# HUMAN PROCESSING OF AUDITORY-VISUAL INFORMATION IN SPEECH PERCEPTION: POTENTIAL FOR MULTIMODAL HUMAN-MACHINE INTERFACES

*Patricia K. Kuhl,[1] Minoru Tsuzaki,[2] Yoh'ichi Tohkura,[2] and Andrew N. Meltzoff[3]*

[1] Department of Speech and Hearing Sciences, University of Washington, Seattle, WA 98195

[2] Advanced Telecommunication Research Laboratories International, Kyoto, 619-02, Japan

[3] Department of Psychology, University of Washington, Seattle, WA, 98195

## ABSTRACT

Speech perception is not a unimodal process. Observers who see and hear a talker take both auditory and visual information into account in determining what the talker said. This is best illustrated by experiments in which discrepant speech information is delivered to the two modalities. In this situation, observers perceive neither the syllable sent to the auditory modality nor the syllable sent to the visual modality, but a combination of the two. Recent research in our laboratories has established two additional facts. First, the auditory-visual effect occurs in both American and Japanese subjects. Moreover, our studies show that the effect interacts with the language spoken by the talker one watches. Japanese subjects show significantly stronger auditory-visual effects when watching a foreign-language speaker than when watching a native-language speaker. In American subjects, this difference is less pronounced. Second, in investigating auditory-visual effects, we have found that minimal visual information is necessary to produce auditory-visual effects. Data from human observers have implications for human-machine interfaces that utilize multimedia technology.

## INTRODUCTION

There is evidence showing that speech perception is a multimodal phenomenon [1]. Psychological tests show that when people watch a speaker they are influenced both by what they see and what they hear. This is demonstrated in a situation in which discrepant auditory and visual speech information is presented. For example, when audio /ba/ is combined with video /ga/, observers report "hearing" a syllable with an intermediate place of articulation, such as /da/, /tha/, or /za/. When audio /ga/ is combined with video /ba/, /bga/ is perceived. In both cases, speech perception is co-determined by the two modalities, rather than determined by a single modality. This effect has been demonstrated even in circumstances in which the information sent to the two modalities could not have derived from a single biological source. When viewers see a male talker but hear the voice of a female talker, they nonetheless show the auditory-visual effect [2]. Even young infants are influenced by visually presented speech information [3]. When multimodal speech information is available, observers seem unable to ignore it — they appear compelled to take visual information into account.

Multimedia technology uses face-voice material when presenting spoken language information. Given the psychological data, the multimedia presentation of speech is preferable not only because it is more "user-friendly," but because multimodal presentation actually improves the perception of speech due to the mind's automatic integration of auditory and visual speech information. Multimedia presentation could be particularly helpful when auditory information is unclear or degraded due to noise.

A question that arises is the extent to which auditory-visual speech effects are culturally universal, and the effects of native- as opposed to foreign-language stimulus materials on auditory-visual speech perception. Multimodal speech perception has been studied primarily in American English subjects using American English stimuli. A number of reports show that Japanese people watching and listening to Japanese speech are much less strongly affected by visual information than American subjects [4], suggesting that there may be cultural differences in the perception of audio-visual speech information. If so, then the utility of multimedia speech presentation might be limited to certain populations.

Three factors could account for the observed differences in auditory-visual speech perception between American and Japanese subjects: (a) stimulus factors, (b) subject factors, or (c) language-experience factors. Regarding stimulus factors, Japanese speech stimuli might contain fewer visual cues (due to less lip movement by Japanese as opposed to American speakers). If this were the case, both Americans and Japanese should show reduced effects when watching Japanese talkers speak. Regarding subject factors, it has been suggested that Japanese adults have less experience watching the talker's face (because it is considered impolite). Japanese people might therefore not integrate visual information to the same degree when compared to Americans. If this were the case, then Japanese subjects should show reduced auditory-visual effects when compared to Americans regardless of the stimulus material.

Finally, regarding language-experience factors, exposure to auditory-visual signals produced by speakers of one's native language could alter the processing of foreign-language auditory-visual speech information. Phonetic units differ across languages, even when they are ostensibly the "same" phonetic unit. Thus, the consonants /b/, /d/, and /g/, while present in both languages, are produced somewhat differently by American and Japanese speakers. Kuhl has argued that exposure to a specific language results in stored auditory and visual representations of native-language units [5]. On this account, American and Japanese stored representations differ. Depending on the degree to which the auditory - visual signals of foreign speakers match native listeners' stored representations, differential auditory-visual results would be expected to be obtained when tested with native- as opposed to foreign-language speech material.

Linguistic experience profoundly affects the auditory perception of speech in adults [6]. Moreover, the effect of auditory language experience occurs as early as 6 months of age [7]. By 4 months of age, infants detect auditory-visual correspondences for native-language speech units, suggesting that they have learned how certain sounds "look" on the face of a talker [3]. Adults have been

watching native speakers for a long period of time. It is plausible that multimodal speech perception has been affected by this long period of exposure to native-language auditory and visual speech.

The present study was designed to systematically investigate which of the three alternative hypotheses — stimulus factors, subject factors, or language-experience factors — best account for the data. We designed an experiment in which American and Japanese adults were tested with identical auditory-visual stimulus materials of two types, American and Japanese. The tapes used in the tests were identical. The testing situations were identical. The only factor that differed was the language experience of the two subject groups. If stimulus factors play a major role in accounting for the results on American and Japanese subjects, then both groups should show larger auditory-visual effects when watching one of the video tapes (presumably the American tape). If subject factors are of primary importance, then one group, presumably the Americans, would be expected to show larger auditory-visual effects with both video tapes. If language experience factors strongly influence auditory-visual speech perception, then interactions between the language background of the subjects and language spoken on the stimulus tape should be observed.

## METHODS

Stimuli.

Two sets of visual stimuli (American and Japanese) were prepared by videotaping two female talkers who produced several instances of the syllables /ba/ and /ga/. The talker on the American tape was a native speaker of American English; the talker on the Japanese tape was a native speaker of Japanese. The stimuli were recorded in color. The talker was seated on a stool placed in front of a black background. The camera was centered on the talker's face and resulted in an excellent view of the talker's mouth and face. The entire face of the speaker, including the hair, was visible. When shown on a 14" color monitor the facial images were life-size. From the video recordings of each speaker, a single /ba/ and /ga/ were selected. The two video tokens were similar in their overall durations and the articulations lacked any extraneous movements.

Auditory tokens of the syllables /ba/ and /ga/, spoken by the same two talkers, were recorded with a high quality microphone and tape recorder. The syllables were digitized at a 20 kHz sampling rate on a digital computer, low-pass filtered at 10 kHz and analyzed with signal processing software. A single /ba/ and /ga/, which closely matched the duration of the corresponding video tokens, were selected for each talker.

Three conditions were run with both the American and Japanese stimuli: auditory-visual (AV), auditory-only (A), and visual-only (V). Once the American and Japanese stimulus tapes were created for the AV condition, the A and V conditions could be run simply by turning the audio signal off and playing only the visual signal in the V condition, and by turning the video signal off and playing the audio portion in the A condition.

American and Japanese AV stimulus tapes were created in the exact same manner. The auditory and visual /ba/ and /ga/ stimuli were combined to create four auditory-visual stimuli. Two of the four stimuli provided conflicting phonetic information. In one case, auditory /ba/ was paired with visual /ga/, which typically results in the perception of a syllable with an intermediate place of articulation (/da, /tha/, or /za/), referred to as a "fusion" response. In the other case, auditory /ga/ was paired with visual /ba/, which typically produces a /bga/ response, referred to as a "combination" response. The final two AV stimuli were controls. The [illegible] providing phonetic information: auditory /ba/ was paired with visual /ba/ and auditory /ga/ was paired with visual /ga/.

When dubbing the auditory stimuli onto the video stimuli, the output of one audio channel from the videotape recorder, which contained the original sound track, was fed into the computer. A marker tone preceded each of the utterances on the original sound track. When the computer sensed the onset of the marker tone, it output one new audio token according to a predetermined order onto the second audio track of the videotape. The new audio tokens were synchronized to occur with the video articulations by using a temporal delay, which was precalculated so that the release burst of the dubbed utterance would match precisely the release burst of the original utterance corresponding to the video token. The syllables were output at 20 kHz and lowpass filtered at 10.0 kHz. This dubbing procedure resulted in a high degree of accuracy in aligning the auditory and visual components for the AV stimuli. Measurement indicated that the range of asynchronies between the A and V tokens was from +3 to -4 msec. The identical randomization schedule was used to construct AV stimuli in America and in Japan.

Subjects

Two independent replications of the full experiment were conducted. In each of the two experiments, 60 adult subjects with normal hearing and normal (or corrected-to-normal) vision were tested; the total N was thus 120. Sixty of these subjects were native speakers of American English; 60 were native speakers of Japanese. The American subjects were students at the University of Washington. The Japanese subjects were students from Doshisha University and staff members of the ATR.

Procedure

Across the two studies, six independent groups of 20 subjects (10 American and 10 Japanese) were randomly assigned to one of the six test conditions: American AV, Japanese AV, American A, Japanese A, American V, and Japanese V. In each test condition, subjects were presented with 8 practice trials followed by 80 test trials in a 25-minute session. Subjects were instructed to watch and listen (AV), watch (V), or listen (A) to each trial, and to identify whether the syllable they perceived began with /b/, /d/, /g/, /th/, /z/ or /bg/. Pilot testing determined that these responses accounted for 96% of subjects' perceptions of the stimuli in the AV condition. Subjects responded verbally to the experimenter, who recorded this information on an answer sheet. This enabled the subjects to keep their attention and/or vision focused on the stimuli at all times.

Testing was conducted in two locations: the Speech Research Laboratory at the University of Washington in Seattle, and the ATR Human Information Processing Research Laboratories in Kyoto. Subjects were tested in a sound-attenuated room. Subjects were positioned approximately 115 cm from the video monitor, which was seated on a table. The videotape was played on a videocassette machine located in an adjoining control room. Contrast and brightness controls were set at their midpoint levels; the audio signal was presented at a comfortable listening level of approximately 65 dB SPL (A scale), measured for the peak intensity of the vowel at the approximate location of the subject's head.

## RESULTS

The results of the two studies provide evidence that language experience affects auditory-visual speech perception. The data from the fusion and combination situations were each submitted to a 3-way analysis of

[illegible] the effects of Language Background (2) x Stimulus Tape (2) x Study (2).

Figure 1 displays the percentage of auditory-visual "fusion" responses that occurred for the two language groups for the two stimulus tapes. Fusion responses occur when audio /ba/ is presented in combination with video /ga/. As shown, Japanese subjects show a rather small percentage of fusion responses when watching a native speaker talk (about 40%); however, when watching the American speaker there is a dramatic increase in fusion responses (to 82%). American subjects show a moderate amount of fusion responses in both cases. These results tends to rule out two of the hypotheses raised at the outset. The stimulus factor is ruled out because neither stimulus tape elicited greater auditory-visual effects across the two subject populations. The subject factor is ruled out as the primary explanation for the data because neither subject population showed reduced auditory-visual speech effects across both stimulus tapes. The results appear to be more adequately accounted for by language experience factors.

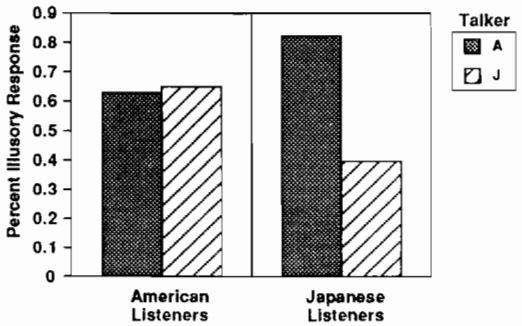


*Figure 1. Percent of "fusion" responses by American and Japanese subjects for American versus Japanese stimulus tapes.*

The analysis on fusion responses supports this conclusion. The analysis revealed a significant main effect of Stimulus Tape, $F$ (1, 72) = 6.41, $p < .05$, and a significant Language Background x Stimulus Tape interaction, $F$ (1, 72) = 7.75, $p < .01$. No other main effects or their interactions approached significance. Simple effects follow-up tests revealed that the Japanese subjects showed a highly significant effect of native- vs. foreign-language stimulus tape, $F$ (1, 72) = 14.13, $p < .001$; American subjects showed no significant difference in response to the American versus Japanese stimulus tapes. In summary, the perception of "fusion" responses showed a strong effect of language experience for Japanese subjects.

Figure 2 displays the data for the "combination" condition. In this condition, audio /ga/ is combined with video /ba/ and the perceived response is /bga/. As shown, both American and Japanese subjects show an effect of language experience. Both are more strongly affected by visual information when watching a foreign-language speaker. American subjects showed higher combination responses when watching and listening to the Japanese speaker. Japanese subjects showed significantly higher combination responses when watching and listening to the American speaker. Inspection of these results again suggests that the data cannot be explained by either overall stimulus or overall subject factors.

The analysis on combination responses revealed a significant main effect of Language, $F$ (1, 72) = 4.37, $p < .05$, and a highly significant Language Background x Stimulus Tape interaction, $F$ (1, 72) = 8.08, $p < .01$. No other main effects or interactions approached significance. Simple effects tests revealed that the Japanese subjects differed significantly in their response to the American and

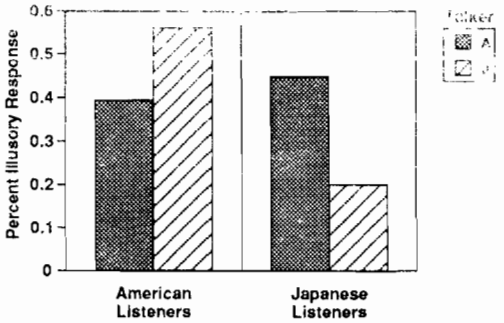


*Figure 2. Percent of "combination" responses by American and Japanese subjects for American versus Japanese stimulus tapes.*

Japanese stimulus tapes, $F$ (1, 72) = 5.75, $p < .05$, while American subjects' responses to the two tapes showed a nonsignificant trend in the same direction, $F$ (1, 72) = 2.63, $p = .109$. In the case of the American subjects, the two replications of the experiment produced different trends: In the first study, Americans revealed a strong tendency to show more auditory-visual effects when watching the foreign as opposed to the American speaker; in the replication experiment, this effect was greatly dampened. Further tests on the American subjects in this condition are therefore warranted.

While not completely analyzed, additional evidence of the effects of language experience on speech perception derives from data gathered in the A-only condition. Japanese subjects differ in their response to the American and Japanese A-only stimuli, with Japanese stimuli being correctly identified significantly more often than American stimuli. This was true for both the /b/ and /g/ consonants. American A-only data also accord with the "better on native-language" pattern for the consonant /g/. There is thus evidence that foreign- as opposed to native-language tokens are less well identified.

## DISCUSSION

In this experiment it was shown that both American and Japanese people combine auditory and visual speech information. However, the data from the current study reveal interactions between the language experience of the subject and the stimulus being presented.

The experiment provided evidence that performance on auditory-visual speech perception can be differentially affected depending upon whether one is watching a native versus foreign speaker. In particular, Japanese subjects were much more strongly affected by the visual information when it was spoken by a foreign speaker (see also [4]). This was true in both the fusion and the combination situations. What might cause this effect?

There are two potential explanations for language-experience effects, one fairly global and one more specific. The global explanation argues that people automatically pay more attention to a foreign speaker's mouth movements, assuming that the speaker will be more difficult to understand. This explanation would account for the results of the Japanese subjects. However, a global attentional hypothesis of this sort cannot account for the results of the American subjects, who, in the fusion situation, showed no significant difference between the American and Japanese stimulus tapes. Moreover, the attentional explanation cannot account for the fact that there are effects of language experience in the auditory-only condition.

A more complete explanation is that these cross-language auditory-visual effects are due to the fact that

speakers of different languages employ mentally stored information about the auditory and visual characteristics of their native language during the perception of speech, and that the speech of foreign speakers fails to match these stored auditory and visual representations [5]. In the case of Japanese listeners, the American auditory tokens of /b/ and /g/ are poorly identified. We propose that the tokens produced by Americans do not match the stored representations of Japanese subjects. This may result in increased attention to the visual signal. In the case of American subjects, the consonant /g/ fits this pattern; it is identified more poorly when spoken by a Japanese speaker as opposed to an American speaker. When auditory /g/ is combined with visual /b/ (the combination condition), Americans show stronger visual effects. We believe that differential auditory-visual speech effects are due to a mismatch between subjects' internal representations of auditory and visual speech information, based on their native-language experience, and the auditory and visual signals they see a foreign speaker produce.

In order to test this hypothesis, our research will focus on the auditory and visual characteristics of "ideal" tokens for American and Japanese subjects. Experiments are now underway in which three aspects of the stimulus are isolated: (a) the facial context that identifies a speaker as a member of one culture as opposed to another; (b) the purely auditory aspects of the phonetic stimulus; and (c) the purely visual aspects of the phonetic stimulus.

In current experiments we are manipulating the visual stimulus to determine which aspects of the face chiefly influence audio-visual speech perception in the two populations. We have developed a method for isolating specific articulatory "parts" of the dynamic facial stimulus. We are currently examining the role they play, individually and in combination, in producing multimodal speech perception effects in the two groups. In one condition, we have isolated the moving lips of the speaker (Fig. 3). The facial context in which the lips occurred was eliminated.

Preliminary results suggest that for American listeners tested in the native-language condition, lips alone are sufficient to reproduce the audio-visual effects. In other words, minimal visual information is sufficient to induce auditory-visual effects. This research on the necessary and sufficient conditions for combining information across modalities will be informative for both theory development and practical applications that use visual information in multimodal speech technology.

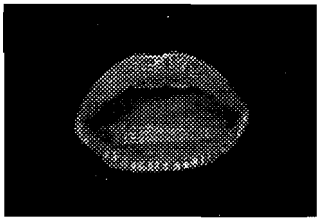

*Figure 3. Visual stimulus used to study the contribution of lip movements in multimodal speech perception.*

REFERENCES


[1] McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264: 746-748. Roberts, M. & Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaption in speech perception is purely auditory. *Perception & Psychophysics*, 30: 309-314. 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. Green, K. P. & Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*, 45: 34-42. 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.

[2] Green, K. P., Kuhl, P. K., Meltzoff, A. N. & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, 50: 524-536.

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

[4] 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.

[5] Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50: 93-107. Kuhl, P. K. (1992). Infants' perception and representation of speech: Development of a new theory. In J. J. Ohala, T. M. Nearey, B. L. Derwing, M. M. Hodge, & G. E. Wiebe (Eds.), *Proceedings of the International Conference on Spoken Language Processing* (pp. 449-456). Edmonton, Alberta: University of Alberta. Kuhl, P. K. (1993). Innate predispositions and the effects of experience in speech perception: The native language magnet theory. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. MacNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 259-274). Boston: Kluwer Academic Publishers. Kuhl, P. K., & Meltzoff, A. N. (In press). Evolution, nativism, and learning in the development of language and speech. In M. Gopnik (Ed.), *The biological basis of language*. New York: Oxford University Press.

[6] W. Strange, ed. (1994). *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research*. Timonium, MD: York Press.

[7] Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N. & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255: 606-608.