Some effects of audio-visual speech in perceiving Korean

Jeesun Kim [1] *& Chris Davis* [2]

[1]Behavioral Science Research Center, Korea University, Seoul, KOREA.

[2]The University of Melbourne, Melbourne, AUSTRALIA

ABSTRACT

The experiments reported here investigated whether seeing a speaker's face (visible speech) affects the *perception* and *memory* of Korean speech sounds. In order to exclude the possibility of top-down, knowledge-based influences on perception and memory, the experiments tested people with no knowledge of Korean. The first experiment examined whether visible speech (Auditory and Visual – AV) assists English native speakers (with no knowledge of Korean) in the detection of a syllable within a Korean speech phrase. It was found that a syllable was more likely to be detected within a phrase when the participants could see the speaker's face. The second experiment investigated whether English native speakers' judgments about the duration of a Korean phrase would be affected by visible speech. It was found that in the AV condition participant's estimates of phrase duration were highly correlated with the actual durations whereas those in the AO condition were not. The results are discussed with respect to the benefits of communication with multimodal information and future applications.

1. INTRODUCTION

This paper is concerned with investigating some perceptual consequences of listening to Korean as foreign speech with and without the visual information of the movements of a speaker's face (visible speech). There are numerous demonstrations that incorporating the speaker's face facilitates understanding a noisy speech signal (Binnie, Montgomery, & Jackson, 1974; Dodd, 1977; Sanders & Goodrich, 1971; Summerfield, 1979). In addition, some recent studies have demonstrated that the production and later memory of foreign speech sounds is also assisted by visible speech (Davis & Kim, in press; Reisberg, McLean, & Goldfield, 1987). However, it is not clear whether visible speech aids in the perception of the speech sounds themselves or acts to facilitate top-down, knowledge-based processes involved in perceiving and integrating speech into a message. To understand which of these options is the case, one can isolate the perceptional process from the knowledge-base process by testing whether visible speech facilitates the perception of Korean speech for people with no knowledge of Korean.

Apart from providing a way of partioning the effects of visible speech, testing participants on non-native speech also raises issues of foreign language learning. Learning the sounds of a foreign language can be difficult; a common experience is that they are indistinct and occur too rapidly to be adequately registered. Experiments aimed at investigating this issue have largely focused on the difficulty adults experience in learning non-native phonemes and have often examined performance using an AX discrimination task using synthesized stimuli that vary on some continuum (e.g., stimuli that vary on a "rock" - "lock" continuum, Strange & Dittman, 1984). Less attention has been paid to the different task of detecting a sound in unfamiliar continuous-speech, although this is a more ecological valid situation (see Flege & Hillenbrand, 1985; Williams, 1979, for a discussion on different techniques of assessment of L2 perception).

The first experiment will examine whether the detection of a target syllable within a Korean

spoken phrase (syllable monitoring) is enhanced by the addition of the speaker's moving face (AV condition) in comparison to a static face (AO condition). Since Korean is a syllable-based language, it is simple to select a suitable syllable to act as a target sound. Focusing on syllable monitoring may be more relevant to other language processing as a syllable is more perceptually stable than a phoneme (c.f., Cox, Norrix, & Green, 1999). The current study was run outside of Korea in order to facilitate finding participants who have no knowledge of Korean.

Study 1

The first experiment will examine both response latency and error data. A straightforward prediction based on the facilitatory effects of visible speech on detection of sounds in noise is that syllable detection should be easier (faster and less errors) when visual information about the speaker's face accompanies the auditory information (the visible speech condition).

2. METHOD

2.1. Participants

16 graduate and undergraduate native speakers of English from the University of Melbourne participated in the experiment. The ages of participants ranged from 27 to 42 years. None of the participants knew Korean.

2.2 Materials & Design

The experimental items consisted of color digital videos of 40 short phrases in Korean that were spoken by a female native Korean speaker. All the items consisted of 8 syllables and were made up or either 3 or 4 morphemes. The duration of stimuli ranged between 1.75 and 2.75 seconds. The speaker was seated in a well-lit anechoic chamber and positioned at a fixed distance from the camera and recorded against a blank background. The speaker was instructed to keep her expression neutral and her whole head movements to a minimum during the recordings. Only the speaker's face was recorded.

Each experimental item consisted of the following: First, a visual alerting signal (the word "ready" displayed in the center of a video monitor for 500 ms). Following this a single target syllable of average duration 620 ms was played within an envelope of approximately 1300 ms of silence. After an additional 500 ms silence either a still photo of the speaker was displayed while the carrier phrase was spoken (the audio - only condition) or the video of the speaker was shown (the auditory and visual condition). The photo and the video were the same screen dimensions (17 cm wide X 15 cm high). It should be noted that the target syllable (the sound presented to participants before the carrier phrase) was recorded separately to the carrier phrase. That is, the target syllable was recorded as a single spoken syllable and was not extracted from the carrier phrase.

Two experimental lists were constructed such that items that appeared in one list in the no-visible speech condition would appear in the other list in the visible speech condition and vice versa. Within any list, half of the items (20) appeared with visible speech and half without. Also half of the items within the visible or no-visible speech condition contained the target syllable and half did not.

2.3 Equipment

The utterances were recorded using a SONY MVC-FD91 digital video camera at 320 x 240 resolution. The resultant files were displayed using the DMDX software (Forster, & Forster, 1990). This software is a Win32 implementation that uses DirectX 5.0 or higher. The stimuli were presented on a 350 cm video monitor by a Pentium 200 MMX PC equipped with a 4 mb Sgram Diamond Stealth II S220 video card and 32 mb ram and playback was at 30 frames/sec. The auditory components of the stimuli were presented to participants via a set of Sony DR-S7 headphone.

2.4. Procedure

Each participant was tested individually in one or other of the two item lists. Participants were given a standard set of instructions that explained that they would first hear a spoken sound (target sound) and then a foreign language phrase (carrier). They were asked to press the left mouse button as soon as they heard the target sound in

the carrier phrase and to do nothing if they did not. They were told that along with the spoken carrier they would either see a still photograph of the speaker or a video of the speaker saying the phrase. They were told in either case to monitor the screen and the experimenter made sure that they did so.

The experiment began with 4 practice items. Subsequent experimental items were presented in a different pseudo-random order for each participant. That is, the visible and no-visible speech conditions were mixed. No feedback as to the time or correctness of the response was given. Items were displayed automatically 1 second after a response or 2.5 seconds after the end of the spoken phrase.

2.5 Response timing and data treatment

Participant's responses were timed from a clock-on marker that was keyed to a particular video frame. The position of the clock-on marker was determined by reviewing the video for the critical syllable in Adobe Premier 5.1. Clock-on for trials that did not contain the target syllable was from the beginning of the phrase. The latency data of each participant was winsorized (i.e., trimmed back to ± 2sd units from an individual's mean, applied to 2% of data). A lower reaction time cutoff was not applied.

3. RESULTS

The mean percentage of target syllables missed as a function of visible or no-visible speech condition are presented in Figure 1. As can be seen, the average number of misses was less when the participants were able to see the face of the speaker (AV condition) compared with when they could not (AO condition). A paired t-test was conducted that showed that this difference was significant, $t(15)^1 = 3.162$, $p < 0.05$.

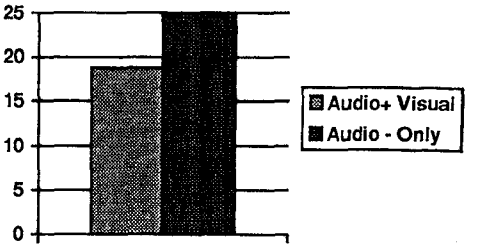

Figure 1: Percentage of target syllable misses as a function of visible (Audio + Visual) or no-visible (Audio - Only) speech condition.

The mean percentages of target syllable errors of commission as a function of visible or no-visible speech condition are presented in Figure 2. Once again, the AV condition produced more accurate detection responses than the AO condition. As before a paired t-test was used to determine whether the difference was significant, and it was, $t(17) = 2.657$, $p < 0.05$.

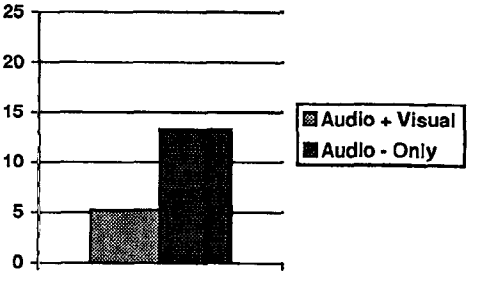

Figure 2: Percentage of target syllable errors of commission as a function of visible (Audio + Visual) or no-visible (Audio - Only) speech condition.

The mean response latencies to correct syllable detection responses (hits) as a function of visible or no-visible speech condition is shown in Figure 3. As can be seen from the figure there was little difference in the mean reaction times for each condition. Indeed a paired t-test indicated that the difference was not significant, $t(15) = 0.08$, $p > 0.05$.

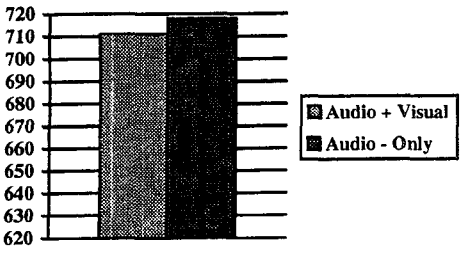

Figure 3: Mean response latencies to correct syllable detection responses as a function of visible (Audio + Visual) or no-visible (Audio - Only) speech condition.

4. DISCUSSION

The results show that the detection of a target syllable within a phrase of Korean (unknown) speech sounds was enhanced by seeing the moving face of the speaker. That is, participants detected a target when it was present more often in the visible speech (AV) condition than in the no-visible speech (AO) condition. Furthermore, when the target was not present, participants in the visible speech condition made less errors of commission. Interestingly, the time to correctly detect that a syllable was in the carrier phrase was not affected by the type of viewing condition. This lack of an effect in response times may be because in the AO condition, errors of omission were made on the hard items, and thus these response times did not contribute to the mean for this condition.

The finding that the perception of Korean (foreign) speech sounds is enhanced with visible speech suggests that the improvement in shadowing of foreign speech (Reisberg, McLean, & Goldfield, 1987) may be in part due to these sounds being perceived more clearly so that they can then be produced more accurately (although observing how to articulate a sound is also important, Catford & Pisoni 1970).

The second experiment was aimed at demonstrating another possible perceptual effect of visible speech that also relates to detecting possible improvements in perceiving and remembering the sound segments of a foreign language phrase. The question of interest was whether judgments about the *duration* of a foreign language phrase would be more accurate with visible speech compared to a no-visible speech condition. The idea being tested stems from the common place observation that unknown speech is often perceived as being too fast to properly apprehend. Given the above demonstration, that visible speech makes syllables more detectable, it may be that people will judge the duration of these phrases to be greater than those with no-visible speech (as the syllable structure of the latter may be indistinct or incomplete).

Study 2

5. METHOD

5.1. Participants

8 graduate and undergraduate native speakers of English from the University of Melbourne participated in the experiment. The participants had not participated in Experiment 1.The ages of participants ranged from 26 to 43 years. None of the participants knew the Korean language.

5.2 Materials & Design

The same 40 digital videos of a speaker pronouncing short Korean phrases used in Experiment 1 were used in this experiment. The duration of each phrase was determined by reviewing it in Adobe Premier 5.1.

Each experimental item consisted of the following: First a visual alerting signal (the word "ready" displayed in the center of a video monitor for 500 ms). After this either a still photo of the speaker was displayed while the phrase was spoken or the video of the speaker was shown. The photo and the video were the same screen dimensions as in Experiment 1. 500 ms after the phrase, the word "START" appeared in red letters 1cm high and 4 cm wide and stayed on the screen until a response was made. Participants pressed the left mouse button when they judged that the same amount of time as the spoken phrase had elapsed. When they pressed the button the word "START" was replaced by the word "STOP".

Two experimental lists were constructed such that items that appeared in one list in the no-visible speech condition would appear in the other list in the visible speech condition and vice versa.

5.3 Procedure

Each participant was tested individually in one or other of the two item lists. Participants were given a standard set of instructions that explained that they would first hear a spoken phrase in a foreign language and that they would be asked to indicate how long it lasted. It was suggested to them that a good way of doing this would be to try to *repeat the phrase* and so judge how long it took. They were told that after they had heard the phrase the word START would appear and that they should press the left mouse button when they thought that the same time had elapsed as the spoken phrase. They were also told that along with the spoken phrase they would either see a still photograph of the speaker or a video of the speaker saying the phrase. They were told in either case to monitor the screen and the experimenter made sure that they did so.

The experimental items were presented in a different pseudo-random order for each participant. That is, the visible and no-visible speech conditions were mixed. As in Experiment 1 no response feed-back was given and items were presented automatically after 1 second after a response.

5.5 Data treatment

In this experiment the latency data was not winsorized as the data represented estimates of the duration of different length utterances and so trimming the data to ± 2 s.d. from a mean value would not be sensible.

6. RESULTS

The mean of participant's estimates of phrase duration as a function of visible and no-visible speech condition are presented in Figure 4. Also presented in the figure is the mean of the actual phrase durations.

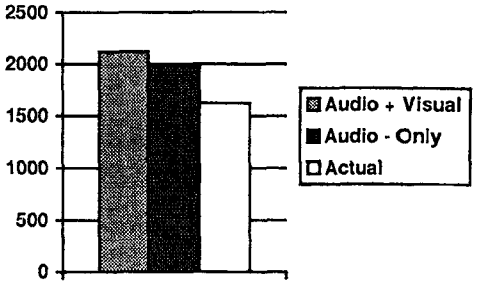

Figure 4: Mean phrase duration estimates (msec) as a function of visible (Audio + Visual) or no-visible (Audio - Only) speech condition. Actual phrase durations are also given.

Somewhat surprisingly, it turned out that the mean estimates of the duration of the phrase were longer than the actual durations. Two paired t-tests were conducted to determine if the estimates in the two conditions differed from the actual durations. They did, $t(39)=9.2$, $p<0.001$, in the AO condition; $t(39) = 17.95$, $p< 0.001$, in the AV condition. Also the estimates of phrase duration were longer in the AV than in the AO condition, $t(39) = 12.669$, $p < 0.05$. That is, the mean estimate for the AO condition was closer to the mean of the actual durations than that of the AV condition (see Figure 4).

In order to determine whether the participant's judgments of duration correlated with the actual durations, Pearson correlation coefficients were calculated for the AV and AO conditions. There was a significant positive correlation between the judgments made with visible speech and the actual judgments, $r = 0.598$, $p < 0.01$ (see Figure 5). However, there was no significant correlation between judgments in the AO condition and the actual phrase durations, $r = 0.083$, $p >0.05$. This indicates that in this condition participants were likely to judge durations without recourse to the actual durations. There was no correlation between the estimates of phrase durations for the AV and AO conditions, $r=-0.226$, $p>005$.

Interestingly, the actual durations correlated significantly with the difference between the estimates of duration in the AV and AO conditions, $r = .344$, $p < 0.05$. This shows that the longer the phrase was the bigger was the

difference between the visible speech estimate and the no-visible speech estimate.

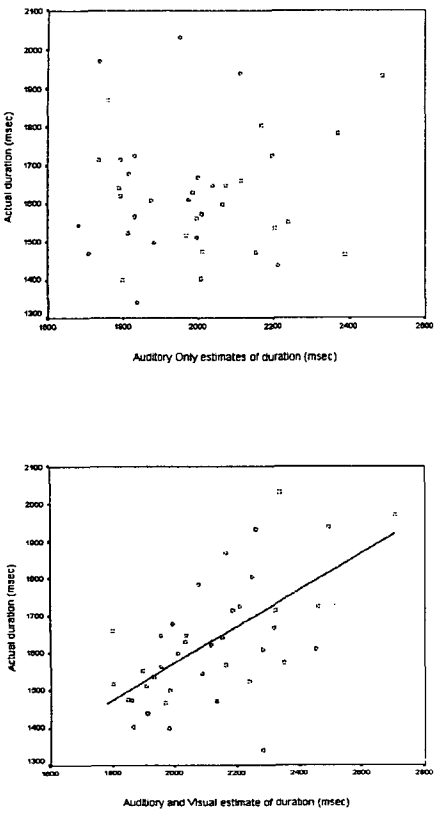

Figure 5: Error in estimates (msec) in the visible (Audio + Visual) and no-visible (Audio - Only) speech conditions as a function of actual phrase durations.

7. DISCUSSION

The motivation for this experiment was based on the familiar notion that foreign speech sounds very rapid and is difficult to remember. It was suggested that the AV condition would produce a more accurate estimate of phrase duration than the AO condition (since visible speech helps the perception of speech in difficult conditions).

The results provided no support for these predictions concerning mean estimates of duration. As it turned out, the means of both AV and AO conditions were larger than the actual durations. Also the estimates of the duration were significantly longer when participants could see the moving face of the speaker (the AV condition) than when they saw a static face. However, an examination of the means of the estimates may be misleading. In terms of whether the AV condition helps improve the accuracy of the phrase duration estimates, what is important is not simply the mean estimates of durations, but whether the estimates are positively correlated with the actual durations. The judgments of phrase duration in the visible speech condition were positively correlated with actual duration. Whereas, there was no correlation between duration judgments and actual duration in the no-visible speech (AO) condition (in spite of the fact that the mean estimates of the durations were relatively closer to the actual duration compared to the AV condition).

The results can best be explained by considering how participants might have made their judgments. In this regard, it is noteworthy to point out that the procedure of the experiment involved suggesting to participants that a good way of performing the task would be to try to *repeat the phrase* and so judge how long it took. If it is assumed that participants followed this instruction and made their estimates by producing internal renditions of the phrases, then there should be a high correlation between these estimates of phrase durations and the actual durations. However, it should also be noted that this method of estimating phrase duration does not require that the means of the predicted and actual duration are similar. This is because people naïve to the sounds of a language may generate an imitation of it at a slower rate. Independent evidence for such a phenomenon comes from Markham (1999).

Although the above account nicely accommodates the findings from the AV condition, where there was a strong correlation between predicted and actual estimates, it does not explain the lack of such a correlation for the AO condition. In this latter case, it must be presumed that judgments of duration were unable to be based upon the repetition of the phrase and were made according to some general strategy, such as responding when a fixed period of time has passed. Evidence

for such a deadline strategy comes from an inspection of the distribution of the errors of the estimates, as it can be seen that these are greatest for the short phrases and least for the long ones. Participants may have been unwilling to respond to the shorter phrases as such decision would require positive evidence (which they did not have in the AO condition), so they simply waited for the expiration of a deadline.

What the above interpretation suggests is that the participants in the AV condition were able to better remember the speech sounds compared with the AO condition. That is, visible speech facilitates the perception of speech segments and in so doing allows for a more faithful memory of them to be made. This way of considering the effect of visible speech is consistent with the work of Davis and Kim (in press) who showed that participants manage to keep a more accurate record of the segments of a phrase in an AV condition compared with an AO one.

8. CONCLUSIONS

The current demonstrations of the effects of visible speech, indicates that a more efficacious learning situation for unfamiliar speech is one that provides the hearer-learner with a range of identification cues, specifically with a view of the speaker's face. This finding has a clear basis in the ecology of human speech communication. Face to face communication provides multiple language recognition cues that the speaker/hearer potentially could make use of in understanding (see Munhall & Vatikiotis-Basteson, 1998).

While these demonstrations of providing multi-modal information show benefits in terms of detecting and remembering speech sounds, there are practical problems with this approach in applying it to language tuition. Single modes of communication, such as text, are cheap and require only minimum input by a human agent. On the other hand, multi-modal communication such as showing the face of the speaker, their body postures and gestures, requires the presence of a speaker for each utterance. This would involve cost and effort in any full-scale learning situation, i.e., filming the speaker for all of the phrases a learner may want to learn. Further, the use of an off-line recording limits the interactivity of the interface by locking the user into predefined exchanges. One way around this would be to replace the human speaker by a computerized talking head that simulates the essential aspects of visible speech (i.e., tongue, teeth, lip and jaw movements, see for instance, Cohen & Massaro, 1994). Indeed, the development of a more practical user interface such as the "talking head" (Cohen, Beskow, & Massaro, 1998; Massaro, 1998), along with automated speech recognition and speech-reading techniques may make possible an automated second language learning system. For example, (Cole, Carmell, Conners, Macon, Wouters, de Villiers, Tarachow, Massaro, Cohen, Beskow, Yang, Meier, Waibel, Stone, Fortier, Davis, & Soland, 1998) describes just such an integrated approach to developing a learning system for language training with profoundly deaf children. The potential of such an agent in a wide range of language teaching situations is enormous; not only because it can provide extensive rehearsal but also because it can show aspects of speech production that are normally occluded (e.g., a view of tongue position etc).

Providing information about the movement of the speaker's face could also help people in understanding the speaker's message and emotion. For example, it has been demonstrated that people are more accurate in rating speaker's mood with both visual and auditory presentation than with either the one or the other presentation only (Cerrato, Leoni, Falcone, & Bordoni, 1998). Indeed, the understanding other people's minds and intentions involves the analysis and exchange of many subtle signals such as eye direction, facial expression, etc. (e.g., Baron-Cohen, 1998; Brother, 1998).

Indeed, the way that aspects of meaning might be conveyed using audio-visual techniques is an interesting and growing area of research (e.g., Pelachaud, Badler, & Steedman, 1996; Poggi & Pelachaud, 1998).

9. REFERENCES

Barron-Cohen, S. (1995). Mindblindness. Cambridge, Mass. : MIT Press.

Binnie, C. A. Montgomery, A. A. & Jackson P. L. (1974). Auditory and visual contributions to the perception of consonants. Journal of Speech & Hearing Research, **17**, 619-630.

Brothers, L. (1998). Friday's footprint: How society shapes the human mind. Oxford University Press: New York.

Catford J. C., & Pisoni D B. (1970). Auditory versus articulatory training in exotic sounds. Modern Language Journal, **54**, 477-481.

Cerrato, Leoni, Falcone, & Bordoni, (1998). Is it possible to evaluate the contribution of visual information to the process of speech comprehension? AVSP'98: International conference on audio-visual speech processing (Sydney, Australia).

Cohen, M. M., & Massaro, D. W. (1994) Development and experimentation with synthetic visible speech. Behavioral Research Methods and Instrumentation, **26**, 260-265

Cohen, M.M., Beskow, J., & Massaro, D.W. (1998). Recent developments in facial animation: An inside view. AVSP'98: International conference on audio-visual speech processing (Sydney, Australia).

Cox, E.A., Norrix, L.W., & Green, K.P. (1999). The contribution of visual information to on-line sentence processing: Evidence from phoneme monitoring. AVSP,99: International conference on audio-visual speech processing (Santa Cruz, USA).

Cole, R.A. Carmell, T., Conners, P., Macon, M., Wouters, J., de Villiers, J., Tarachow, A., Massaro, D.W., Cohen, M.M., Beskow, J., Yang, J., Meier, U., Waibel, A., Stone, P., Fortier, G., Davis, A., Soland, C. (1998). Intelligent animated agents for interactive language training STiLL - ESCA Workshop on Speech Technology in Language Learning (Stockholm, Sweden).

Davis, C. & Kim, J. (in press). Repeating and remembering foreign language words: Implications for language teaching systems. Artificial Intelligent Review.

Dodd, B. (1977). The role of vision in the perception of speech. Perception, **6**, 31-40.

Flege, J. E., & Hillenbrand, J.(1985). Differential use of temporal cues to the [s-z] contrast by native and non-native speakers of English. Journal of the Acoustical Society of America, **79**, 708-721.

Forster, K.I., & Forster, J.C. (1990). DMDX: Laboratory software for mental chronometry. Tucson, AZ: University of Arizona. http://www.u.arizona.edu/~jforster/dmdx.htm

Markham, D. (1999). Naïve imitation of second-language stimuli: Duration and F0. ICPhS'99: Internation Conference of Phonetic Science, San Francisco.

Massaro, D. W. (1998) Perceiving talking faces: From speech perception to a behavioral principle. Cambridge, Mass.: MIT Press.

Munhall, K. G. & Vatikiotis-Bateson, E. (1998) The moving face during speech communication. In R. Cambell, B. .Dodd, & D. Burnham (Eds.), Hearing by eye, Part 2: The psychology of speechreading and audiovisual speech. London: Taylor & Francis, Psychology Press.

Reisberg, D, McLean, J., & Goldfield, A. (1987). Easy to hear to understand: A lip-reading advantage with intact auditory stimuli. In B.Dodd and R. Cambell, (Eds.) Hearing by eye: The psychology of lip-reading. London: Lawrence Erlbaum, pp 97 -113.

Sanders, D. & Goodrich, S. (1971). The relative contribution of visual and auditory attention. Journal of Speech and Hearing Research, **14**, 154-159.

Strange, W., & Dittman, S. (1984). "Effects of discriminability training on the perception of /l-r/ by Japanese adults learning English. Perception and Psychophysics, **36**, 131-145.

Summerfield. A. Q. (1979). Use of visual information for phonetic perception. Phonetica, **36**, 314-331.

Williams, L. (1979). The modification of speech perception and production in second-language learning. Perception & Psychophysics, **21**, 95-104.

[1] Note: Due to equipment failure, one item in each condition had to be discarded from the analysis.

The authors thank Dr. Cheol-Woo Jo (Changwon National University, Korea) for his assistance in preparing the video files used in this experiment, for his time and use of laboratory space.