Developmental Science

Developmental Science 11:2 (2008), pp 306–320 DOI: 10.1111/j.1467-7687.2008.00677.x

PAPER

Impact of language on development of auditory-visual speech perception

Kaoru Sekiyama[1,3] and Denis Burnham[2]

1. School of Systems Information Science, Future University, Hakodate, Japan
2. MARCS Auditory Laboratories, University of Western Sydney, Australia
3. Division of Cognitive Psychology, Kumamoto University, Japan

Abstract

The McGurk effect paradigm was used to examine the developmental onset of inter-language differences between Japanese and English in auditory-visual speech perception. Participants were asked to identify syllables in audiovisual (with congruent or discrepant auditory and visual components), audio-only, and video-only presentations at various signal-to-noise levels. In Experiment 1 with two groups of adults, native speakers of Japanese and native speakers of English, the results on both percent visually influenced responses and reaction time supported previous reports of a weaker visual influence for Japanese participants. In Experiment 2, an additional three age groups (6, 8, and 11 years) in each language group were tested. The results showed that the degree of visual influence was low and equivalent for Japanese and English language 6-year-olds, and increased over age for English language participants, especially between 6 and 8 years, but remained the same for Japanese participants. This may be related to the fact that English language adults and older children processed visual speech information relatively faster than auditory information whereas no such inter-modal differences were found in the Japanese participants' reaction times.

Introduction

It is now widely understood that speech perception is an auditory-visual phenomenon (e.g. Campbell, Dodd & Burnham, 1998; Dodd & Campbell, 1987; Massaro, 1987, 1998; Stork & Hennecke, 1996; Vatikiotis-Bateson, Perrier & Bailly, in press). This can be seen both in speech conditions when visual information about lip and facial movements compensates for degradation of acoustic information (Sumby & Pollack, 1954), and also in the McGurk effect (McGurk & MacDonald, 1976) in which auditory [ba] dubbed onto the face movements for [ga] are perceived as 'da' or 'tha'. This latter 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 the present study, auditory-visual speech perception is investigated in a developmental plus cross-linguistic framework. Research relevant to each aspect is discussed in turn. 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, despite the fact that the three age groups identified the auditory-only stimuli equally accurately (McGurk & MacDonald, 1976). This reduced visual influence in children's auditory-visual speech perception is robust, as it has been confirmed in later studies with 4- to 6-year-old children, compared with adults (Massaro, 1984; Massaro, Thompson, Barron & Laren, 1986); and also in a gradual developmental increase in visual influence across childhood in 5-, 7-, 9-, and 11-year-olds to adults (Hockley & Polka, 1994).

The developmental increase in visual influence over age in these studies is possibly related to experience in articulating speech sounds. It has been found that preschool children who make substitution errors in articulation are less influenced by visual cues than are children who can correctly produce consonants (Desjardins, Rogers & Werker, 1997): Compared with non-substituter children, substituter children were poorer at speechreading, and had a lower degree of visual influence in auditory-visual speech perception. As non-substituter and substituter children had equivalent auditory-only speech perception, it appears that experience in correctly producing consonants impacts upon the representation of visible speech. This interpretation is supported by a study showing that cerebral palsied adults, lacking in experience of normal speech production, tend to show less visual influence in speech perception under some conditions than non-impaired adults (Siva, Stevens,

Address for correspondence: Kaoru Sekiyama, Division of Cognitive Psychology, Faculty of Letters, Kumamoto University, 2-40-1 Kurokami, Kumamoto, 860-8555, Japan; e-mail: sekiyama@kumamoto-u.ac.jp

Kuhl & Meltzoff, 1995). Thus, it appears that articulation experience affects speechreading and the degree of visual influence in auditory-visual speech perception.

On the basis of these age- and experience-related findings we may conclude that auditory-visual speech perception is affected by the amount of experience with speech input, perhaps largely due to poorer speechreading ability of children (Hockley & Polka, 1994; Massaro *et al*., 1986), and by the degree of proficiency in speech output (Desjardins *et al*., 1997; Siva *et al*., 1995). However, no conclusions can be drawn from these studies regarding the type of 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: Native speakers of Japanese are less subject to visual influence in the McGurk effect than native speakers of English (Kuhl, Tsuzaki, Tohkura & Meltzoff, 1994; Sekiyama & Tohkura, 1991, 1993; but also see Massaro, Tsuzaki, Cohen, Gesi & Heredia, 1993). Moreover, as it has been shown that Japanese speakers more frequently notice the incompatibility between auditory and visual cues in McGurk-type stimuli than do English speakers (Sekiyama, 1994), it appears that the source of this weaker McGurk effect, weaker visual influence, may be a lowered rate of integration of auditory and visual cues compared with English perceivers. If so, then this would suggest that there are differences in auditory-visual integration rather than in perception of visual cues as a result of learning Japanese vs. English. 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 in auditory-visual speech perception for adult speakers of Japanese 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 approaches, the developmental and the cross-linguistic, need to be combined (Burnham & Sekiyama, in press; Burnham, Kitamura & Mattock, 2007). Such an approach is used here with children (6-, 8-, and 11-year-olds) and adults from Japanese and Australian English language backgrounds.

With respect to the stimuli to be used, some previous studies indicate that non-native stimuli induce more visual influence than native stimuli (de Gelder, Bertelson, Vroomen & Chen, 1995; Fuster-Duran, 1996; Grassegger, 1995; Kuhl *et al*., 1994; Sekiyama & Tohkura, 1993). Given these findings, in comparing the two language groups here, speech materials were prepared with multiple talkers – each participant was shown stimuli from two talkers in his/her native language and two talkers in his/her non-native language. Two talkers were used in each language to reduce possible talker differences that are often observed even within a language (e.g. Gagné, Masterson, Munhall, Bilida & Querengesser, 1994; Sekiyama, Braida, Nishino, Hayashi & Tsuyo, 1995).

Two behavioral measures were recorded: Standard identification responses in a three-alternative forced-choice (3AFC) task, and reaction times. Previous research has shown that reaction times (RTs) are longer for mismatching auditory-visual stimuli than matching auditory-visual or auditory-only counterparts (Burnham & Dodd, 2004; Green & Kuhl, 1991; Massaro & Cohen, 1983), presumably indicating the involvement of some process to overcome mismatch during integration. Such a relationship between processing time and stimulus type was also examined to understand the nature of developmental and inter-language differences.

Needless to say, it is not easy to compare auditory-visual speech perception cross-linguistically: Because of different phonological systems between languages, stimuli and response alternatives should be carefully chosen, and the McGurk effect, while useful, may not be the perfect solution for the examination. In this study, such difficulties were compensated for by recording two types of data, response frequency and reaction time.

Experiment 1 employed adult participants only, half with Japanese language and half with English language background. In Experiment 2 children of 6, 8, and 11 years were tested (again half with Japanese language and half with English language background) and the results were compared with those of the adults from Experiment 1.

Experiment 1: Adult participants

The adults' data are presented first as Experiment 1 in order to (a) confirm earlier Japanese vs. English findings in the current experimental setting, and to (b) examine the effects of the experimental factors in detail.

Method

Participants

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 language participants were tested at MARCS Auditory Laboratories at the University of Western Sydney, 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 in the two countries.

Stimuli

The stimuli consisted of [ba], [da], [ga] uttered by four talkers (two English language and two Japanese language 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 equivalent between the two stimulus languages. The utterances were videotaped, digitized, and edited on computer to produce audio-only (AO), video-only (VO), and audiovisual (AV) stimuli. Video digitizing was done at 29.97 frames/s in 640 × 480 pixels, and audio digitizing was at 32 kHz in 16 bit; each stimulus was created as a 2.3 s movie of a monosyllabic utterance. The duration of acoustic speech signals in each movie was approximately 330 ms on average. The movie file was edited with frame unit accuracy (33.3 ms), and the sound portion was additionally edited with 1 ms accuracy (for more details, see Sekiyama, Kanno, Miura & Sugita, 2003). Half of the AV stimuli were matching (e.g. auditory [ba], visual [ba], i.e. AbVb) and the other half McGurk-type mismatching (e.g. auditory [ba], visual [ga], i.e. AbVg). Three kinds of mismatching AV stimuli were created by combining within-talker auditory and visual components (AbVg, AdVb, AgVb). The VO stimuli, one each for [ba], [da], and [ga], were created by cutting out the audio track. In the AO 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 × 4 talkers), 12 visual stimuli (3 consonants × 4 talkers), and 24 audiovisual stimuli (3 auditory consonants × 2 congruity types [matching/mismatching: AbVb/AbVg; AdVd/AdVb; AgVg/AgVb] × 4 talkers).

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

Procedure

The stimuli were presented from computer (Sharp MJ730R) onto a 17-in CRT monitor (Sony 17GS) and through a loudspeaker (Aiwa SC-B10). Experimental conditions were blocked depending on the modality (AO, VO, AV) and the SN ratio of the auditory stimuli (–4, +4, +12 dB, and Clear), and there were two repetitions of each stimulus in a block (2 × 12 stimuli = 24 trials in each block of the AO and VO conditions, and 2 × 24 stimuli = 48 trials for each block of the AV condition). Each participant was presented with the AV condition first. Half of the subjects were presented with the stimuli in an AV, AO, VO order, and the other half in an AV, VO, AO order. In the AV and AO conditions, the speech was presented at 65 dB and the SN ratios, −4, +4, 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. This block design was employed in order to maintain participants' interest between blocks by introducing changes, especially bearing in mind that children, with reduced attention span compared to adults, were to be tested in Experiment 2.

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' (AdVb, AgVb, which sometimes produce 'combination responses', e.g. 'bda' or 'bga'). We did not allow for such responses because we intended to measure reaction times for which the number of alternatives should be two or three. Limiting the response alternatives was appropriate in order to make the task less confusing for young children in Experiment 2. Although combination responses were reported in the original report by McGurk and MacDonald (1976) and have been replicated especially in a forced-choice paradigm (e.g. Kuhl *et al*., 1994), it should be noted that non-combination responses are often major response categories in an open-choice paradigm (MacDonald & McGurk, 1978; Sekiyama & Tohkura, 1993). Based on these previous results, it was anticipated that responses that were not identical to the auditory stimulus could be regarded as 'visually influenced'.

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.

Before starting the first block of each of the AV, AO, and VO conditions, six practice trials were given to each participant. Excluding these practice trials, the total number of trials per participant was 312 (24 trials × 4 SN ratios for AO, 24 trials for VO, and 48 trials × 4 SN ratios for AV conditions). The experiment took an average of 25 minutes per participant.

Results and discussion

Before analyzing the data, we examined whether the order of SN ratio (increasing or decreasing) made any differences to the results. An analysis of variance (ANOVA) was carried out on the percent correct data in the AO condition with two between-subjects (language backgrounds, and order of SN ratio) and two within-subjects (native vs. non-native stimuli, and SN ratios) factors. The main effect of the order of SN ratio was not significant [$F(1, 44) = 0.982$, $p < .327$]. Thus, this factor was simply collapsed in the following analyses.

Response frequency

For each participant, responses were averaged across syllables and the two talkers within a stimulus language. In order to provide a consistent metric against which to judge performance cross-linguistically, and given that in auditory-visual trials there is no objectively correct

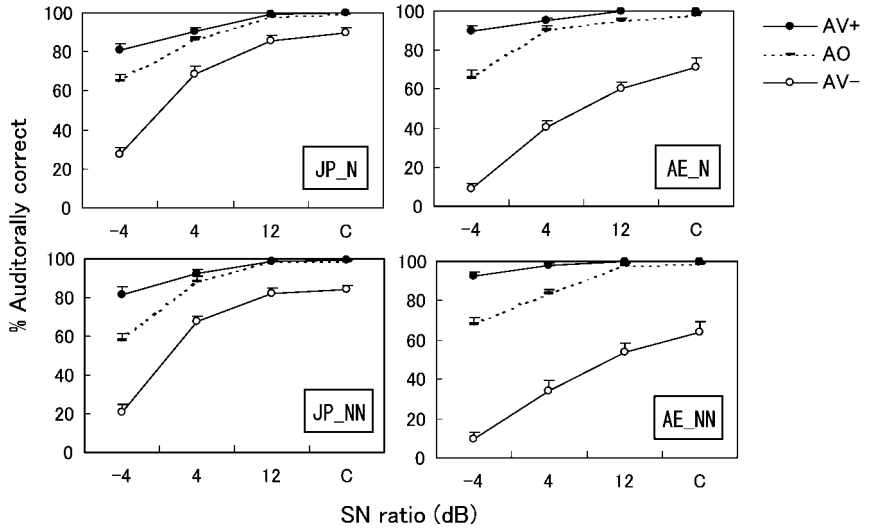

Figure 1 *For adults in Experiment 1, percent auditorally correct responses in the matching AV (AV+), mismatching AV (AV–), and AO conditions as a function of the SN ratio of the auditory stimulus. Data are shown for the Japanese participants with native stimuli (JP_N) and with non-native stimuli (JP_NN) and for the Australian English participants with native stimuli (AE_N) and with non-native stimuli (AE_NN). The bars show standard errors.*

response, both auditory-only and auditory-visual responses were judged in relation to the auditory component of the stimulus, so that the relative degree of visual influence could be compared. This follows an established method of reporting such results (e.g. Massaro & Cohen, 1996; Sekiyama, 1997). Percent correct responses in the matching AV (AV+), mistatching AV (AV–), and AO conditions are shown as a function of the SN ratio in Figure 1. The results are shown for the two language groups (Japanese vs. Australian English; JP vs. AE) and two stimulus languages (native vs. non-native; N vs. NN). As anticipated, the auditorally 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 AO; and the degree of the visual interference (negative) effect can be seen as the difference between the percent correct AO 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' (shown in Figure 2), 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 revealed that the negative and positive effects showed similar tendencies.

The size of the visual influence was submitted to a 2 language groups (Japanese/English) × 2 stimulus languages (native vs. non-native) × 4 SN ratios ANOVA, with repeated measures on the last two factors, and with planned comparisons within levels of the factors (Harris, 1994; planned comparisons were used throughout this

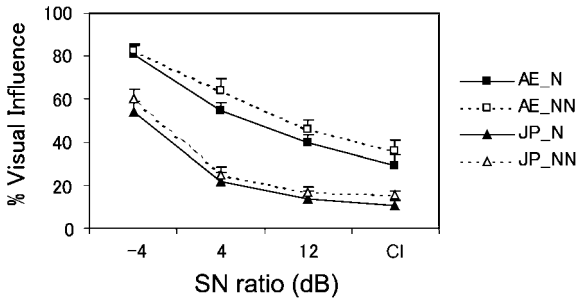

Figure 2 *For adults in Experiment 1, percent visual influence (difference between AV+ and AV– in percent auditorally correct responses) as a function of the SN ratio of the auditory stimulus. Data are shown for the Australian English participants with native stimuli (AE_N) and with non-native stimuli (AE_NN) and for the Japanese participants with native stimuli (JP_N) and with non-native stimuli (JP_NN). The bars show standard errors.*

study). The main effect of language group was highly significant [$F(1, 46) = 53.777$, $p < .001$]. Visual influence for English language participants was greater than for Japanese language participants, thus confirming previously reported inter-language differences (Kuhl *et al.*, 1994; Sekiyama & Tohkura, 1991, 1993). The main effect of the stimulus languages was significant [$F(1, 46) = 11.360$, $p < .01$], but there was no language group × stimulus language interaction, indicating that both groups showed greater visual influence for non-native stimuli. As shown in Figure 2, this non-native effect was relatively small compared with the differences between the two language groups. The main effect of noise [$F(1, 46) = 86.415$, $p < .001$] with significant linear [$F(1, 46) = 179.944$,

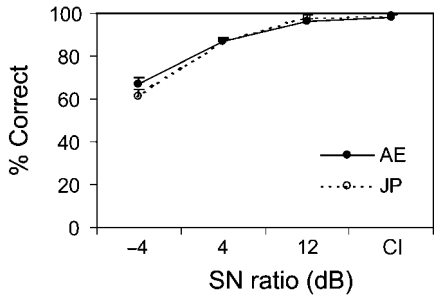

Figure 3 *For adults in Experiment 1, percent correct responses in the AO condition as a function of SN ratio. Data are shown for the Australian English (AE) and Japanese (JP) participants. The bars show standard errors.*

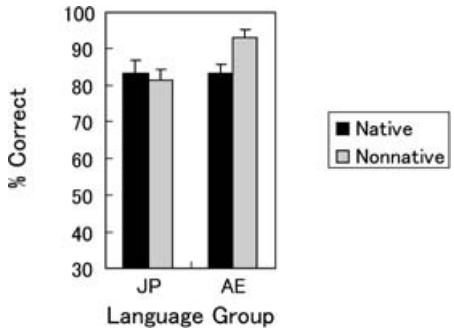

Figure 4 *For adults in Experiment 1, percent correct responses in the VO condition for the Japanese (JP) and Australian English (AE) participants perceiveing native and non-native stimuli. The bars show standard errors.*

$p < .001$] and quadratic [$F(1, 46) = 14.161$, $p < .001$] trend components showed exponentially increasing visual influence at lower SN ratios; and a significant language group × quadratic SN ratio trend interaction [$F(1, 46) = 5.663$, $p < .05$] indicated that the sharpest increase in visual influence over SN ratio was for the Japanese participants between the two lowest SN ratios.

Turning to the unimodal conditions, a 2 language groups × 2 stimulus languages × 4 noise levels ANOVA found no difference in AO performance between the two language groups [$F(1, 46) = 0.268$] (see Figure 3), or the stimulus languages [$F(1, 46) = 0.603$], nor was there any interaction between stimulus languages and language groups. There were significant main effects and linear and quadratic trends of SN ratio [$F(1, 46) = 288.996$, $p < .001$; $F(1, 46) = 216.286$, $p < .001$, $F(1, 46) = 22.120$, $p < .001$, respectively], indicating poorer auditory performance at lower SN ratios. These results indicate that audio-only performance of both language groups depended only on SN ratios.

Figure 4 shows that percent correct responses in the VO condition (speechreading score) was high for both language groups (82.3% for Japanese, and 88.1% for English language participants). ANOVA (2 language groups × 2 stimulus languages) revealed no significant main effects of language groups [$F(1, 46) = 2.929$, $p < .10$] or stimulus languages [$F(1, 46) = 3.680$, $p < .10$], but there was a significant interaction of the two factors [$F(1, 46) = 8.680$, $p < .01$] mainly due to better speechreading for non-native than native stimuli by the English language participants. This indicates that the Japanese language stimuli (native for the Japanese participants, non-native for the English language participants) were visibly more informative than those of the English talkers, at least to the English language participants.

Collectively, these response frequency data show a weaker visual influence in auditory-visual speech perception in Japanese language adults than English language adults, replicating previous findings (Kuhl *et al.*, 1994; Sekiyama & Tohkura, 1991, 1993). Visual influence was consistently stronger for non-native than native speech stimuli, but this non-native effect was relatively small compared with the differences between the language groups (Figure 2). There were no inter-language differences in AO performance. The inter-language differences in the size of the visual influence may be partly related to the differences in the VO performance, but only so when Japanese stimuli were presented.

Reaction time (RT)

Figure 5 shows RTs (averaged across stimulus languages) for the AV conditions as a function of the SN ratio,

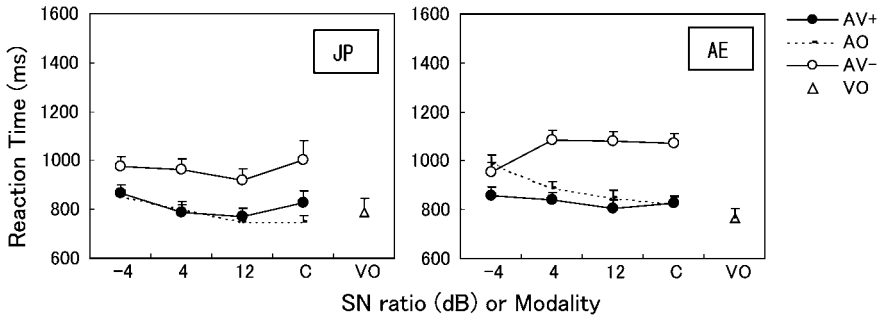

Figure 5 *For adults in Experiment 1, mean reaction time in the AV–, AO, AV+ conditions as a function of the SN ratio of auditory stimulus, together with that in the VO condition. Data are shown for the Japanese (JP) and Australian English (AE) participants. The bars show standard errors.*

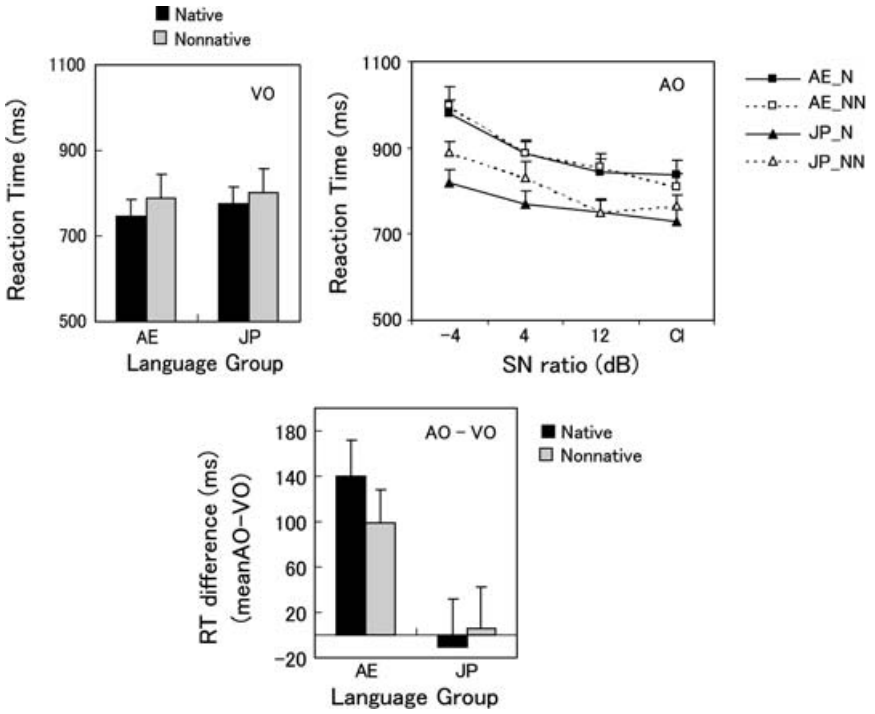

Figure 6 *For adults in Experiment 1, mean reaction time in unimodal conditions. (a) RTs in the VO condition for the Australian English (AE) and Japanese (JP) participants perceiving native and non-native stimuli. (b) RTs in the AO condition for AE participants perceiving native (AE_N) or non-native (AE_N) stimuli and for JP participants with native (JP_N) or with non-native stimuli (JP_NN). (c) RT difference between AO (averaged across SN ratios) and VO conditions. The bars show standard errors.*

together with those for the AO and VO conditions. For simplicity, the results are combined across native and non-native stimuli here. In the AV conditions, RTs were obviously shorter for the matching (AV+) than mismatching (AV−) condition. The longer RTs in the mismatching condition replicate previous results (Green & Kuhl, 1991; Massaro & Cohen, 1983), probably indicating some process involved in integrating mismatching auditory and visual information. To investigate this further, the effect of AV interaction measured by the RT differences between AV− (open circles) and AV+ (filled circles) conditions were calculated for each participant, and submitted to an ANOVA (2 language groups × 2 stimulus languages × 4 SN ratios). There was a significant main effect of language group [$F(1, 46) = 6.366$, $p < .05$], indicating that the effect of AV interaction was larger for the English language participants than for the Japanese participants. The main effect of the stimulus languages was not significant [$F(1, 46) = 0.736$], but the language group × stimulus language interaction was significant [$F(1, 46) = 6.312$, $p < .05$], indicating slightly higher AV interaction for the non-native than native stimuli in Japanese, but not English language participants. Indicative of the smaller RT differences at the lower SN ratios, the linear and quadratic trends of the SN ratio were significant [$F(1, 46) = 22.608$, $p < .001$; $F(1, 46) = 9.215$, $p < .01$, respectively], showing that the mismatch is less easily detected when auditory speech is less intelligible. Additionally a language group × linear trend for SN ratio trend interaction [$F(1, 46) = 9.320$, $p < .01$] indicates that the decrease in RT differences at lower SN ratios is more prominent for English than for Japanese language participants.

With respect to the relationship between the AV+ and AO conditions (filled circles vs. dotted lines in Figure 5), noticeable inter-language differences were found: the additional matching visual cues improve identification latency for the English, but not the Japanese participants. This was tested in an ANOVA (2 language groups × 2 stimulus languages × 4 SN ratios) using the RT differences between the AO and AV+ conditions at each SN ratio for each subject. The main effect of language group [$F(1, 46) = 9.123$, $p < .01$] revealed greater facilitation of identification by visual information for the English language participants, indicating greater auditory-visual integration in speech perception by English language participants, compared with relatively more separate auditory and visual processing in Japanese language participants.

Turning to the unimodal conditions, the upper panels in Figure 6 show mean RTs in the VO and AO conditions. ANOVA (2 language group × 2 stimulus language) for the VO RTs revealed a significant effect of stimulus language [$F(1, 46) = 5.770$, $p < .05$], but no significant language background group [$F(1, 46) = 0.116$] or stimulus language × group interaction [$F(1, 46) = 0.341$]. Thus, VO response latencies were about the same for the two language groups, but native stimuli were speechread more quickly than non-native stimuli (Figure 6a).

ANOVA (2 language groups × 2 stimulus languages × 4 SN ratios) for the AO RTs revealed significant main effects of language background group [$F(1, 46) = 7.424$, $p < .01$] and stimulus language [$F(1, 46) = 7.200$, $p < .01$], indicating that AO responses were faster in Japanese than English language participants (the mean group difference was as great as 100 ms) and faster for native than non-native stimuli. Additionally, the significant language background group × stimulus language interaction [$F(1, 46) = 6.343$, $p < .05$] indicates that the native language advantage occurred mainly for the Japanese participants. There was also a significant main effect [$F(1, 46) = 27.149$, $p < .001$] and linear trend for SN ratio [$F(1, 46) = 54.893$, $p < .001$], indicating that the AO responses were slower at lower SN ratios, but this did not interact with language background group or stimulus language (Figure 6b).

In audiovisual integration, it seems that the relationship between processing speed for auditory vs. visual judgments plays a crucial role. In normal speech, preparatory mouth movements always precede auditory speech. Given this, plus the faster physical speed of optic than acoustic wave transmission, there is a possibility that visual speech cues are processed ahead of auditory speech cues. If so, then visual information may serve to prime the subsequent auditory judgment. With this possibility in mind, RT differences between AO (averaged across SN ratios) and VO were calculated for each participant and submitted to a 2 language group × 2 stimulus language ANOVA (Figure 6c). The mean AO minus VO differences for native and non-native stimuli were 140 ms and 99 ms for the English language participants, and −11 ms and 6 ms for the Japanese participants. There was no significant effect of stimulus language [$F(1, 46) = 0.770$], but a significant main effect of the language group [$F(1, 46) = 6.435$, $p < .05$] and an interaction of the two factors [$F(1, 46) = 4.313$, $p < .05$]. These results indicate that the AO minus VO effect, the degree to which responses were faster to VO than AO, was significantly greater for English language than Japanese participants.

Taken together, the RT data show that AV interaction was stronger in the English than Japanese language participants: In the AV+ condition there was greater facilitation of speech perception by visual information for the English language participants, and in the AV– condition visual interference was also stronger for English language participants. In the unimodal conditions, AO responses were faster in Japanese than English language participants, which may be one of the reasons why they are less influenced by visual speech cues. Examination of RT differences between the AO and VO conditions showed there was a large visual processing advantage for English language participants, but no such intermodal differences for the Japanese participants. Assuming that such unimodal processing speeds also occur in the AV conditions, then the temporal precedence of visual judgments in English language participants could increase the probability of visual influence on the subsequent auditory judgment. Because Japanese participants were much faster in AO judgments than English language participants, such visual precedence was not the case in the Japanese. Thus the differences between Japanese and English language participants in response frequency data (Figure 1) may be seen as a result of a different *time course* in their auditory, visual, and auditory-visual speech processing: the weaker visual influence for the Japanese participants could be due to the auditory and visual information being available at a similar point in time for them (due to their faster AO latencies, Figure 6), but visual information at an earlier point in time than auditory information for English language participants.

Experiment 2: Cross-linguistic developmental comparisons

Method

Participants

Besides the 48 adult monolinguals described in Experiment 1 (24 English and 24 Japanese speakers, between the ages of 18 and 29), three age groups of children participated (6-, 8-, and 11-year-olds). In each age group, there were 16 English and 16 Japanese monolingual children, thus the total number of participants was 144 including the adults. The mean ages of these groups were 6.66 years ($SD = 0.28$ years), 8.40 years ($SD = 0.27$ years), and 11.32 years ($SD = 0.26$ years) for English speaking children and 6.55 years ($SD = 0.34$ years), 8.54 years ($SD = 0.35$ years), and 11.63 years ($SD = 0.20$ years) for Japanese speaking children. All participants reported normal hearing and normal or corrected-to-normal vision. The English language participants were tested at MARCS Auditory Laboratories at the University of Western Sydney, Australia or at nearby elementary schools, and the Japanese speakers at Future University Hakodate, Japan or a nearby elementary school.

Stimuli and procedure

The stimuli and procedure were identical to those in Experiment 1. While the adults took about 25 min to complete the experiment, younger children needed more time: For 6-year-olds, one intermission was usually necessary and the experiment took a maximum of an hour including this intermission.

Results and discussion

Response frequency

As in Experiment 1, each participant's responses were averaged across syllables and the two talkers within each

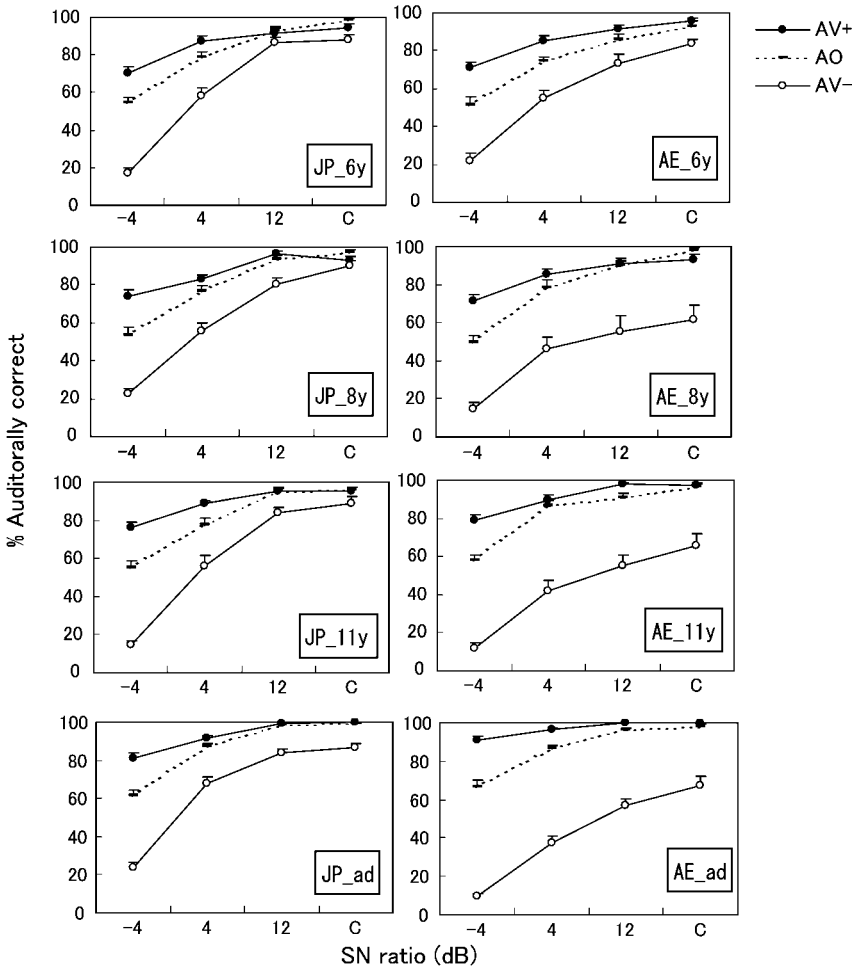

Figure 7 *In Experiment 2, percent auditorally correct responses in the AV+, AV–, and AO conditions as a function of SN ratio of auditory stimulus. Data are shown for each language (JP vs. AE) and age (6 years, 8 years, 11 years, and adults) group. The bars show standard errors.*

stimulus language. Figure 7 shows the percent correct responses in the AV+, AV–, and AO conditions as a function of SN ratio combined across stimulus languages (native and non-native). The size of visual influence was examined via the difference in percent auditorally correct responses between the AV+ and the AV– conditions (differences between the filled and open circles in Figure 7), which can be seen more clearly averaged across SN ratios in Figure 8a. As can be seen, at 6 years the size of visual influence was relatively small and equivalent for Japanese and English language children. For the English language participants, visual influence increased between 6 and 8 years, whereas it remained almost constant over age for the Japanese participants.

ANOVA of the degree of visual influence (2 language groups × 4 age groups × 2 stimulus languages × 4 SN ratios) showed a main effect of language group [$F(1, 136) = 46.991$, $p < .001$], indicating greater visual influence for English language participants. Significant age-related main effects and interactions with language group were found for (i) children vs. adults [$F(1, 136) = 9.303$, $p < .01$, and $F(1, 136) = 5.728$, $p < .05$, respectively], and (ii) 6 years vs. 8 and 11 years [$F(1, 136) = 10.813$, $p < .01$ and $F(1, 136) = 5.178$, $p < .05$, respectively], indicating that the degree of visual influence increased between 6 and 8 years, but only for the English language children.

There was a significant main effect of stimulus language [$F(1, 136 = 6.905, p < .01)$], and a significant stimulus language × 11 years vs. adults interaction [$F(1, 136) = 4.164$, $p < .05$], indicating that over and above the greater visual influence for non-native stimuli, the non-native visual influence advantage increased between 11 years and adulthood.

Finally, regarding the SN ratio factor, there were significant main effects of noise vs. clear [$F(1, 136) = 219.728$, $p < .001$], and the linear and quadratic trends of SN ratio [$F(1, 136) = 364.849$, $p < .001$; and $F(1, 136) = 15.385$, $p < .001$, respectively], indicating that the size of visual influence was larger at lower SN ratios.

In the VO condition (Figure 8b), speechreading scores increased over age in both language groups. An ANOVA (2 language groups × 4 age groups × 2 stimulus languages) revealed significant age-related effects for children vs. adults [$F(1, 136) = 32.781$, $p < .001$], 6 vs. 8 and 11 years

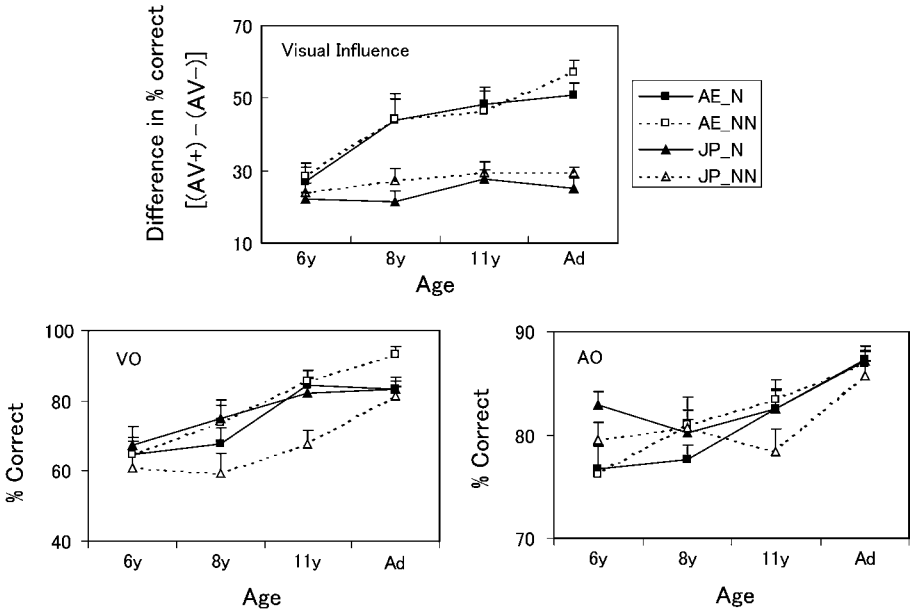

Figure 8 *In Experiment 2, as a function of age: (a) Percent visual influence (difference between AV+ and AV– in percent auditorally correct responses); (b) Percent correct responses in the VO condition; (c) Percent correct responses in the AO condition. The data are shown for each language background and stimulus language. The bars show standard errors.*

[$F(1, 136) = 11.315$, $p < .01$], and 8 vs. 11 years [$F(1, 136) = 9.842$, $p < .01$], indicating an increase in speechreading performance up to 11 years. There was a significant main effect of language group [$F(1, 136) = 4.441$, $p < .05$], indicating that the speechreading performance was generally better in English than in Japanese language participants. Given that this inter-language difference was not significant when tested only with the adult participants (Experiment 1), individual tests at each age were conducted here in separate 2 language group × 2 stimulus language ANOVAs, revealing that the main effect of the language group was significant only at 11 years [for 6, 8, and 11 years, $F(1, 30) = 0.011$; $F(1, 30) = 0.308$; $F(1, 30) = 6.217$, $p < .05$, respectively]. Thus, it seems that inter-language differences in speechreading accuracy, if any, do not emerge until 11 years.

Concerning stimulus language, VO performance did not differ overall between native and non-native stimuli [$F(1, 136) = 3.088$, $p < .10$], but a language group × stimulus language interaction [$F(1, 136) = 20.689$, $p < .001$] indicated that native stimuli were better speechread by Japanese participants whereas non-native stimuli were better speechread by English language participants. This suggests that the Japanese talker stimuli were visibly more intelligible than those presented by the English talkers, presumably due to some individual talker differences. (In our pilot experiments, it was found that the advantage for native stimuli is relatively small compared to individual talker differences.) Significant age × stimulus language interactions were found only for children vs. adults × stimulus language, and 11 years vs. adults × stimulus language [$F(1, 136) = 7.493$, $p < .01$; $F(1, 136) = 6.620$, $p < .05$, respectively], indicating that the intelligibility differences among talkers are more reliably detectable at later stages of development.

In the AO condition (Figure 8c), accuracy of auditory speech perception generally increased over age in both language groups. ANOVA (2 language groups × 4 age groups × 2 stimulus languages × 4 SN ratios) showed significant age-related effects for children vs. adults [$F(1, 136) = 38.803$, $p < .001$], and 11 years vs. adults [$F(1, 136) = 13.449$, $p < .001$], indicating developmental improvement especially between 11 years and adulthood. In younger children, however, a language group × 6 vs. 8 and 11 years interaction was found [$F(1, 132) = 4.377$, $p < .05$], indicating better AO performance in Japanese than English language children at 6 years, but not at later ages. This early auditory superiority in the Japanese may be related to a greater dependence on auditory cues and less dependence on visual cues in audiovisual speech perception at older ages.

There was no main effect of stimulus language in the AO condition [$F(1, 136) = 0.767$], but there was a significant language group × stimulus language interaction [$F(1, 136) = 4.113$, $p < .05$], indicating an advantage for native AO stimuli only for the Japanese participants. Auditory speech perception was poorer at lower SN ratios (Figure 7), as indicated by significant effects of noise vs. clear [$F(1, 136) = 987.795$, $p < .001$], linear and quadratic SN ratio trends [$F(1, 136) = 771.207$, $p < .001$; $F(1, 136) = 38.174$, $p < .001$, respectively].

Reaction time (RT)

Figure 9 shows RTs (averaged across stimulus languages) for the AV conditions as a function of SN ratio, together with

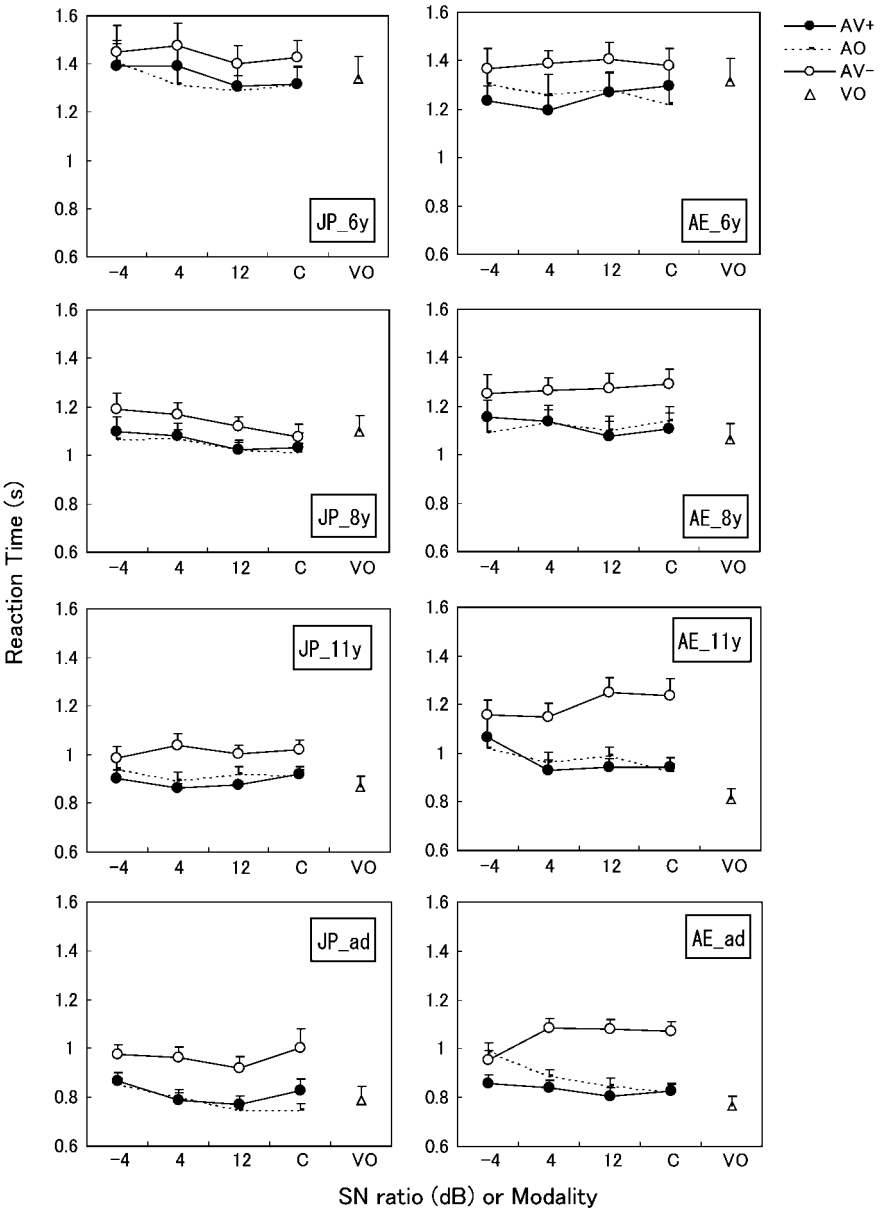

Figure 9 *In Experiment 2, mean reaction time in the AV–, AO, AV+ conditions as a function of SN ratio of the auditory stimulus, together with that in the VO condition. Data are shown for each age and language group. The bars show standard errors.*

those for the AO and VO conditions. In general, RTs for the AV+ condition are less than for the AV– condition across ages and language groups, but the size of this RT difference is smaller in children than in adults. As in Experiment 1, RT differences between AV– (open circles) and AV+ (filled circles) conditions were calculated for each participant as a measure of the effect of AV interaction on speed of processing, and submitted to ANOVA (2 language groups × 4 age groups × 2 stimulus languages × 4 SN ratios). As can be clearly seen in Figure 10a, the main effect of language group was significant [$F(1, 136) = 25.363$, $p < .001$], indicating greater AV interaction for the English than the Japanese language participants. Significant children vs. adults and 8 vs. 11 years effects [$F(1, 136) = 9.523$, $p < .01$; $F(1, 136) = 6.896$, $p < .01$, respectively], along with no significant interactions between language group and age showed that there were equivalent age-related increases in visual effects on RT after 8 years, irrespective of language group. The main effect of the stimulus languages was not significant, nor was the interaction between language group and the stimulus language. There were a few significant noise-level-related terms, but they are not scrutinized here.

The lower panels in Figure 10 show RTs in the VO (Figure 10b) and AO (Figure 10c) conditions. ANOVA (2 language groups × 4 age groups × 2 stimulus languages) for the VO RTs revealed that only age-related terms were significant: Children vs. adults [$F(1, 136) = 40.022$, $p < .001$], 6 vs. 8 and 11 years [$F(1, 136) = 41.412$, $p < .001$], and 8 vs. 11 years [$F(1, 136) = 11.617$, $p < .001$]. Thus,

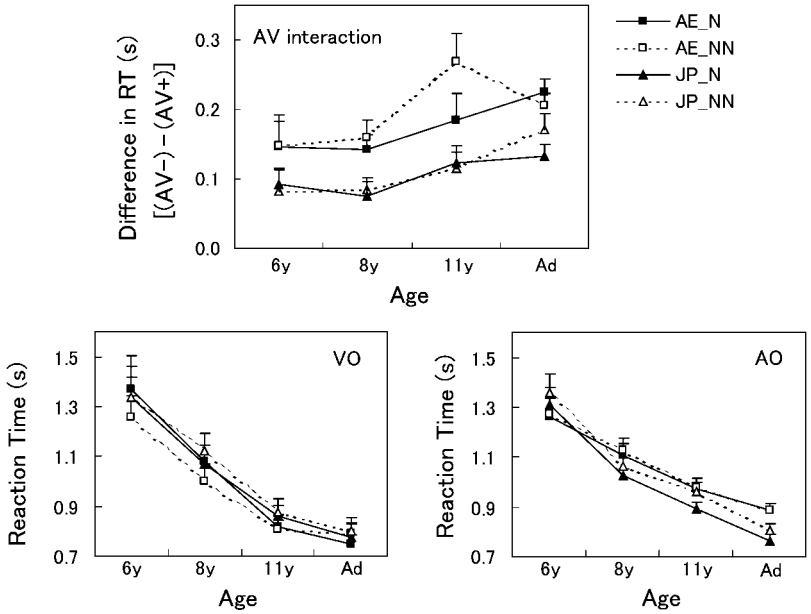

Figure 10 *In Experiment 2, as a function of age: (a) Visual influence in reaction times (difference between AV– and AV+ in RT); (b) RTs in the VO condition; (c) RTs in the AO condition. The data are shown for each language background and stimulus language. The bars show standard errors.*

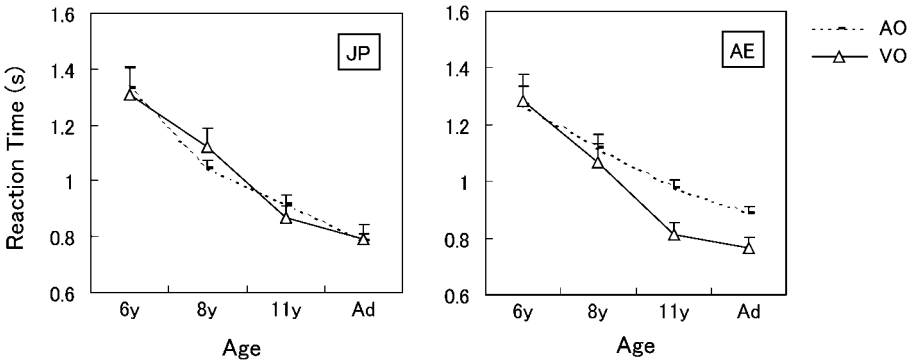

Figure 11 *In Experiment 2, mean RT in the VO and AO conditions as a function of age for each language group. The bars show standard errors.*

VO response latencies decreased over age up to 11 years, irrespective of language group or stimulus language (Figure 10b).

ANOVA (2 language groups × 4 age groups × 2 stimulus languages × 4 SN ratios) for the AO RTs revealed that AO response latencies gradually decreased over age: Significant effects were found in children vs. adults [$F(1, 136) = 74.081$, $p < .001$], 6 vs. 8 and 11 years [$F(1, 136) = 53.314$, $p < .001$], 8 vs. 11 years [$F(1, 136) = 8.150$, $p < .01$], and 11 years vs. adults [$F(1, 136) = 7.808$, $p < .01$]. The main effect of stimulus language was significant [$F(1, 136) = 13.3022$, $p < .001$], indicating shorter RTs for native stimuli. Although the two language groups did not differ on average [$F(1, 136) = 1.594$], the language group × stimulus language interaction was significant [$F(1, 136) = 6.600$, $p < .05$], indicating that the advantage of the native stimuli was mainly for the Japanese participants (Figure 10c).

Figure 11 shows the relationship between RTs in the AO and VO conditions for each language group. As in Experiment 1, RT differences between AO (averaged across SN ratios) and VO were calculated for each participant and submitted to ANOVA (2 language groups × 2 stimulus languages). The main effect of language group was significant [$F(1, 136) = 6.174$, $p < .05$], indicating that the visual over auditory identification advantage was greater for English language than Japanese language participants. Over age, significant effects of 6 vs. 8 and 11 years, and 8 vs. 11 years [$F(1, 136) = 4.798$, $p < .05$; $F(1, 136) = 4.004$, $p < .05$, respectively] indicated a developmental increase in visual over auditory advantage, and a significant language group × 6 vs. 8 and 11 years interaction [$F(1, 136) = 3.957$, $p < .05$] showed that this increase especially occurred for English language participants between 6 and 8 years. This is consistent with the notion that English, but not Japanese, language

participants' speech perception becomes more visually tuned after 6 years in terms of response frequency (see Figure 8a for the greater visual influence in the AV conditions for the English language children after 6 years).

Concerning stimulus language, there was a significantly larger AO – VO difference for non-native stimuli [$F(1, 136) = 4.576$, $p < .05$], perhaps due to the Japanese participants' faster auditory identification for the native than non-native stimuli (Figure 10c).

Considering the RT results as a whole, the larger increase of AO – VO difference over age for the English language than the Japanese participants is in accord with increasing degree of visual influence in the AV conditions over age for the English language participants, suggesting that the increasing temporal precedence of visual over auditory judgments in older English language speakers causes the greater visual influence.

General discussion

In this study, two experiments were conducted, one with adults only and one comparing adults and children. In Experiment 1, Japanese adults showed weaker visual influence than the English language adults. This was found not only in the response frequency data (Figure 1 and Figure 2), thus confirming the results of previous studies, but also in RT data (Figure 5): There was less effect of AV interaction on speed of processing for Japanese than English language participants thus extending the results of previous studies. These inter-language differences in visual influence are not attributable to unimodal accuracy in either the AO or VO conditions because Japanese and English language participants were equally accurate on both (Figure 3 and Figure 4), but may be related to processing speed: Japanese adults were 100 ms faster than the English language adults in the AO condition, and accordingly, English language adults were relatively much faster in the VO than in the AO condition (Figure 6). If such unimodal processing speeds are also the case in the AV conditions, then the temporal precedence of visual judgments for English language adults could increase the probability of the visual influence on the subsequent auditory judgment. Thus it can be concluded from the adult data that there is no essential difference in the *availability* of auditory and visual information for Japanese and English participants, but rather in the relative *time course* for the availability of auditory and visual unimodal information.

Besides studies conducted in our laboratories, a few studies have also found a weaker visual influence for Japanese than English language adults in auditory-visual speech perception: Kuhl *et al.* (1994) found that Japanese participants showed a weaker visual influence than American English language adults especially when presented with native stimuli, and the data by Massaro *et al.* (1993) indicated that the visual influence of artificial speech stimuli was weaker for Japanese participants than for speakers of Spanish or American English (the group difference was significant by ANOVA although these authors' emphasis was on another point). The present studies provide the first evidence for another aspect of the inter-language difference, that is, the relative *time course* for the availability of auditory and visual unimodal information. With these additional data, the Japanese–English inter-language difference has been firmly confirmed.

In Experiment 2, when the adult data were combined with those of children, it was found that this inter-language difference emerges between 6 and 8 years. The 6-year-olds in both language groups showed a relatively weak visual influence and no group differences in response frequency, but between 6 and 8 years visual influence increased for English but not Japanese language children, resulting in inter-language differences from 8 years onwards (Figure 7 and Figure 8a). Such a developmental shift was similarly observed in the RT data although AV interaction (RT difference [AV–] – [AV+]) tended to be stronger for the English language participants at each age and it gradually increased over age after 8 years in *both* language groups (Figure 10a). The greater visual influence for English language participants after 6 years (Figure 8a) was related to the relative speed of unimodal processing (AO–VO): This difference was almost zero at 6 years in both language groups, then for the English language participants, the precedence of visual processing emerged at 8 years and was maintained across age, whereas for Japanese participants, no such visual precedence was observed (Figure 11). In speechreading (VO), English language participants tended to be more accurate than Japanese participants, but the group difference was significant only at 11 years (Figure 8b). So it appears that it is the English and Japanese children's relative speed in unimodal processing (VO vs. AO), evident from 8 years, rather than the more accurate VO processing by English language children, evident only at 11 years, that drives the developmental shift in auditory-visual speech perception between 6 and 8 years for the English language children. Indeed the Japanese 6-year-olds are 'immune' to such a shift due to their greater accuracy in the AO condition, compared to the English language 6-year-olds (Figure 8c), and this early auditory superiority in accuracy by Japanese participants is transformed into superiority in auditory latencies over age to adulthood (Figure 6).

These results clearly show that language experience has an impact on AV speech perception, and that there is a significant developmental dimension to this. We confirm that AV integration increases over age for English language participants (McGurk & MacDonald, 1976; Massaro, 1984; Massaro *et al.*, 1986), and also show that this is differentially manifested cross-linguistically: Not only is AV integration developmentally promoted more in the English than the Japanese language environment, but such an effect of language environment on visual influence in speech perception is already evident at around 8 years. The emergence of this cross-linguistic

difference over development may be related to the relative speed of processing of auditory and visual information, as set out below.

Aspects of inter-language differences can be noted in both the AO and VO unimodal accuracy data. In the AO condition, the Japanese participants were more accurate than the English language participants at 6 years, and this difference disappeared at later stages. This early auditory superiority in Japanese children could obviate the need for visual supplementation, and result in greater auditory-dependent processing in AV speech perception at later stages. The fact that Japanese adults respond to auditory information more quickly than do English language adults may be a consequence of this early proficiency in their auditory processing.

The reason why the Japanese 6-year-olds are more accurate than their English counterparts in the AO condition is still unknown, but it may be due to the less crowded phoneme space in the Japanese language. Japanese syllable identification may be less difficult with auditory information alone due to the smaller number of vowels (5 vs. around 14 in English), and lack of some consonant contrasts that occur in English (e.g. /r/ vs. /l/; /b/ vs. /v/; and /s/ vs. /θ/). English syllable identification via auditory information alone may thus be more difficult, which could result in (a) greater processing time required to develop the same accuracy level of AO identification and (b) more susceptibility to augmentation of speech perception by visual cues.

In accord with this argument, English language participants were generally more accurate than the Japanese participants in the VO condition. This difference was significant only at 11 years, and this relatively late visual superiority in English language participants may be regarded as a consequence of relatively more visually tuned AV processing in younger English language participants.

The results for the relative speed of unimodal processing, namely, AO–VO RT differences, supplement the accuracy data: It appears that the inter-language differences in visual influence (Figure 8a) are related to relative speed of unimodal processing. Temporal precedence of visual over auditory processing emerged at 8 years in the English but not the Japanese participants, suggesting that English language perceivers become more visually tuned in speech perception after 6 years whereas Japanese perceivers do not. Such a shift in the English speakers could result in the heightened use of visual speech information after 8 years, not because there is any greater availability of visual speech information for English language children, but because it is available relatively earlier in the time course of auditory-visual speech perception. Why might this shift in attention to visual vs. auditory speech cues occur for English, but not Japanese children?

In many instances of crossmodal binding, an early developmental preference for auditory input is later supplanted by visual dominance (Burnham, 1993; McGurk & MacDonald, 1976; Querleu & Renard 1981; Querleu, Renard, Versyp, Paris-Delrue & Crepin, 1988; Robinson & Sloutsky, 2004). This is possibly related to the early functionality of the auditory system *in utero* and the effective prenatal transmission of auditory stimulation to the fetus (Aldridge, Stillman & Brown, 2001; Burnham, 1993; Jusczyk, 1995; Mastropieri & Turkewitz, 1999; Sansavini, Bertoncini & Giovanelli, 1997; Tharpe & Ashmead, 2001; Trehub, 2001). Does this auditory to visual developmental trend also obtain for speech perception?

It is possible that this developmental trend for a dominant modality shift may only be manifested in speech perception under certain specific phonological conditions. At school, between 6 and 8 years, children experience the perceptually challenging exposure (Ryalls & Pisoni, 1997) to a huge population of speakers. This challenge of talker variability may be greater for English due to phonological differences: Compared to Japanese, English has more varied and complex syllable structures, more vowels, more consonant contrasts, many syllable-initial consonant clusters (at least 31 vs. none in Japanese), and syllable-final consonant clusters. Additionally, English incorporates visually but not so auditorally distinct consonant contrasts, e.g. labiodental-interdental-alveolar – as in four-thaw-saw, or vat-that-sat, or other contrasts such as in bat-vat, or lead-read, which do not exist in Japanese. The phonological complexity of English could provide pressure to seek extra sources of information, with the visual distinctiveness of English phonology providing an effective source of such information.

As auditory-visual speech integration may be considered to require attention (Alsius, Navarra, Campbell & Soto-Faraco, 2005), the less complex phonology and less distinctive visual information in Japanese may not provide the necessary preconditions and impetus for attention to visual cues in noiseless conditions. By the same token, the heightened use of visual speech information by English children from sometime between 6 and 8 years onwards may be an adaptive strategy for attending to salient information in their phonologically complex and visually distinctive language. In Japanese, talker variability intensification around school age may also be present (as shown in increasing visual effects on RT [AV–] – [AV+] differences), but the phonological demands may not, especially if the simpler structure of their language promotes better AO speech perception early in development (which in turn seems to result in Japanese adults' faster AO speech perception compared with English language adults). Thus, while Japanese children may experience AV *interaction* to some extent, they may not *integrate* visual information because the phonological environment does not demand it. There is faster processing of visual information as a function of increasing age for *both* language groups, but it is only in the English language environment that the visual information is translated into auditory-visual perceptual decision, perhaps due to the temporal precedence of visual over auditory processing. Thus the availability of visual speech information temporally prior to auditory information for English

langauge children, but at similar points in time for Japanese language children, may be the critical issue here – auditory-visual speech integration may require this temporal precedence of visual over auditory information.

The combined response frequency and latency data in this study converge to suggest that the English language environment promotes the earlier availability and *use of* visual speech information resulting in greater auditory-visual speech integration; whereas the Japanese language environment promotes earlier availability of auditory speech information, superior auditory-only speech perception, and relatively lower auditory-visual speech integration. These results point to the need for further research regarding the relative time course of auditory and visual processing in auditory-visual speech perception (Bernstein, Burnham & Scwartz, 2002). Neverthless, it can be concluded that in the course of development, the nature of the surrounding language impacts upon auditory-visual integration in speech perception especially during the early school years.

Acknowledgements

The authors thank the children and adults who participated in this and pilot studies, and their parents and teachers; the Akagawa Elementary School, and the NSW Department of Education and Training. We are grateful for the assistance of Yasuko Nagasaki, Thomas Stainsby, Eleanor Gittins, and David Goddard in running pilot studies; and Yumiko Saito, Helen Tam, and Doğu Erdener for running the main study; Dr Michael Tyler for phonetic advice; and Professors Cathi Best and Barbara Dodd for comments on an earlier draft. The work was supported by grants from the Hayao Nakayama Foundation, the Japanese Ministry of Education, Culture, Sports, Science, and Technology (10610070; 18530563); the Australian Research Council (DP0211947, DP0558698), Future University Hakodate, and the University of Western Sydney.

References

Aldridge, M.A., Stillman, R.D., & Bower, T.G.R. (2001). Newborn categorization of vowel-like sounds. *Developmental Science*, **4**, 220–232.

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demand. *Current Biology*, **15**, 839–843.

Bernstein, L.E., Burnham, D., & Schwartz, J.-L. (2002). Special session: Issues in audiovisual spoken language processing (when, where, and how?). *Proceedings of the 7th International Conference on Spoken Language Processing, Denver, USA*, pp. 1445–1448.

Burnham, D. (1993). Visual recognition of mother by young infants: facilitation by speech. *Perception*, **22**, 1133–1153.

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by pre-linguistic infants: perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, **44**, 204–220.

Burnham, D., Kitamura, C., & Mattock, K. (2007). Psycho-LINGUISTICS meets PSYCHOlinguistics: different emphases, sustainable collaborations. In W. Aroonmanakun (Ed.), *Unfolding linguistics* (pp. 23–39). Bangkok: Chulalongkorn University Press.

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

Campbell, R., Dodd, B., & Burnham, D. (1998). *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech*. Hove: Psychology Press.

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.

Desjardins, R.N., Rogers, J., & Werker, J.F. (1997). An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks. *Journal of Experimental Child Psychology*, **66**, 85–110.

Dodd, B., & Campbell, R. (1987). *Hearing by eye: The psychology of lip-reading*. London: Lawrence Erlbaum Associates.

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.

Gagné, J.P., Masterson, V., Munhall, K.G., Bilida, N., & Querengesser, C. (1994). Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech. *Journal of the Academy of Rehabilitative Audiology*, **27**, 135–158.

Grassegger, H. (1995). McGurk effect in German and Hungarian listeners. *Proceedings of the International Congress of Phonetic Sciences, Stockholm, Vol. 4* (pp. 210–213). Stockholm: Congress organizers at KTH and Stockholm University.

Green, K.P., & Kuhl, P.K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, **17**, 278–288.

Harris, R.J. (1994). *ANOVA: An analysis of variance primer*. Itasca, IL: F.E. Peacock.

Hockley, N., & Polka, L. (1994). A developmental study of audiovisual speech perception using the McGurk paradigm. *Journal of the Acoustical Society of America*, **96**, 3309.

Kuhl, P.K., Tsuzaki, M., Tohkura, Y., & Meltzoff, A.N. (1994). Human processing of auditory-visual information in speech perception: potential for multimodal human–machine interfaces. In the Acoustical Society of Japan (Ed.), *Proceedings of the International Conference of Spoken Language Processing* (pp. 539–542). Tokyo: the Acoustical Society of Japan.

Jusczyk, P.W. (1995). Language acquisition: speech sounds and phonological development. In J.L. Miller & P.D. Eimas (Eds.), *Handbook of perception and cognition: Vol.11. Speech, language, and communication* (pp. 263–301). Orlando, FL: Academic Press.

MacDonald, J., & McGurk, H. (1978). Visual influence on speech perception processing. *Perception and Psychophysics*, **24**, 253–257.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746–748.

Massaro, D.W. (1984). Children's perception of visual and auditory speech. *Child Development*, **55**, 1777–1788.

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

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

Massaro, D.W., & Cohen, M.M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, **9**, 753–771.

Massaro, D.W., & Cohen, M.M. (1996). Perceiving speech from inverted faces. *Perception and Psychophysics*, **58**, 1047–1065.

Massaro, D.W., Thompson, L.A., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology*, **41**, 93–113.

Massaro, D.W., Tsuzaki, M., Cohen, M.M., Gesi, A., & Heredia, R. (1993). Bimodal speech perception: an examination across languages. *Journal of Phonetics*, **21**, 445–478.

Mastropieri, D., & Turkewitz, G. (1999). Prenatal experience and neonatal responsiveness to vocal expressions of emotion. *Developmental Psychobiology*, **35**, 204–214.

Querleu, D., & Renard, X. (1981). Les perceptions auditives du fetus humain. *Medicine & Hygiene*, **39**, 2101–2110.

Querleu, D., Renard, X., Versyp, F., Paris-Delrue, L., & Crepin, G. (1988). Fetal hearing. *European Journal of Obstetrics and Gynaecology and Reproductive Biology*, **29**, 191–212.

Robinson, C.W., & Sloutsky, V.M. (2004). Auditory dominance and its change in the course of development. *Child Development*, **75**, 1387–1401.

Ryalls, B.O., & Pisoni, D.B. (1997). The effect of talker variability on word recognition in preschool children. *Developmental Psychology*, **33**, 441–452.

Sansavini, A., Bertoncini, J., & Giovanelli, G. (1997). Newborns discriminate the rhythm of multisyllabic stressed words. *Developmental Psychology*, **33**, 3–11.

Sekiyama, K. (1994). Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *Journal of the Acoustical Society of Japan (E)*, **15**, 143–158.

Sekiyama, K. (1997). Cultural and linguistic factors in audio-visual speech processing: the McGurk effect in Chinese subjects. *Perception and Psychophysics*, **59**, 73–80.

Sekiyama, K., Braida, L.D., Nishino, K., Hayashi, M., & Tsuyo, M. (1995). The McGurk effect in Japanese and American perceivers. *Proceedings of the 13th International Congress of Phonetic Sciences, Vol. 3* (pp. 214–217). Stockholm: Congress organizers at KTH and Stockholm University.

Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neuroscience Research*, **47**, 277–287.

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*, **90**, 1797–1805.

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

Siva, N., Stevens, E.B., Kuhl, P.K., & Meltzoff, A.N. (1995). A comparison between cerebral-palsied and normal adults in the perception of auditory-visual illusions. *Journal of the Acoustical Society of America*, **98**, 2983.

Stork, D.G., & Hennecke, M.E. (1996). *Speechreading by humans and machines: Models, systems, and applications*. New York: Springer.

Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**, 212–215.

Tharpe, A.M., & Ashmead, D.H. (2001). A longitudinal investigation of infant auditory sensitivity. *American Journal of Audiology*, **10**, 104–112.

Trehub, S.E. (2001). Musical predispositions in infancy. *Annals of the New York Academy of Science*, **930**, 1–16.

Vatikiotis-Bateson, E., Perrier, P., & Bailly, G. (in press). *Advances in auditory-visual speech processing*. Cambridge, MA: MIT Press.

Received: 5 October 2006
Accepted: 27 April 2007