Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Perceiving foreign-accented auditory-visual speech in noise: the influence of visual form and timing information

Saya Kawase, Jeesun Kim, Vincent Aubanel, Chris Davis

The MARCS Institute, Western Sydney University, Sydney, NSW, Australia
s.kawase@westernsydney.edu.au

Abstract

The present study examined the extent to which visual form and timing information assisted in the perception of native English and Japanese-accented English speech in noise. We also examined whether the degree of visual facilitation would be mediated by the talkers' English experience. Thirty native Australian English listeners performed a speech perception in noise task with English sentences produced by inexperienced and experienced Japanese talkers as well as a native English talker. The Japanese speakers were selected from a previous study where acoustic analyses showed that the speech rhythm of the inexperienced talker was more influenced by their native language than to the experienced one. The stimulus sentences were presented under the three conditions: Audio-only, Audio-visual (visual form and timing) and Audio-visual with mouth covered (visual timing only). The results showed a visual timing facilitation effect for the stimuli produced by the experienced but not in the inexperienced Japanese talker. A facilitative form effect was found for all the talker groups but the size of this effect decreased as the degree of the non-native experience decreased. Our findings illustrate the influence of L2 talker's experience on the effectiveness of their visual form and timing cues.

Index Terms: Visual speech, form and timing, auditory and visual speech processing, foreign-accented speech perception

1. Introduction

Seeing a talker's face/head movements (visual speech) helps speech perception in face-to-face conversation, and is particularly effective when speech is presented in background noise [e.g., 1, 2, 3]. This is possibly because visual speech provides extensive gestural cues associated with speech production, namely, form and timing information [4]. Visual form information is transmitted by changes in articulatory gestures (with mouth, lip, tongue) and can provide phoneme information. The effect of visual form is clear with respect to phonemic perception. A showcase of such is the McGurk effect where the perceptual integration of visual /g/ and auditory /b/ cues result in the perception of /d/ [5]. Robust enhancement of intelligibility due to the visual form was shown in numerous studies [e.g., 1-3], enlightening the benefits of "lip-reading" for the perception.

Although many studies have been conducted to examine visual speech effect at the segmental levels, few have looked at how prosodic information is delivered by visual speech. Recent studies have provided evidence of "visual timing cues" in speech perception. Here timing information is hypothesized to be carried by the cyclic opening and closing of the mouth, as well as changes in peri-oral regions – these gestures carry information about speech onset, offset, and rhythmic information [4-9]. A way to isolate visual speech timing cues has recently been developed and the effect of visual timing has been found [4, 9]. For example, Kim et al [9] presented a stimulus talker's face with the talker's mouth region covered (i.e., cover segmental information) and found improvements in speech identification in noise compared to when speech was presented without visual speech. Furthermore, it has been suggested that visual timing cues help speech segmentation, particularly in noise [7]. Thus, the visual timing effect seems to be robust, particularly in an auditorily degraded condition.

It is also important to consider the usage of visual cues with regards to the talker-listener's language background. That is, the above-mentioned visual form and timing effects were observed when the taker (stimuli) and the perceivers shared a same linguistic experience. Here, we are interested in when their linguistic system does not entirely match between the talker and the listener, such as foreign-accented speech perception. Foreign-accented speech refers to the speech production of non-native talkers [10], differing from native speech in both segments [e.g., 11-13] and timing [e.g., 14-21]. Recently, research showed that visual speech can also help foreign-accented speech perception in noise [22, 23], yet it still remains to be understood to what extent the foreign-accented visual form and timing effect appears. Thus, the current study aims to address this question.

A recent study illustrates different visual form effects in foreign-accented speech processing compared to native speech. Kawase et al. [24] found visual facilitative effects on the perception of English consonants (/b, v, θ/), produced by the non-native Japanese speakers of English. Conversely, an inhibitory effect was also found in the Japanese-produced /ɹ/. A follow-up analysis suggested that this negative visual effect may be due to the Japanese speakers' different articulation with the influence of L1 counterpart (i.e., Japanese flap /ɾ/) which does not involve a lip-rounding. In turn, native English perceives were more likely to identify /la/ instead of /ra/ when the visual speech was available. These findings indicate that similarities between native and non-native visual categories may determine to what extent the visual form information can be perceived correctly.

The differences between native and non-native speech are also shown in timing characteristics. For example, stress-timed languages (e.g., English, Dutch) have stress and vowel reduction but not in syllable-timed languages (e.g., Spanish, French) nor in mora-timed languages (e.g., Japanese) [14], and the different rhythm types between L1 and L2 have also been shown to affect how L2 rhythm is produced in previous acoustic findings [e.g., 14-21]. Our concurrent research showed the influence of L1 on L2 rhythm production with the

 doi: 10.21437/SpeechProsody.2016-99

non-native talker's second language experience, such that the native English talkers and experienced Japanese talkers produced larger variability in vowel duration compared to the inexperienced Japanese talkers [21]. The reduced variability of the inexperienced Japanese production can be due to an influence of their L1 timing characteristics, as Japanese does not employ stress or vowel reduction, resulting in less successful in producing vowels of variable duration in English [25]. Given the existence of native language (acoustic) rhythm influence on L2 production, this effect may appear in visual speech to be perceived as foreign-accented visual timing.

Thus, the current study followed up [16], adopting the mouth masking methods used in Kim and Davis [4], in order to investigate the influences of visual timing as well as form information on the perception of Japanese-accented English in noise. In addition, we examine the effect of L2 experience (experienced and inexperienced) in the visual speech by selecting two Japanese talkers based on rhythmic measures conducted in our concurrent study [21] (i.e., one talker's speech rhythm was much closer to that of native speaker's than the other talker's speech rhythm). The speech stimuli were presented in Audio-only (AO: with a static figure; no mouth movement), in Audio-visual (AV: with mouth movement) and in Audio-visual with the mouth covered (AVm: with mouth covered) conditions (the details is described in Section 2.2.4). It was expected to observe different intelligibility levels across conditions with the intelligibility of the inexperienced Japanese being lower than that of experienced Japanese and native English, and the experienced Japanese to be lower than the native English in AO, AV and AVm. More importantly, we predicted there would be a reduced facilitative effect in visual timing as well as form on the perception of Japanese-accented English compared to native English speech. Among the Japanese-accented English speech, we also predicted a further decrease in these visual speech effects as degree of dissimilarity increase.

2. Methods

2.1. Participants

Forty native Australian English perceivers (34 female, 6 male; M_{age} = 21.6) participated in this study. They were recruited from the University of Western Sydney using the university's research participation system. All of the participants reported normal hearing and normal or corrected-to-normal vision. Our questionnaire also confirmed that none of the participants were familiar with Japanese-accented English. Data from five participants were excluded due to their language background (i.e., simultaneous bilingual).

2.2. Stimuli and Experimental Design

2.2.1. Materials/Talkers

The materials consisted of 234 IEEE Harvard Sentences produced by two Japanese talkers and one Australian English talker (all females; M_{age} = 24.0 years) who resided in Sydney. The Japanese talkers consisted of one 'inexperienced talker' (I-NJ) of English whose mean length of residence in Australia was relatively short (LOR = 4.5 months). Our additional foreign accent rating study by native Australian English listeners (n = 15) confirmed that her English is 'strongly foreign-accented' (8.1 out of 9). The second Japanese talker was an 'experienced talker' (E-NJ) who had lived in Sydney for more than a year at the time of testing (LOR = 12.5 months). Her English was rated as 'mildly foreign-accented' (4.7 out of 9). Both Japanese participants started learning English as a foreign language in Japan at approximately age 13 (i.e., late learners of English). The monolingual Australian English talker was born and raised in Sydney, and was recruited at the University of Western Sydney. All talkers reported no history of speech, vision or hearing problems.

2.2.2. Stimulus recording

The audio and video recordings were conducted in a sound-treated recording booth. The Japanese and English talkers were given instructions regarding facial expression (neutral) and pose (forward facing, at camera). They were asked to read a list of sentences, one at a time, out loud in a neutral tone whilst being recorded. The set of sentences were recorded twice for each participant, but only the first production was used unless errors or disfluencies occurred in the first production. Each sentence was presented for participants to utter on a 17" LCD computer monitor using DMDX software. The videos were recorded using a Sony HXR-NX30P video camera. Separate audio recordings were made using an externally connected lapel microphone, (an AT4033a audio-technica microphone) in 44.1 kHz, 16-bit mono.

2.2.3. Stimulus Editing

The recorded auditory signals were mixed with the associated speaker's speech-shaped noise at a signal-to-noise (SNR) ratio of -4dB, using Praat [26]. The speech shape noise was produced on the basis of the entire track of the talker's production and was added to the entire stimuli. The RMS level of each mixture was fixed at the value of 0.04.

For the video files, the location of the talkers' lips was tracked using Sensarea software [27] and the videos were edited to ensure that the stimulus talkers' face appeared in approximately the same position across trials. The portion of the video was trimmed to show only the lower region of the face (as in [4]). The files were played at a screen resolution of 640 x 480 with 32-bit in colour (for the moving face) at 50 fps or grayscale (for static face).

2.2.4. Stimulus condition (AV, AVm, AO)

Three types of stimuli were prepared: an AV condition (test trials, n = 60; practice trials, n = 6) where the lower region of the face was presented with visible face and mouth motion; an AVm condition (test trials, n = 60; practice trials, n = 6) where the lower region of the face was presented with covered mouth motion by a gray circular patch (radius 20 of visual arc); and an AO condition (test trials, n = 60; practice trials, n = 6) where only a static face of the talker was presented (See Figure 1). Each condition consisted of the recordings from each of the three talkers (n = 20 each), and none of the stimulus sentences were repeated.

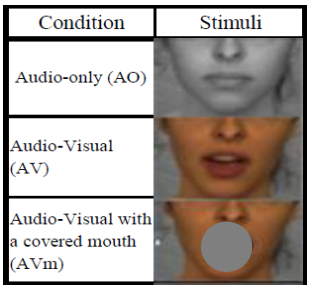

Figure 1: *Illustration of stimulus conditions*

2.3. Procedure

The participants were tested individually in a sound-treated booth. They were instructed to see a talker's face (either moving or static) while listening to the speech in noise over the headphones. A set of MATLAB scripts based on Psychtoolbox were used for stimulus presentation and response collection. In the task, the participants were asked to type in what they heard using a keyboard. A few catch trials were prepared (n = 6 per a condition) to ensure that the participants would watch visual stimuli throughout the entire experiment. In the case of the catch trials, the participants were instructed not to respond to the sentence, and instead were asked to press 'x' when a red cross appeared on a screen. The participants who did not respond to the catch trials (n = 2) were excluded in the analyses. Overall, each participant completed the three modality conditions and the three talker groups, and the presentation order of the condition and talker group was randomized. In total, the perception task lasted approximately one hour.

3. Results

The results of the catch-trials showed that five participants did not pay attention to the visual presentation. Therefore, the data from these participants were removed and the data reported here is from the remaining 30 participants. Firstly, we ran a generalized linear mixed effects logistic regression where keyword identification was the binomial dependent variable (correct vs. incorrect). The fixed effect included modality of stimulus presentation (AO, AVm, AV) as well as the stimulus talker groups (NE, NJ-E, NJ-I). The model also contained random intercepts for participant and keyword. Model comparisons were performed to assess whether the inclusion of each fixed effect as well as their interaction made a significant contribution to the model. The data analyses were performed using the lme4 1.1–7 package in R 3.2.1 [28].

The model comparison revealed that there was significant main effects on the stimulus talker groups ($\chi^2(6) = 165.68$, $p < .0001$) as well as modality of stimulus presentation ($\chi^2(6) = 473.43$, $p < .0001$). In addition, a significant interaction between the stimulus talker groups and modality was found ($\chi^2(4) = 78.52$, $p < .0001$). In order to understand (a) to what extent available visual cues (form and timing) would impact on the perception of spoken sentences and (b) what extent visual effect is different across the stimulus talker groups in each condition, further analyses (a) across modalities (AO, AVm, AV) as well as (b) across stimulus talker groups (native English, experienced Japanese, inexperienced Japanese) were conducted.

3.1. AO, AVm, AV

Separate analyses for each modality was conducted with a generalized linear mixed effects logistic regression where keyword identification was the binomial dependent variable (correct vs. incorrect). Contrast-coded fixed effect included stimulus talker groups (NE, NJ-E, NJ-I). Each model also contained random intercepts for participant and keyword.

Figure 2 shows mean proportion of correctly identified keywords in each modality condition across the stimulus talkers. As clearly observed in the figure, the NE stimuli was perceived significantly more intelligible compared to NJ-E and NJ-I in AO (NE vs NJ-E: $\beta = -0.90$, SE = 0.23, z = -3.9, $p < .0001$; NE vs NJ-I: $\beta = -1.97$, SE = 0.23, z = -8.50, $p < .0001$), AVm (NE vs NJ-E: $\beta = -0.94$, SE = 0.23, z = -4.11, $p < .0001$; NE vs NJ-I: $\beta = -2.25$, SE = 0.23, z = -9.69, $p < .0001$), as well as AV conditions (NE vs NJ-E: $\beta = -1.17$, SE = 0.22, z = -5.22, $p < .0001$; NE vs NJ-I: $\beta = -2.68$, SE = 0.23, z = -11.87, $p < .0001$). In addition, there were significant differences among the Japanese groups, with NJ-E being more intelligible than NJ-I in AO ($\beta = -1.07$, SE = 0.23, z = -4.65, $p < .0001$), AVm ($\beta = -1.31$, SE = 0.23, z = -5.68, $p < .0001$), and AV conditions ($\beta = -1.52$, SE = -0.22, z = -6.80, $p < .0001$).

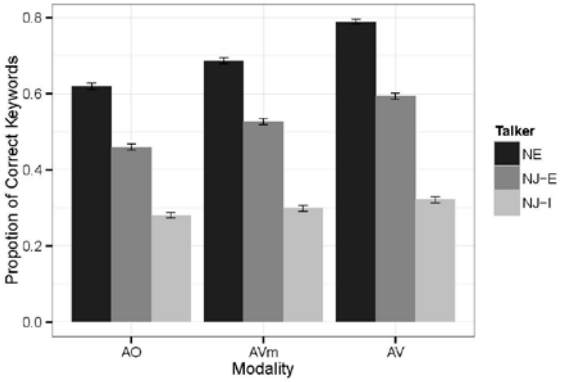

Figure 2: *Mean proportion of correct keywords in the Audio-only (AO) conditions, the Audio-visual with mouth covered (AVm) and the Audio-visual (AV) conditions. produced by native English (NE), experienced (NJ-E) and inexperienced (NJ-I) Japanese talkers. Error bars indicate +/- one standard error.*

3.2. Visual Effects across NE, NJ-E and NJ-I

Furthermore, additional analyses for each stimulus talker group was conducted with a generalized linear mixed effects logistic regression with keyword identification as the binomial dependent variable (correct vs. incorrect). Contrast-coded fixed effect included modality of stimulus presentation (AO, AVm, AV). Each model also contained random intercepts for participant and keyword.

Figure 3 shows mean proportion of correctly identified keywords as a function of stimulus talker groups. As shown in the figure, there was a significant main effect on the modality of stimulus presentation on the perception of native English (NE) stimuli, with the AV being perceived higher than AVm and AO (AV vs. AVm: $\beta = 0.64$, SE = 0.06, z = 10.31, $p < .0001$; AV vs. AO: $\beta = 1.02$, SE = 0.06, z = 16.46, $p < .0001$). The AVm was also perceived significantly more correctly compared to the AO ($\beta = 0.37$, SE = 0.06, z = 6.45, $p < .0001$). A similar pattern was also observed among the experienced

Japanese (NJ-E) stimuli, showing that AV being perceived higher than AVm and AO (AV vs. AVm: $\beta = 0.35$, SE = 0.05, z = 6.461, $p < .0001$; AV vs. AO: $\beta = 0.71$, SE = 0.06, z = 12.80, $p < .0001$), and AVm being perceived higher than AO ($\beta = 0.35$, SE = 0.05, z = 6.47, $p < .0001$).

As for the inexperienced Japanese (NJ-I), there was significant differences in AV vs. AVm as well as AV vs AO, with the AV being perceived higher than AVm and AO (AV vs. AVm: $\beta = 0.18$, SE = 0.06, z = 2.95, $p < .001$; AV vs. AO: $\beta = 0.28$, SE = 0.06, z = 4.55, $p < .0001$). However, there was no significant difference between AVm and AO ($p > .05$).

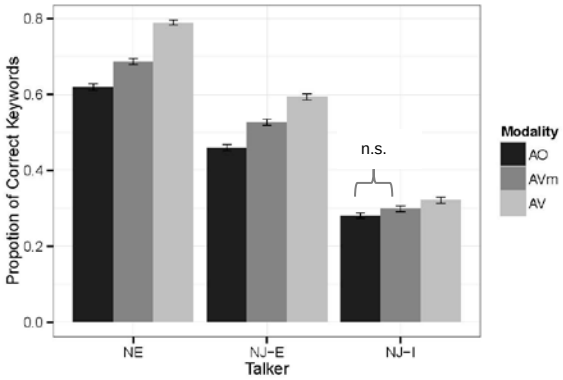

Figure 3: *Mean proportion of correct keywords by native English (NE), experienced (NJ-E) and inexperienced (NJ-I) Japanese in the Audio-only (AO) conditions, the Audio-visual with mouth covered (AVm) and the Audio-visual (AV) conditions. Error bars indicate +/- one standard error.*

4. Discussion and conclusions

The aim of this study was to investigate visual form and timing effects on the perception of Japanese-accented English. In particular, this study examined how the effects would change as a function of the degree of the talkers' foreign accents. Whilst prior studies showed the positive effects of visual form (available in lip-movements) and timing information (gestural changes in perioral regions) with native speech (e.g., 1-9), it was unknown to what extent the visual form and timing information in foreign-accented speech affects the perception of sentences. The current study examined this by manipulating stimuli (audio-only, audio-visual with mouth covered, and audio-visual with the full face). We also considered the talker's L2 proficiency factor by comparing the speech produced by the experienced and inexperienced Japanese talkers (selected from [21]) whose speech timing in English production is different acoustically.

While it is clear that native English speech is more intelligible compared to the non-native English speech, with being more intelligible in the experienced Japanese compared to inexperienced Japanese talkers, visual speech information in foreign-accented speech can also provide facilitation on the perception in noise. For the visual timing, there was a positive visual timing effect in the experienced Japanese talker's speech as well as native English speech. In prior studies, visual timing information has been suggested to produce a facilitative effect such as assisting speech segmentation [7] and such facilitative effect might appear when the foreign-accented speech is produced with similar timing information. Indeed, our prior analyses showed more similar acoustic timing characteristics between native English and the experienced Japanese compared to the inexperienced Japanese talker's speech [21], thus the positive visual timing effect can be attributed to their similarity in the production.

On the other hand, no significant visual timing effect was found in the inexperienced Japanese. Given that the inexperienced talker produced speech timing less similar to the native speech, the visual timing would not be matched with the perceiver's expected timing, thus used less efficiently. Furthermore, the speech rhythm of foreign accented speech may lead listeners to wrong segmentation of speech, and in such case, visual timing information may result in an inhibitory effect by reinforcing rhythmic cues for wrong segmentation.

For the visual form, there was a clear facilitative effect on the perception effects across the stimulus talker groups. Consistent with studies showing visual form effect in native speech [1-3], seeing the talker's lip movements seems to provide robust information even in connected speech. It is worth pointing out that even though visual form effect was shown in the inexperienced Japanese talker's speech, it is relatively minor compared the experienced Japanese and native English stimuli. For native speech processing, visual speech effects tend to be larger as the auditory signal is weaker [29], this may not be the case for processing foreign accented speech processing. That is, it may be that a minimal level of auditory intelligibility is necessary in order to benefit from visual speech information. This requires further investigation.

Overall, the current study introduces the effective use of visual form and timing information even in foreign-accented speech perception, and indicates that visual speech effects interact with listeners' linguistic knowledge of speech form and timing, i.e., the similarity between the native and the foreign accented speech. The greater is the listeners' knowledge for the given speech form and rhythm, the greater visual form and timing effects. In the current study, the visual effect of foreign accented speech is measured in terms of accuracy of speech recognition in a fixed noise level, but further studies are necessary for comprehensive understandings of visual form and timing effect in foreign-accented speech perception. Finally, the current findings offer valuable insights into the usefulness of the face-to-face conversation with non-native talkers as a function of the talker's experience.

5. Acknowledgements

This study was supported by an ARC Discovery grant (DP130104447). The first author acknowledges the support of an Australian Endeavour Scholarship.

6. References

[1] H. McGurk, and J. MacDonald, "Visual influences on speech perception processes," *Perception & Psychophysics, 24*, pp. 253-257, 1978.

[2] K. Nielsen, "Segmental differences in the visual contribution to speech intelligibility," *The Journal of the Acoustical Society of America*, p. 2574, 2004.

[3] W. H. Sumby, and I. Pollack, "Visual contribution to speech intelligibility in noise," *The Journal of the Acoustical Society of America, 26*, pp. 212-215, 1954.

[4] J. Kim, and C. Davis, " How visual timing and form information affect speech and non-speech processing," *Brain and*

Language, 137, pp. 86-90, 2014.
[5] A. Macleod, and Q. Summerfield, "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, pp. 29-43, 1990.
[6] C. Chandrasekaran, A. Trubanova, S. Stillittano, and A. Caplier, A., Ghazanfar, "The Natural Statistics of Speech," *PLoS Computational Biology*, 5, pp. 1-18, 2009
[7] C. Davis, and J. Kim, "Audio-visual speech perception off the top of the head," *Cognition*, 100, pp. B21-B31. J. 2006
[8] S. Greenberg, H. Carvey, L. Hitchcock, and S. Chang, "Temporal properties of spontaneous speech - a syllable-centric perspective," *Journal of Phonetics*, 31, pp. 465-485, 2003
[9] J. Kim, V. Aubanel, and C. Davis, "The effect of auditory and visual signal availability on speech perception," *Proceeding of 18th International Congress of Phonetic Sciences*, Glasgow, UK. 2015.
[10] M. J. Munro, and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning*, 45, pp. 73-97, 1995.
[11] J. E. Flege, "Second language speech learning: Theory, findings, and problems," In Strange, W. (ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research*. Timonium MD: York Press, pp. 233-277, 1995.
[12] J. E. Flege, O.-S. Bohn, and S. Jang, "Effects of experience on non-native speakers' production and perception of English vowels," *Journal of Phonetics, 25*, pp. 437-470, 1997.
[13] N. Takagi, "Perception of American English /r/ and /l/ by adult Japanese learners of English: A unified view," University of California, Irvine, 1993.
[14] R. M. Dauer, "Stress-timing and syllable-timing reanalyzed," Journal of Phonetics, vol.11, pp. 51-62, 1983.
[15] I. Grenon, and L. White, "Acquiring rhythm: A comparison of L1 and L2 speakers of Canadian English and Japanese," Proceedings of the 32nd Boston University conference on language development, 2008.
[16] C. Lai, K. Evanini, and K. Zechner, "Applying rhythm metrics to non-native spontaneous speech," Proceedings of SLaTE 2013, 2013.
[17] H. Lin, and Q. Wang, "Vowel quantity and consonant variance: A comparison between Chinese and English," Proceedings of between stress and tone. Leiden, June 2005.
[18] P. P. Mok, and V. Dellwo, "Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English," Proceedings of Speech Prosody. 2008.
[19] A. Tortel, and D. Hirst, "Rhythm metrics and the production of English L1/L2." Proceedings of Speech Prosody, 2010.
[20] L. White, and S. L. Mattys, "Calibrating rhythm: First language and second language studies," Journal of Phonetics, vol. 35(4), pp. 501-522, 2007.
[21] S. Kawase, J. Kim, and C. Davis, C, "The influence of second language experience on Japanese-accented English rhythm," *Proceeding of Speech Prosody 2016*, Boston, USA. Submitted.
[22] H.G,Yi, J.E.B. Phelps, R. Smiljanic, and Chandrasekaran. B, "Reduced efficiency of audio visual integration for nonnative speech," *The Journal of the Acoustical Society of America*, 134, EL387–EL393, 2014.
[23] B. Banks, E. Gowen, K. J. Munro, and P. Adank, "Audiovisual cues benefit recognition of accented speech in noise but not perceptual adaptation," *Frontiers in human neuroscience*, 9, pp.1-13, 2015.
[24] S. Kawase, B. Hannah, and Y. Wang, "The influence of visual speech information on the intelligibility of English consonants produced by non-native speakers," *The Journal of the Acoustical Society of America*, 136, pp. 1352-136. 2014.
[25] Mochizuki-Sudo, and S. Kiritani, "Production and perception of stress-related durational patterns in Japanese learners of English," *Journal of Phonetics*, vol. 19(2), pp. 231-248, 1991.
[26] P. Boersma, and D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 5.3.51, retrieved 2 June 2013 from http://www.praat.org/.
[27] P. Bertolino, "Sensarea: An authoring tool to create accurate clickable videos," *In 10th workshop on content-based multimedia indexing* (pp. 1–4). Annecy, France. 2012.
[28] D. Bates, M. Maechler, B. Bolker, and S. Walker, "lme4: Linear mixed-effects models using Eigen and S4 (Version R package version 1.1-7)," Retrieved from http://CRAN.R-project.org/package=lme4, 2014.
[29] Jongman, R. Wayland, and S, Wong, "Acoustic characteristics of English fricatives *The Journal of the Acoustical Society of America*. 108, pp. 1252–1263, 2000.