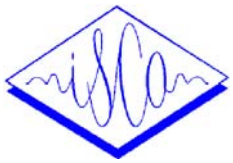
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Issues in the Development of Auditory-Visual Speech Perception: Adults, Infants, and Children

Kaoru Sekiyama[1] *and Denis Burnham*[2]

1. Future University, Hakodate, Japan
2. MARCS Auditory Laboratories, University of Western Sydney, Australia
<sekiyama@fun.ac.jp>, <d.burnham@uws.edu.au>

Abstract

In this paper we review research by us and others on the development of auditory-visual speech perception in order to draw out issues of interest, pinpoint unanswered questions, and highlight future research directions. Studies using the McGurk Effect have shown that Japanese adults use visual speech information less than do English language adults. Infants perceive the McGurk Effect, and for English language children, it appears that auditory-visual speech perception improves over age, especially between 6 and 8 years. However, the same is not the case for Japanese children.

10.21437/Interspeech.2004-412

1. Introduction

1.1. Auditory-Visual Speech Perception

Optical information from facial movements of a talker contributes to speech perception not only when acoustic information is degraded [1] or when the listener is hearing-impaired, but also when the acoustic information is clearly audible. This is most clearly shown in the classic "McGurk effect" or "fusion illusion", in which dubbing the auditory speech syllable [ba] onto the lip movements for [ga] results in the emergent perception of "da" or "tha". This occurs both when the observer is aware, or unaware of the conflicting sources of information [2]. The beauty of this effect is not simply that it results in an illusion, but that it unequivocally shows that visual information is used in speech perception even when auditory information is clear and undegraded.

1.2. Developmental Methods

Developmental studies are vital to determine whether and to what extent perceptual, linguistic, or cultural experience on the one hand, and cognitive capacity on the other affect auditory-visual speech processing. There are two methods of interest here. The ontogenetic method involves comparing the abilities of individuals of different ages brought up in functionally identical environments on a common task in order to investigate the effect of the *amount* of experience or maturation on development. The differential experience method involves comparing the ability of individuals of the same age brought up in functionally different environments on a common task in order to investigate the effect that the *type* of exposure or experience may have on development [3].

The McGurk Effect has been used in auditory-visual speech perception research because it is an effective and standard tool. Here, studies using this tool in both differential experience and ontogentic paradigms are discussed in order to elucidate issues, questions, and future research directions in auditory-visual speech perception.

2. Differential Experience Studies with Adults

2.1. Empirical Studies

In an attempt to replicate the original McGurk effect results, we tested native speakers of Japanese with speech stimuli, /ba/, /da/, /ga/, /pa/, /ta/, /ka/, /ma/, /na/, /ra/, and /wa/, articulated by a native Japanese talker [4]. In a pilot experiment, the evidence of the McGurk effect was very weak. So we set up two conditions. In a noise-added condition, white noise was added to the auditory component of the stimuli. In a quiet condition, no noise was added. The Japanese participants reported a weak McGurk effect in the quiet condition, but a strong effect in the noise-added condition. It was as if the native Japanese talkers used visual information in speech perception only when the auditory component of the stimuli was not perfectly intelligible.

Subsequently, we conducted cross-language comparisons [5] with two sets of stimuli. One set was Japanese (identical to the stimuli in the above study), and the other was English (stimuli articulated by a native English talker). The two sets of stimuli were presented to two language groups, native speakers of Japanese and American English. On average the Japanese participants showed a weaker McGurk effect than did the Americans. So again a language-influenced McGurk effect was found. Nevertheless, the Japanese reported a stronger McGurk effect with the non-native (English) speech stimuli than with the native (Japanese) stimuli. This weaker visual influence in Japanese perceivers has also been replicated in another laboratory [6, but see also 7].

Further data were obtained from native speakers of Chinese [8]. The same Japanese and English stimuli were used as above, so both stimulus sets were non-native for the perceivers. Chinese participants showed a weak McGurk effect for both Japanese and English stimuli. Comparison with Sekiyama's Japanese and American results [4,5] revealed that the McGurk effect was weakest in the Chinese, intermediate in the Japanese, and strongest in the Americans.

2.2. Issues

A number of possible reasons for this difference between Chinese, Japanese and English speakers can be considered. First, one might think that this language influenced McGurk effect might occur because Japanese adult perceivers are poor speechreaders. However, this is unlikely because the Japanese adults *did* use visual information in the noise-added condition [4]. Thus, in the quiet condition they must have processed visual information, but did not make use of it. In fact,

Sekiyama [9] showed that the Japanese perceivers more frequently noticed the incompatibility between auditory and visual cues in McGurk stimuli than did the Americans.

A second possibility is that Japanese perceivers do not incorporate visual cues into speech perception so much because they do not pay much attention to visual speech in their everyday lives. This may be related to the Japanese cultural habit of avoiding staring at the person to whom one is talking. (Note that this seems to be diminishing as younger generations of Japanese become more Westernised). It may be also related to the fact that articulatory movements for Japanese syllables are not as visibly distinct as those for English ones. Phonologically, there are no visually-distinct consonants such as /v/ or /D/ (as in then) in Japanese. Furthermore, it appears that the extent of mouth/face movements by Japanese speakers is less than English speakers. (However, this issue requires further empirical investigation.)

Thirdly, in Japanese there may be less need to incorporate visual information because the phonology is less complicated - there are no consonant clusters, and only five vowels [5]. This may be countered by the fact that native Spanish speakers respond to McGurk stimuli much the same way as English speakers [7]. Final resolution of this option may rely upon systematic extensive comparisons of auditory-visual speech perception in languages with different phonologies.

Finally, it may be that the degree of visual influence in the McGurk effect is inversely proportional to the use of tonal information in the perceivers' native language. Consider the following. In Chinese the meanings of spoken words are determined not only by vowel and consonant combinations, but also by the pitch pattern (tone) of the word. A similar device, pitch-accent, operates in Japanese on multisyllabic words. Now, while it has recently been found that there is some reliable visual information for tone [10, 11], it may reasonably be assumed that tone and pitch-accent are most strongly carried by the auditory modality. If so, then auditory information should be weighted more strongly and visual information less strongly in tonal languages. This proposal is consistent with the observed results: auditory information is used less (and visual more) in English (no tones), than in Japanese (two pitch-accents), and in turn than Chinese (four tones in Mandarin, and six tones in Cantonese).

3. Studies on Infants' Perception of the McGurk Effect

3.1. Empirical Studies

The status of auditory-visual speech perception *at* birth is not explicitly known, though there is evidence for auditory-visual speech perception abilities shortly *after* birth. Infants match faces and voices on the basis of whether face and voice are in synchrony or not by 2½ months of age [12,13]; and on the basis of vowel colour [14, 15] and consonant-vowel combinations [16] by 4 and 6 months respectively.

These results show that by at least 3 months infants perceive structural correspondences between seen and heard speech and match visual and auditory speech. However, the studies cannot distinguish between two possible mechanisms: (a) that infants perceive speech as an integrated auditory-visual event, or (b) that infants match the characteristics of one modality with those of the other. If infants were found to perceive an emergent percept in the McGurk effect paradigm, then it could be more confidently claimed that the first alternative is the case.

There is good evidence from three studies that infants integrate auditory and visual speech information. The first two rest on the observation that adults perceive auditory [ba], visual [va] as "va", but auditory [da], visual [va] as "da". 5-month-old infants' visual fixation to auditory-visual [va] and was habituated and then they were tested on different auditory-visual combinations [17]. There was generalisation of habituation to an auditory [ba], visual [va] presentation, showing that infants perceived this to be the same as, or at least similar to, auditory-visual [va]. However, habituation did *not* generalise to an auditory [da], visual [va] presentation, showing that infants perceived this to be different from auditory-visual [va]. These results show that 5-month-old infants appear to perceive both auditory [ba], visual [va] "va", whereas auditory [da], visual [va] is not perceived as "va", and possibly as "da". Thus in certain combinations the auditory information dominates, and in others the visual information dominates, just as is the case with adults. In a similar study 4-month-old infants habituated to auditory-visual [vi], showed no recovery of visual fixation (i.e., no novelty response) when presented with auditory [bi], visual [vi], but infants habituated to auditory-visual [bi] *did* show recovery of visual fixation to auditory [bi], visual [vi] (although this was the case only for female infants) [18]. These results support the view that there is a visual influence in speech perception by infants.

Using the traditional McGurk effect, we [13] tested whether 4.5-month-old infants perceive an emergent percept, [da] or [Da], when presented with auditory [ba], visual [ga], just as children and adults do. Infants habituated to auditory [ba], visual [ga] showed a greater novelty response to [ba] than to [da] or [Da], indicating that this habituation stimulus was *not* perceived as [ba]. On the other hand, infants habituated to matching auditory-visual [ba], showed a greater novelty response to [da] and [Da] than to [ba], suggesting that this habituation stimulus *was* perceived as [ba]. This is evidence for an emergent or fusion percept in infancy, and thus for integrated auditory-visual perception of speech by at least 4.5 months of age. Together these studies provide strong evidence that infants use both auditory and visual information in speech perception.

3.2. Issues in Infant Auditory-Visual Speech Perception

An issue of interest in the infant studies concerns the degree of auditory-visual integration. Werker and Desjardins [18] suggest that auditory-visual integration in infancy is not mandatory, that infants may perceive auditory and visual information separately under some conditions. We [13] have evidence for a similar position – in a subsidiary analysis there was found evidence for some residual memory for the auditory component of the McGurk stimulus, despite perception of the emergent phoneme.

4. Ontogenetic Studies with Children

4.1. Empirical Studies

The original report of the McGurk effect included ontogenetic development data [2] with pre-schoolers (3 to 5 years), school

children (7 to 8 years), and adults (18 to 40 years). Mean visually influenced responses were 59%, 52% and 92% for pre-schoolers, school children, and adults respectively, in spite of the fact that the three groups identified the auditory-only stimuli equally accurately (91, 97, & 99%). Similar developmental increases have been found with children (4 to 6 years) compared with adults [19], and also across childhood - a gradual developmental increase from children (5-, 7-, 9-, 11-year-olds) to adults [20].

The developmental increase in visual influence in these three studies is possibly related to experience in articulating speech sounds. It has been found that preschool children who make articulation errors are less influenced by visual cues than are children who can correctly produce consonants [21]. Based on an articulation test they divided preschoolers (3 to 5 years of age) into two groups: those who made substitution errors and those who did not. Subsequent perception tests revealed that the substituter children were the poorest at speechreading, and had the lowest degree of visual influence in auditory-visual speech perception, followed by the non-substituter children and then the adults. The three groups did not differ, however, in auditory-only speech perception. The authors concluded that experience in correctly producing consonants impacts upon the representation of visible speech. It is, of course, possible that the opposite causal effect is the case - that poor speech reading impacts upon articulation ability. However, it has been shown 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 [22]. Thus, it appears that articulation experience affects speechreading, rather than the other way around.

4.2. Issues

While it is clear that infants perceive speech in an auditory-visual fashion, it is also clear that there is a developmental increase in the use of visual information in speech perception. However, it is unclear why this might be the case. Is there more need for visual information to disambiguate speech input as the level of speech complexity and processing increases? Or is it the case that as articulation improves over development, that the use of visual information increases over age due to improvements in articulation ability? In order to answer these questions, more information is required regarding this developmental improvement, such as its generality across languages, and the specific age at which development occurs.

5. An Ontogenetic / Differential Experience Study with Children

5.1. The Data

Using a combination of the 'amount' and 'type' developmental methods, we [23] tested Japanese and English language 6-, 8-, and 11-year-old children and adults (144 participants in total) with carefully-piloted English and Japanese McGurk stimuli (created by editing /ba/, /da/, /ga/ uttered by two Japanese and two English language speakers). There was equivalent visual influence for 6-year-old English and Japanese participants, and for the Japanese the degree of visual influence remained constant over age. However, for the

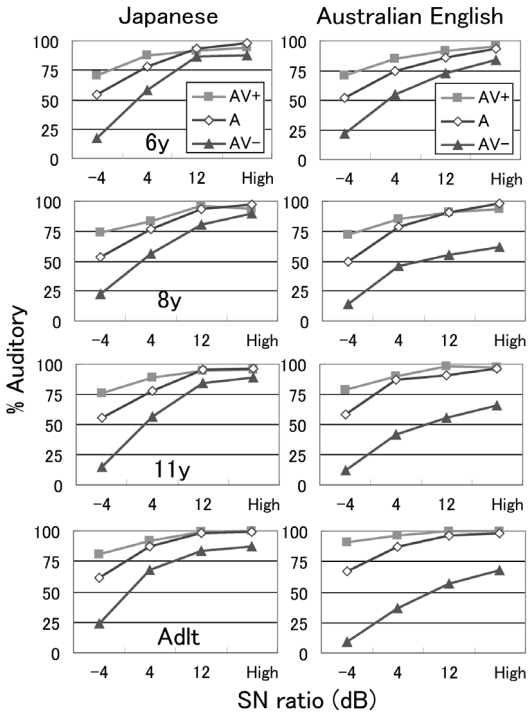


Fig. 1 Percent auditory in AV matching (AV+), AV mismatching (AV-), and audio only (A) presentations.

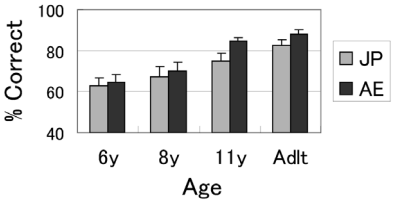


Fig. 2 Percent correct in lip-reading (V).

English language participants, there was a sharp and significant increase in visual influence between 6 and 8 years, which was maintained over age. Figure 1 shows the percent of auditorally correct responses in audiovisual matching (AV+), mismatching (AV-) and audio-only (A) conditions as a function of signal-to noise (SN) ratio of auditory speech (addition of band noise of 300-12000Hz). The size of visual influence can be seen as the difference between AV+ and AV- at each SN ratio. The data indicate that English language experience results in significantly greater visual influence in speech perception than Japanese language experience but only after 6 years of age. The so-called Japanese McGurk effect is thus actually an *English* McGurk effect, Japanese perceivers do not change over age whereas English speakers do. Why?

Another language difference obtained in this developmental study was that speechreading performance in the visual-only condition was better in English speakers than in Japanese speakers on average (87.2% vs 83.0%). However, the difference was only significant at 11 years (Fig. 2), suggesting that this difference is a consequence, rather than a cause of the increased use of visual cues in auditory-visual speech perception. Thus the origin of the sharp divergence in

auditory-visual speech perception at 6 and 8 years by English language children is still unclear.

4.2 Issues

One of the most significant external events that occur between 6 and 8 years is the onset of schooling. Might what is learned at school influence visual speech perception? In another context we have found that children's reading ability is related to their language specific speech perception, measured by perceptual ability for native vs non-native speech contrasts [24]. Above it was noted that the use of visual information in speech perception may be related to articulation ability [21, 22]. In order to examine whether the onset of formal language training and linguistic variables might account for the onset of greater visual influence in English speakers between 6 and 8 years, experiments are required to test school age children on the McGurk effect task along with tests of articulation, reading, and language specific speech perception.

6. Summary and Conclusions

Developmental studies show that auditory-visual speech perception is present in infancy, and differs between Japanese and English language perceivers. Now it has been found that the increased use of visual information by English language users originates between 6 and 8 years. We know that visual information improves speech perception. Investigation of the concomitants of the increase in the use of visual information between 6 and 8 years for English speakers will shed light on the functional load of visual information in speech perception.

7. References

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

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

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

[4] Sekiyama, K. & 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.

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

[6] Kuhl, P.K., Tsuzaki, M., Tohkura, Y., & Meltzoff, A.N. Human processing of auditory-visual information in speech perception: Potential for multimodal human-machine interfaces. *Proceedings ICSLP*. Tokyo: *Acoust. Soc. Japan.*,539-542, 1994.

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

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

[9] Sekiyama, K. Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *J. Acoust. Soc. Japan*, 15:143-158, 1994.

[10] Burnham, D., Ciocca, V., & Stokes, S. (2001) Auditory-visual perception of lexical tone. In P. Dalsgaard, B. Lindberg, & H. Benner (Eds) *Proc.Eurospeec, Aalsborg, Denmark,* ISCA, Bonn, Germany, 395-398, 2001

[11] Burnham, D., Lau, S., Tam, H., & Schoknecht, C. (2001) Visual discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by non-tonal language speakers. In Massaro, D., Light, J., & Geraci, K. (Eds) *Proc. AV Speech Perception,* Causal, 155-160, 2001.

[12] Dodd, B. Lip reading in infants: Attention to speech presented in- and-out-of-synchrony. *Cognitive Psychology,* 11:478-484, 1979

[13] Burnham, D. & Dodd Familiarity and novelty in infant cross-language studies: Factors, problems, and a possible solution. In C. Rovee-Collier (Ed) *Advances in Infancy Research*, 12: 170-187, 1998.

[14] Kuhl, P.K. & Meltzoff, A.N. The bimodal perception of speech in infancy. *Science, 218*, 1138-1141, 1982.

[15] Patterson, M.L., & Werker, J.F. Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, 6:191-196, 2003.

[16] MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. Infant intermodal speech perception is a left-hemisphere function. *Science*, 219: 1347-1348, 1983.

[17] Rosenblum, L.D., Schmuckler, M.A., & Johnson, J.A. The McGurk effect in infants. *Perception & Psychophysics*, *59*, 347-357, 1997.

[18] Desjardins, R.N. & Werker, J.F. 4-month-old female infants are influenced by visible speech. *International Conference of Infant Studies*, Providence RI, 1996

[19] Massaro, D.W., Thompson, B.B., & Laren, E. Developmental changes in visual and auditory contributions to speech perception. *J. Experimental Child Psychology*, 41: 93-113, 1986.

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

[21] Desjardins, R.N., Rogers, J. & Werker, J.F. An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks. *J. Experimental Child Psychology*, 66: 85-110, 1997.

[22] Siva, N., Stevens, E.B., Kuhl, P.K., & Meltzoff, A. N. A comparison between cerbral-palsied and normal adults in the perception of auditory-visual illusions. *J. Acoust. Soc. Am.*, 98:2983, 1995

[23] Sekiyama, K., Burnham, D., Tam, H., & Erdener, D. Development of auditory-visual speech perception in Japanese and English speakers. J-L Schwartz, F. Berthommier, M-A.Cathiard, & D. Sodoyer (Eds) *Proc. AV Speech Processing, St. Jorioz, France,* Institut de la Communication Parlée, Grenoble, 43-46, 2003

[24] Burnham, D. Language specific speech perception and the onset of reading. *Reading and Writing: An Interdisciplinary Journal,* 16: 573-609, 2003.