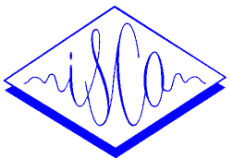

AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Auditory-Visual Speech Perception Development in Japanese and English Speakers

Kaoru Sekiyama[1], *Denis Burnham*[2], *Helen Tam*[2], *Dogu Erdener*[2]

[1]Future University Hakodate, Japan
[2]University of Western Sydney, Australia
sekiyama@fun.ac.jp, d.burnham@uws.edu.au

Abstract

Development of auditory-visual speech perception was investigated in a cross-linguistic developmental framework, using the McGurk effect. Stimuli consisting of /ba/, /da/, and /ga/ utterances were presented to participants who were asked to make syllable identifications on audiovisual (congruent and discrepant), audio-only, and video-only presentations at various signal-to-noise levels. The results of Experiment 1 with 24 adult native speakers of English and 24 of Japanese supported previous reports of a weaker visual influence for Japanese participants. Experiment 2 was a short version of Experiment1 in which 16 Japanese and 14 English language 6-year-olds, and new groups of 24 Japanese and 24 English adults were tested. The results showed that the degree of visual influence was low but statistically equivalent for Japanese and English language 6-year-olds, and that there was a significant increase in visual influence over age for the English but not the Japanese language groups. Nevertheless, both the Japanese and English language groups showed an increase in speechreading performance (in the visual-only condition). It appears that the developmental increase of speechreading performance is related to the increase of the size of the visual influence in the English language participants, whereas such a straightforward relationship is not the case for the Japanese participants.

1. Introduction

It is now well-known that speech perception is an auditory-visual phenomenon, and this can be seen both in degraded speech conditions when visual information about lip and facial movements of a talker compensates for degradation of acoustic information [1], and also in the so-called McGurk effect [2] in which auditory [ba] dubbed onto the face movements for [ga] are perceived as “da” or “tha”. This effect shows that speech perception is an auditory-visual phenomenon even in clear undegraded conditions, and provides a useful tool for investigating various processes of auditory-visual speech processing.

In this paper, the developmental course of auditory-visual speech processing is investigated. The original report of the McGurk effect included both adults and children and showed that children of 3 to 5 years, and 7 to 8 years have less visual influence in their perception of the McGurk effect than do adults [2]. This reduced visual influence in children’s auditory-visual speech perception is robust, as it has been confirmed in later studies [3,4]. On the basis of these age-related findings we may conclude that the *amount* of experience affects auditory-visual speech perception. However, no conclusions can be drawn regarding the *type* of experience, or more specifically, linguistic experience, which is important, because all these studies were conducted with English language stimulus items and English language participants.

Studies with adults suggest that the type of linguistic experience may indeed affect auditory-visual speech processing. Cross-linguistic studies with adults have shown that native speakers of Japanese and Chinese are less subject to visual influence in the McGurk effect than native speakers of English [5, 6, 7, 8, but also see 9]. Together these developmental and cross language results are intriguing: on the one hand we know that the use of visual information increases over age in English language perceivers, and on the other that there is less visual influence for adult speakers of Japanese and Chinese than for adult speakers of English.

In order to obtain detailed knowledge of the role of linguistic experience in the development of auditory-visual speech processing, these two methods, the developmental and the cross-linguistic, need to be combined [10]. Such an approach is used here with Japanese and Australian English children and adults. In Experiment 1 Japanese and Australian English adults were tested using the cross-linguistic framework, in order to ascertain whether the previous reports of inter-language differences were obtained. Following this, Japanese and Australian English 6-year-old children and adults were tested in Experiment 2, which was a shorter more child-friendly version of Experiment 1.

1. Experiment 1

1.1. Method

1.1.1. Subjects

Forty-eight monolingual university students (24 English and 24 Japanese speakers) participated. All participants had normal hearing and normal or corrected-to-normal vision, and were between the ages of 18 and 29. The English participants were tested at MARCS Auditory Laboratories at the University of Western Sydney, Australia, and the Japanese speakers at Future University Hakodate, Japan. At an early stage of the experiment, a bilingual experimenter confirmed that the equipment, procedure, and instructions were equivalent for the two countries.

1.1.1. Stimuli

The stimuli consisted of /ba/, /da/, /ga/ uttered by four talkers (two English and two Japanese talkers, one male

and one female in each language). These talkers were selected in a pilot experiment such that the average intelligibility of auditory and visual speech was approximately the same between the two stimulus languages. The utterances were videotaped, digitized, and edited on a computer for audio-only (A), video-only (V), and audiovisual (AV) stimuli. Half of the AV stimuli were congruent (e.g., auditory /ba/, visual [ba]) and the other half McGurk-type incongruent (e.g., auditory /ba/, visual [ga]). Three kinds of incongruent AV stimuli were created by combining within-talker auditory and visual components (auditory /ba/ with visual [ga], auditory /da/ or /ga/ with visual [ba]). The V stimuli, one each for [ba], [da], and [ga], were created by cutting out the audio track. In the A stimuli, one each for /ba/, /da/, and /ga/, the video of the talking face was replaced by a still face of the talker with the mouth neutrally closed. In total, there were 12 auditory stimuli (3 consonants x 4 talkers), 12 visual stimuli, and 24 audiovisual stimuli (3 auditory consonants x 2 congruity types x 4 talkers).

To obtain a wide range of data, we introduced several levels of auditory intelligibility by adding a band noise (300 Hz – 12000 Hz) with signal-to-noise (SN) ratios of -4, 0, +4, +8, and +12 dB, together with a no-noise condition.

1.1.1. Procedure

The stimuli were presented from computer (Sharp MJ730R) onto a 17-in CRT monitor (Sony 17GS) and through a loud speaker (Aiwa SC-B10). Experimental conditions were blocked depending on the modality (A, V, AV) and the SN ratio of the auditory stimuli (-4, 0, +4, +8, +12 dB, and Clear), and there were 4 repetitions of each stimulus in a block. Each participant was given the AV condition first. Half of the subjects were presented with the stimuli in an AV, A, V, order, and the other half in an AV, V, A order. In the AV and A conditions, the speech was presented at 65 dB and the SN ratios, -4, 0, +4, +8, and +12 dB, were determined by the intensity of the added band noise. There was also a 'Clear' condition in which no noise was added. SN ratios varied across blocks in an increasing manner for half of the subjects, and in a decreasing manner for the remaining subjects.

Within each block, the stimuli were presented in random order. The subjects were asked to watch and listen to each stimulus, decide what they perceived, and press one of three buttons for a "ba," "da," or "ga" response accurately and without delay. The stimuli included the so-called "combination presentations (auditory /da/ or /ga/ combined with visual [ba]" which often produce "combination responses", e.g., "bda" or "bga"). We did not allow such responses, in order to make the response alternatives less confusing for young children in later experiments. It should be noted that some previous studies have shown that these combination stimuli often produce non-combination responses [6, 11]. For example, MacDonald & McGurk [11] results showed 58% "da", 17% "ba", 17% "bda", and 8% "pta" responses for auditory /da/ combined with visual [ba]. Among these responses, those which are not identical to the auditory stimulus (/da/ in this example) can be regarded as 'visually-influenced', and in like fashion, with our restriction on response alternatives here, "ba" responses were anticipated for visually-influenced responses.

After each movie file was played, the last frame remained on the screen until one of the three buttons was pressed. The onset of the next stimulus was 1.5 s after the button press. Responses were made on a game controller, which input to the computer such that the responses were stored. The experiment took an average of 50 minutes per participant.

1.1. Results

Responses were averaged across stimuli. Percent auditorily correct responses in the AV congruent (AV+), AV incongruent (AV-), and audio-only (A) conditions are shown as a function of the SN ratio for English language (E_sub), and Japanese language (J_sub) participants in Figures 1a, and 1b, respectively. As was anticipated, the auditorily correct responses decreased as the SN ratio decreased. At each SN ratio, the degree of augmentation due to visual information (positive effect of visual information) can be seen from the difference between the percent correct in AV+ versus percent correct in A. The size of the visual interference (negative) effect can be seen as the difference between the percent correct A and the percent correct AV-. Combining both the positive and negative visual effects, the total degree of visual influence can be taken as the difference between the percent correct AV+ and the percent correct AV-. This total measure, "the size of visual influence", is used in the analyses below. Although this size of visual influence collapses the distinction between the positive and negative effects of visual information, our preliminary analyses showed that the negative and positive effects followed similar tendencies, that is, the same conclusions can be drawn from the visual influence results as the positive or negative effects in most of the following ANOVAs.

At the highest SN ratio, the maximum size of visual influence was about 33% (Figure 1b). This is because errors from the point of view of the auditory stimulus (here due to the occurrence of the McGurk effect) mainly occurred in the auditory /ba/ visual [ga] stimuli, which comprised just one-third of the AV incongruent stimuli. The English language participants showed the McGurk effect almost 100% of the time for this auditory /ba/ visual [ga] set at the highest SN ratio, resulting in scores of around 33% across the three sets of AV incongruent stimuli. For the other incongruent AV stimuli, visually influenced responses occurred mainly in the lower SN ratios.

The size of the visual influence was submitted to a participant language x stimulus language x SN ratio analysis of variance (ANOVA), with repeated measures on the last two factors. The main effect of participant language was highly significant [$F(1, 46) = 17.518$, $p < .0001$]. Visual influence for English language participants was greater than for Japanese language participants, thus replicating the previously reported inter-language difference [5,6,7,8]. For the repeat factors, there was no significant effect of language of the talkers, but there was a significant main effect of SN ratio [$F(5, 230) = 96.906$, $p < 0.0001$], with greatest visual influence at lower SN ratios. As there were no interactions between these factors the results show that there was less visual influence for the Japanese speakers than the English speakers at each SN ratio.

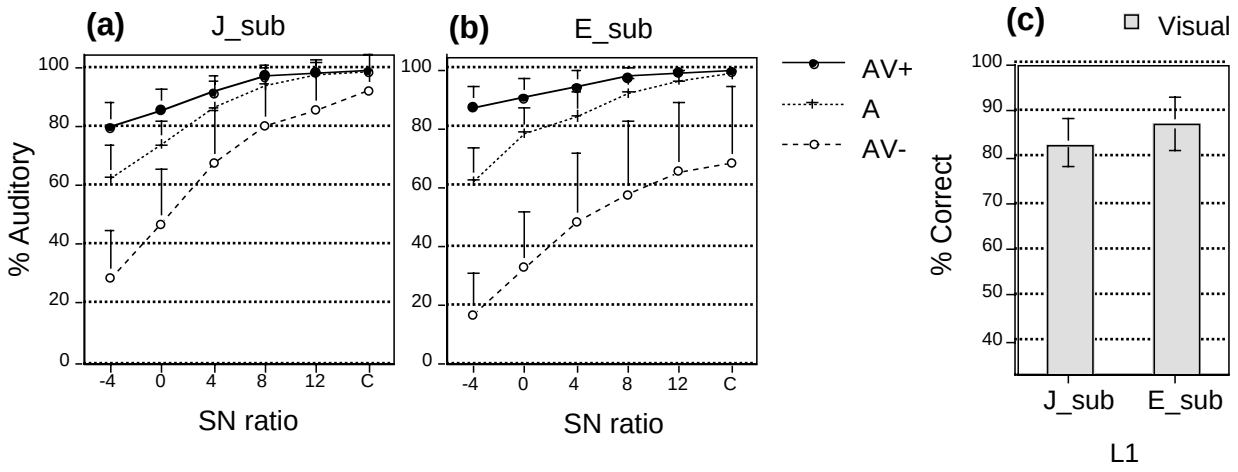

Fig. 1 Response frequency in each group in Exp. 1 (adults). Error bars show standard deviations.

Turning to the unimodal conditions, A-only performance did not differ between the two language groups [F(1, 46)=0.205, p<0.6351]. Figure 1c shows percent of correct responses in the V-only condition (speechreading score). Mean percent correct was 83.0% for the Japanese and 87.2% for the English language participants, and ANOVA revealed that this difference was significant, [F (1, 46)=6.759, p<0.0125]. Thus the weaker visual influence for the Japanese participants might be related to their slightly poorer speechreading performance.

2. Experiment 2

2.1. Method

2.1.1. Subjects

The subjects were newly recruited monolingual adults (24 Japanese and 24 English language participants, age ranged between 18 and 29) and 6-year-old children (16 Japanese and 14 English language participants).

2.1.2. Stimuli and procedure

The same stimuli as in Experiment 1 were used. The procedure was almost the same as in Experiment 1 except that the number of SN ratios was reduced to four, -4, +4, +12 dB, and Clear. Another procedural modification from Experiment 1 was that the number of repetitions of each stimulus was reduced from 4 to 2. Experiment 2 took 20 minutes for adults, and for the 6 year olds, from 30 to 60 minutes (including intermissions in some cases), depending on the child.

2.2. Results

The results for each of the four subject groups (Japanese and English children and adults) are shown in Figure 2 (E: English language participants, 6y: 6 years). Group differences in the size of the visual influence were analyzed via planned contrasts in a 2 (language background) x 2 (age) x 4 (S/N ratio) ANOVA. The main effects of language [F(1,74) = 32.49, p<0.01], and age [F(1,74) = 26.52, p<0.01] were significant, as was their interaction, [F(1,74) = 14.19, p<0.01]. Inspection of simple main effects revealed the source of this interaction: In confirmation of the Experiment 1 results, English adults showed more visual influence than Japanese adults, but for the children this difference was not significant [F(1,46) = 58.37, p<0.01, F(1,28) = 1.51, p>.05, respectively]; and there was an increase in visual influence from 6 years to adulthood for English, but not Japanese speakers [F(1,36) = 38.18, p<0.01, F(1,38) = 1.00, p>.05, respectively]. These results indicate that the size of the visual influence is the same for English and Japanese speakers at age 6, and that there is then a developmental increase in visual influence for the English language, but not for the Japanese language groups.

The noise levels influenced performance: participants showed more visual influence in noise than in the clear [F(1,74) = 135.64, p<0.01]. As noise levels increased the increase in visual influence was more linear for English than Japanese groups [F(1,74) = 4.03, p<0.01], with the Japanese adults showed a markedly more quadratic trend [F(1,74) = 4.34, p<0.01]. Thus the Japanese adults show greater resistance to the effect of noise at the lower noise ratios, 12, and 4 dB than either the English groups or the Japanese children. This underlines the relatively weak visual influence in Japanese adults' auditory-visual speech perception.

With respect to speechreading performance, only age-related differences were found. There was a main effect of age [F(1,74)= 49.50, p<0.01], no main effect of language, and no interaction of these two factors [F(1,74)= 1.03, p>.05, F(1,38)= 0.88, p>.05, respectively], indicating a statistically equivalent increase in speech reading over age for Japanese and English language groups, and no difference in speech reading for either the children or the adults. For adults, this lack of a difference due to language background here is inconsistent with the results in Experiment 1. This difference in results is perhaps due to the large variances here (compare standard deviations in Figure 1c with those in 2e for the adults), which in turn may be due to the reduced number of trials in Experiment 2.

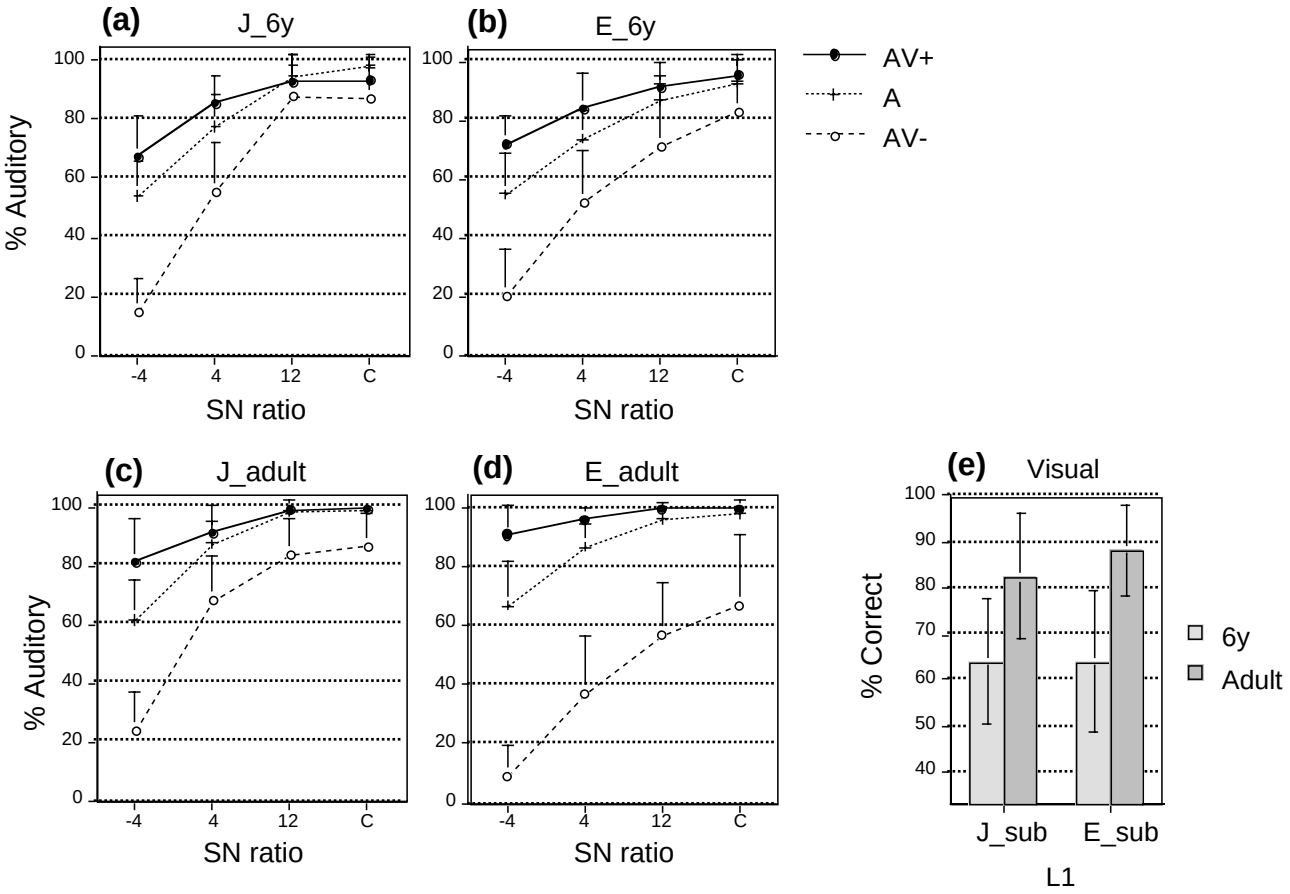

Fig. 2 Response frequency in each group in Exp. 2. Error bars show standard deviations.

The smaller variances in Experiment 1 for speechreading may be due to practice on the preceding AV condition, which contained many more trials in Experiment 1 than here.

The A-only performance was also compared both within languages and within ages. There were no inter-language differences, but the age-related difference was significant, as was the interaction of age and language [$F(1,74)= 2.43$, $p>.05$, $F(1,74)= 39.31$, $p<.01$, $F(1,74)= 4.42$, $p<.01$, respectively]. Simple main effects indicated that auditory identification was better in adults than in 6-year-olds [$F(1, 38)= 9.06$, $p< .01$, $F(1, 36)= 33.66$, $p<.01$, respectively for the Japanese and Australian English participants], and interestingly that the Japanese 6-year-olds showed better auditory identification than their English language counterparts [$F(1,38)= 5.43$, $p< .01$]. This higher A-only performance by the Japanese children is evident in higher SN ratios (see Figure 1a, 1b, Clear, 12dB, 4dB).

3. Discussion

In both Experiments 1 and 2 the Japanese showed less visual influence than did the English language participants, but this difference was not apparent in their 6-year-old counterparts. The results may be summarized as follows: visual influence was equivalently low in Japanese and English language 6-year-olds, and the degree of visual influence increased significantly by adulthood for the English language but not the Japanese language participants. Thus we now know that the inter-language difference between English and Japanese adults observed here and elsewhere develops some time between 6 years and adulthood (20 years).

In English, developmental studies have shown that the visual influence is weaker for children than for adults [2, 3, 4]. The children tested were 3 to 8 years old [2], 2 to 6 years old [3], and 5 to 11 years old [4]. The developmental increase in visual influence seems continuous [4], but it may be concluded from these studies that children under age 8 are remarkably less subject to visual influence than adults. The present results indicate that this developmental increase may occur only in English. It does not appear to be substantial in Japanese.

Previous research has also shown a developmental increase of speechreading performance with children under age 9 being found to be poorer than adults [3, 4]. Our results confirm this tendency for both the English and Japanese language groups. One might attribute the poorer speechreading ability of children to the weaker visual influence in audiovisual speech perception. However, our results show that such a straightforward relationship is true only for the English, and not the Japanese language

participants. Despite appropriately high speechreading performance by Japanese adults, they do not incorporate visual information into perceived speech. What makes Japanese adults behave so?

The causes of this weaker visual influence in Japanese adults are not clear, although several possibilities may be considered. A cultural explanation is that it is related to the Japanese cultural habit of avoiding staring at the person to whom one is talking, though it should be noted that this cultural habit seems to be diminishing as younger generations of Japanese become more Westernized. There are also a number of possible linguistic explanations. Spoken Japanese has less visually identifiable elements than English, and lacks some labio-dentals (e.g., /v/, /_/). It is also said that the mouth movements of Japanese speakers in speech are generally smaller than those of English speakers. In the stimuli used in this study, it was observed that the auditory duration of monosyllables is longer for the English articulations (for the 2 English language stimulus talkers). Similarly, the preparatory movements before the actual articulation took longer for English talkers. These timing differences in everyday experience may play a role in inter-language differences. A final possible explanation arises from our previous and current results showing that the McGurk effect is weakest in native speakers of Chinese, intermediate in the Japanese, and strongest in American and Australian English speakers [8]. Given that Chinese has 4 (Mandarin) or 6 (Cantonese) tones, Japanese 2 pitch-accents, and English has no tones or pitch-accents, which when differentially applied to a particular string of consonants and vowels, changes the meaning of that string, it is distinctly possible that the role of tonal information in the perceiver's native language might be critical with respect to the use of visual speech information. In Experiment 2, we observed that the Japanese 6-year-olds performed better in the A-only condition than their English language counterparts. This could be related to the fact that Japanese children have been exposed to a linguistic environment in which pitch-accent information is of relatively greater importance.

4. Conclusions

The previous finding of a developmental increase in visual influence in audiovisual speech perception from childhood to adulthood for English speakers was repeated here, and was found to be related to an increase of speechreading performance. However, while Japanese 6-year-olds showed the same degree of visual influence as their English language counterparts, there was *no* developmental increase for the Japanese, despite the fact that the Japanese adults' speechreading performance was equivalent to that of the English-speaking adults. One of two possibilities appear to be the case and should be pursued in future research: either there is some aspect of English language development or the English language environment that encourages increased visual processing, and which is lacking in the Japanaese language environment; or there is a general tendency for visual speech processing to increase over age, which is suppressed by some aspect of Japanese language development or the Japanese language environment.

5. Acknowledgements

We are grateful to Yasuko Hayashi-Nagasaki, Thomas Stainsby, Eleanor Gittins, and David Goddard for their efforts in an earlier stage of this project. This study was supported by a Grant-in-Aid of the Japanese Ministry of Education, Sports and Culture (10610070) and a Special Grant from the Science and Technology Agency to KS, an Australian Research Council Large Grant to DB (A79917254), and an Australian Research Council Discovery Grant to DB and KS (A79917254).

6. References

[1] Sumby, W.H. and Pollack, I. "Visual contribution to speech intelligibility in noise", *J. Acoust. Soc. Am.*, 26:212-215, 1954.

[2] McGurk, H. and MacDonald, J. "Hearing lips and seeing voices", *Nature,264:746-748, 1976.*

[3] Massaro, D.W., Thompson, L.A., Barron, B., and Laren, E. "Developmental changes in visual and auditory contribution to speech perception", *J. Exp. Child Psychol., 41:93-113.*

[4] Hockley, N. and Polka, L. "A developmental study of audiovisual speech perception using the McGurk paradigm", *J. Acoust. Soc. Am.*, 96:3309, 1994.

[5] Sekiyama, K. and Tohkura, Y. "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility", *J. Acoust. Soc. Am.*, 90:1797-1805, 1991.

[6] Sekiyama, K. and Tohkura, Y. "Inter-language differences in the influence of visual cues in speech perception", *J. Phonetics*, 21:427-444, 1993.

[7] Kuhl, P.K., Tsuzaki, M., Tohkura, Y., and Meltzoff, A.N. "Human processing of auditory-visual information in speech perception: Potential for multimodal human-machine interfaces", In: Acoust. Soc. Japan (Eds), *Proceedings of the International Conference of Spoken Language Processing*. Tokyo: Acoust. Soc. Jpn., Pp. 539-542, 1994.

[8] Sekiyama, K. "Cultural and linguistic factors in audiovisual speech processing: the McGurk effect in Chinese subjects", *Percept. Psychophys.*, 59:73-80, 1997.

[9] Massaro, D.W., Tsuzaki, M., Cohen, M.M., Gesi, A., and Heredia, R. "Bimodal speech perception: an examination across languages", *J. Phonetics*, 21:445-478, 1993.

[10] Burnham, D. & Sekiyama, K. (in press) Investigating auditory-visual speech perception development using the ontogenetic and differential language methods. In Vatikiotis-Bateson, E., Perrier, P., & Bailly, G. (Ed.). (in preparation). *Advances in auditory-visual speech processing.* Cambridge: MIT Press.

[11] McDonald J. & McGurk, H. Visual influences on speech perception processing. *Perception & Psychophysics*, 24, 253-257, 1978.