelligibility, *Proceedings of Interspeech 2004*. (pp.2533-2536). Jeju Island, Korea.

Pinet, M., Iverson, P., & Huckvale, M. (2011). Second-language experience and speech-in-noise recognition: Effects of talker-listener accent similarity. *The Journal of the Acoustical Society of America, 130*(3), 1653-1662. doi:10.1121/1.3613698

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In Keith Johnson and John W. Mullennix (eds.), *Talker Variability in Speech Processing*. San Diego: Academic, 9-32.

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. 97–114). Hove, UK: Lawrence Erlbaum Associates.

Saito, K., & Brajot, F. (under revision). Scrutinizing the role of length of residence and age of acquisition in the interlanguage pronunciation development of English /r/ by late Japanese bilinguals. Bilingualism: Language and Cognition.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics, 59*(1), 73-80. doi:10.3758/BF03206849

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science, 11*(2), 306-320. doi:10.1111/j.1467-7687.2008.00677.x

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech-perception. *Journal of Phonetics, 21*(4), 427-444.

Soto-Faraco, S., & Alsius, A. (2007). Conscious access to the unisensory components of a cross-modal illusion. *Neuroreport, 18*(4), 347-350.

Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: Evidence from the speeded classification task. *Cognition, 92*(3), B13-B23. doi:10.1016/j.cognition.2003.10.005

Stevens, K.N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In Denes, P.B. and David Jr., E.E. (eds.), Human Communication, A Unified View, 51-66. New York, McGraw-Hill.

Stevens, K.N. and Blumstein, S.E. 1981. The search for invariant acoustic correlates of phonetic features. In P.D. Eimas and J.L. Miller (Ed.). *Perspectives on the Study of Speech*. Hillsdale: Lawrence Erlbaum.

Stekelenburg, J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19*(12), 1964-1973. doi:10.1162/jocn.2007.19.12.1964

Sumby, W., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212. doi:10.1121/1.1907309

Summerfield, Q. (1983) Audio-visual speech perception, lipreading and artificial stimulation. In Hearing Science and Hearing Disorders, edited by M. E. Lutman and M. P. Haggard, Academic (pp. 131–182). London, UK.

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London.Series B: Biological Sciences, 335*(1273), 71-78. doi:10.1098/rstb.1992.0009

Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition, 119*(1), 131-136. doi:10.1016/j.cognition.2010.10.018

Swerts, M., & Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics, 36*(2), 219-238. doi:10.1016/j.wocn.2007.05.001

Takagi, N. (1993). Perception of American English /r/ and /l/ by adult Japanese learners of English: A unified view. Ph.D. Dissertation, University of California, Irvine.

Tiippana, K., Andersen, T., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology, 16*(3), 457-472. doi:10.1080/09541440340000268

Traunmüller, H., & Öhrström, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics, 35*(2), 244-258. doi: 10.1016/j.wocn.2006.03.002

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of United States of America*, *102*(4), 1181–1186.

Sumby, W., & Irwin, P. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212. doi:10.1121/1.1907309

Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America, 124*(3), 1716-1726. doi:10.1121/1.2956483

Wang, Y., Behne, D. M., & Jiang, H. (2009). Influence of native language phonetic system on audio-visual speech perception. *Journal of Phonetics, 37*(3), 344-356. doi:10.1016/j.wocn.2009.04.002

Yoshida, K., & Hirasaka, F. (1983). The lexicon in speech perception. *Sophia Linguistica*, 11, 105–116.

Appendices

Appendix A: Speaker Background

Japanese speakers

Code	Gender)	Age	Native Language	2nd Language	Fluency of 2nd Language	Age of Learning L2	Duration of L2 Learning
401	Female	19	Japanese	English	Moderate	13	7 year
402	Male	24	Japanese	English	Moderate	13	11 years
403	Female	31	Japanese	English	Moderate	12	19 years
404	Female	21	Japanese	English	Moderate	13	8 years
405	Female	22	Japanese	English	Moderate	18	10 years
406	Male	29	Japanese	English	Poor	13	12 years
407	Female	26	Japanese	English	Moderate	12	14 years
408	Female	20	Japanese	English	Moderate	13	7 years
409	Male	28	Japanese	English	Moderate	13	8 years
410	Female	26	Japanese	English	Poor	12	3 years
412	Female	29	Japanese	English	Poor	13	8 years
413	Male	28	Japanese	English	Moderate	12	16 years
414	Male	26	Japanese	English	Moderate	13	13 years
415	Male	27	Japanese	English	Poor	12	18 years

416	Male	24	Japanese	English	Poor	12	6 years

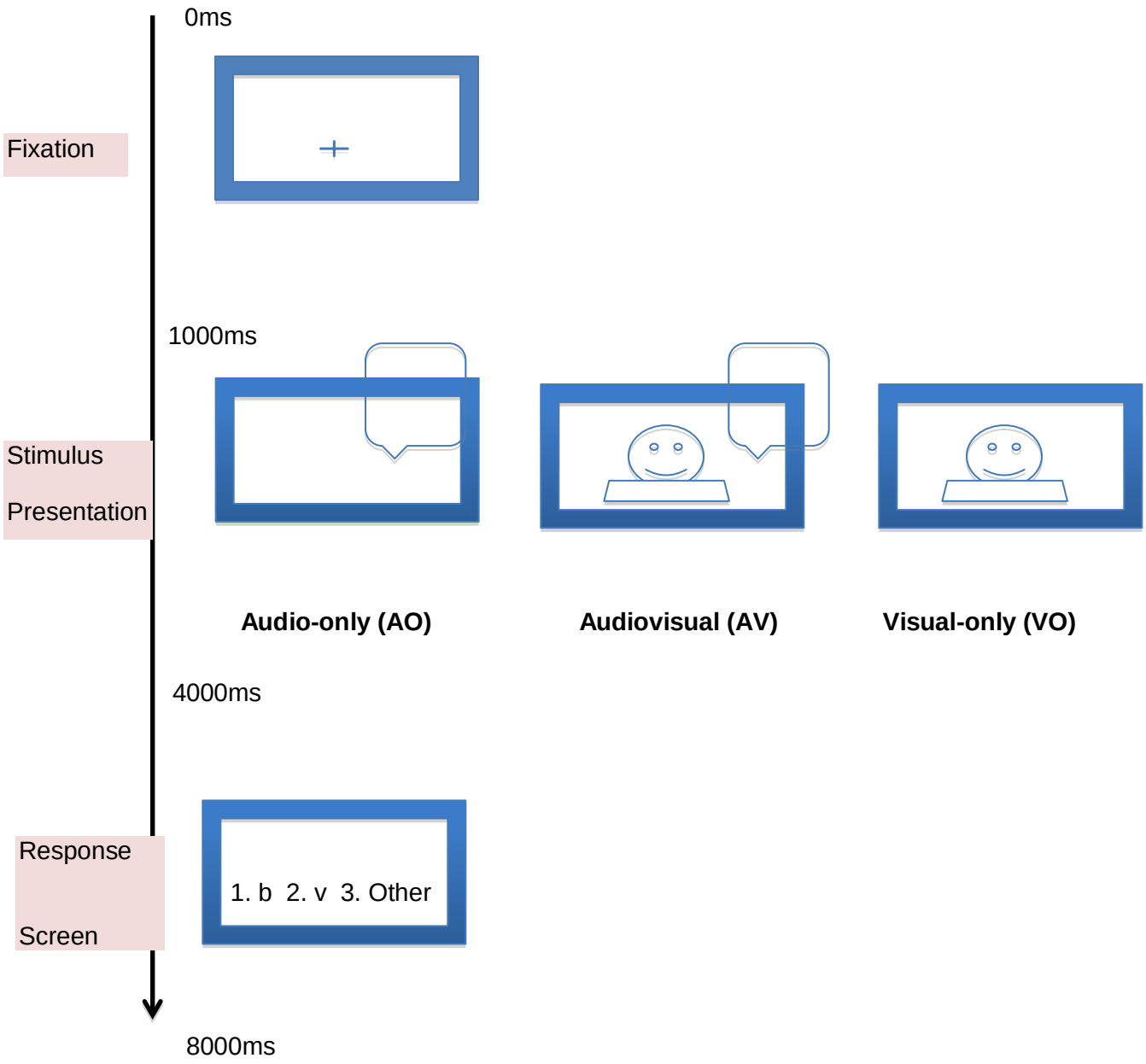
English Speakers:

Code	Gender)	Age	Native Language	2nd Language	Fluency of 2nd Language	Age of Learning L2	Duration of L2 Learning
501	Male	19	English	N/A			
502	Male	24	English	Korean	Poor		
503	Female	18	English	N/A			
504	Female	24	English	French	Poor	7	14
505	Male	21	English	N/A			
506	Female	20	English	Chinese	Moderate	18	20
508	Female	21	English	N/A			
509	Male	23	English	Tagalog	Moderate		
510	Female	19	English	Mandarin	Poor	18	19
511	Female	20	English	Cantonese	Poor	7	10
512	Male	21	English	French	Advanced	5	12

513	Female	26	English	Cantonese	Poor	0	2
514	Female	23	English	Cantonese	Poor	1	6
515	Female	20	English	Cantonese	Moderate	2	
516	Male	21	English	Cantonese	Poor	2	19

Appendix B: Perception tasks

Identification task



Goodness rating task