The McGurk Effect in Infants

Lawrence D. Rosenblum
University of California, Riverside

Mark A. Schmuckler
University of Toronto, Scarborough Campus

Jennifer A. Johnson
University of California, Riverside

Published as: Rosenblum, L.D., Schmuckler, M.A., & Johnson, J.A. (1997). The McGurk effect in infants. Perception & Psychophysics, 59(3), 347-357.

Abstract

In the McGurk effect, perceptual identification of auditory speech syllables is influenced by simultaneous presentation of discrepant visible speech syllables. While this effect has been shown in subjects of different ages and various native language backgrounds, no McGurk tests have been conducted with pre-linguistic infants. A series of experiments tested for the McGurk effect in 5-month-old English-exposed infants. Infants were first gaze-habituated to an audiovisual /va/. They were then presented two different dishabituation stimuli: audio /ba/-visual /va/ (perceived by adults as /va/); and audio /da/-visual /va/ (perceived by adults as /da/). The infants showed generalization from the audiovisual /va/ to the audio /ba/-visual /va/ stimulus but not to the audio /da/-visual /va/ stimulus. Follow-up experiments revealed that these generalization differences were not due to either a general preference for the audio /da/-visual /va/ stimulus or to the auditory similarity of /ba/ to /va/ relative to /da/. These results suggest that the infants were visually influenced in the same way as English-speaking adults.

A great deal of research has demonstrated the salience of audiovisual speech perception. Seeing the face of a speaker can significantly embellish a degraded or noisy auditory speech signal so that it functionally raises the signal-to-noise ratio as much as 22 dB (Sumby and Pollack, 1954; for a review see Rosenblum, Johnson, and Saldaña, under review). Visual speech information can also enhance comprehension of clear speech signals containing complicated content or produced with a heavy foreign accent (Reisberg, McLean, & Goldfield, 1987). There is also evidence that visual speech information helps facilitate the acquisition of certain segments in young children (e.g., Mills, 1987; see also Legerstee, 1990).

The salience of audiovisual speech is also evidenced in findings that its integration is automatic. For example, a phenomenon known as the McGurk effect demonstrates that neither informational stream can be ignored (McGurk and MacDonald, 1976). For the effect, auditory syllables are synchronously dubbed with discrepant visual syllables. These dubbed syllables are then presented to subjects who are asked to report what they hear. Most subjects report hearing a syllable which has been visually influenced in some way so that there is either visual dominance (e.g., audio /ba/-visual /va/ is heard as 'va') or a fusion of the auditory and visual syllables (e.g., audio /ba/-visual /ga/ is heard as 'da' or 'tha'). Subjects in these experiments usually have no awareness of the audiovisual discrepancy in the stimuli and cannot discern the auditory and visual contributions to what they 'hear'. Furthermore, integration occurs even when subjects are explicitly told of the dubbing procedure, or when they are asked to attend to only one of the information streams (Massaro, 1987).

Recent research has shown that the McGurk effect is sustained under substantial changes in the visual stimulus. The effect works with both elaborate and schematic synthetic visual stimuli (e.g., Massaro and Cohen, 1990; Summerfield, Macleod, McGrath, and Brooke, 1989). Also,

integration can occur over auditory and visual signals generated by speakers of different gender (Green, Kuhl, Meltzoff, and Stevens, 1991) indicating that the two information streams need not appear to emanate from the same source. Finally, images which involve no identifiable facial features but are comprised only of a few illuminated dots on a darkened face can also influence heard speech (Rosenblum and Saldaña, 1996). This last study also revealed that visual influences can occur for observers who do not recognize the images as a face. Thus, the McGurk effect is robust to the extent that it holds over substantial visual stimulus changes and is maintained regardless of what the observer knows about the stimuli.

Research on the McGurk effect has also tested young children (McGurk and MacDonald, 1976; Massaro, 1984; Massaro, Thompson, Barren, and Laren, 1986; Massaro, 1987; and Boliek, Green, Fohr, & Obrzut, submitted). Interestingly, evidence suggests that the effect is somewhat weaker in children. McGurk and MacDonald (1976) reported that two groups of English speaking children (3-5 years and 7-8 years old) displayed a smaller overall visual influence than adult subjects. However for some audiovisual tokens, the younger children displayed as many fused responses as adults. Similarly, Massaro, and his colleagues (Massaro, et al 1986) have shown that preschool children show less of a visual influence of /ba/ on /da/ (and /da/ on /ba/) than adults. However, these researchers also showed that the reduced effects are attributable to developmental differences in visual information sensitivity rather than differences in integration processes.

Thus, while the strength of the McGurk effect might increase with development, there is evidence that audiovisual speech integration does occur in young children. Clearly, the issue of the development of the McGurk effect would benefit from research testing younger, prelinguistic infants. This is the goal of the current paper.

Infant sensitivity to audiovisual correspondence

While there have been no direct tests of the McGurk effect in pre-linguistic infants[1], there is a good deal of evidence that infants are sensitive to audiovisual correspondences in speech (e.g., Aronson and Rosenbloom, 1971; Dodd, 1979; Spelke and Owsley, 1979; Kuhl and Meltzoff, 1982; 1984; Kuhl, Williams, and Meltzoff, 1991). Early research showed that infants are sensitive to spatial correspondences in audio and visual speech (e.g., Aronson and Rosenbloom, 1971; Spelke and Owsley, 1979). In one example, Aronson and Rosenbloom (1971) observed that 1-2 month olds become visibly distressed when a mother's voice is displaced away from her face. Research has also shown that infants are sensitive to temporal synchrony in audiovisual speech (e.g., Dodd, 1979; Pickens, Field, Nawrocki, Martinez, Soutullo, & Gonzalez, 1994; Spelke & Cartelyou, 1980; Walker, 1982). Using a gaze preference procedure, Dodd (1979) demonstrated that 3-4 month olds attend longer to audiovisual speech which is in synchrony than to speech presented out-of-synchrony by 400 ms. She interprets this as evidence that infants are aware of some congruence between lip movements and speech sounds.

More recent research suggests that infants are sensitive to phonetic correspondences in audiovisual speech. Kuhl and Meltzoff (1982; 1984) used a preferential gaze procedure to test whether 4 month old infants were sensitive to audiovisual correspondences for the vowels /i/ and /a/. They found that for both vowels, infants looked longer at the face which matched the vowel presented auditorily. Additional research has replicated these findings with /i/ and /u/ vowels (Kuhl and Meltzoff, 1988), and disyllables such as /mama/ and /lulu/ (MacKain, Studdert-Kennedy, Spieker, & Stern, 1983). Similar effects were found using an operant choice sucking procedure (Walton and Bower, 1993). This study showed that infants ranging in age from 1 to 14 months perform more sucks to audiovisual compatible than incompatible vowels. Finally, Legerstee (1990) has found that 3-4 month old infants will only imitate audiovisually compatible vowels (/a/ - /u/) and not stimuli that are dubbed to be incompatible. She interprets this finding as evidence that multimodal information is useful for speech acquisition.

Kuhl, Meltzoff, and their colleagues have also found that infants show no match preference if the auditory component is comprised of simple sinewaves or sinewave complexes (Kuhl and Meltzoff, 1984; Kuhl, Williams, and Meltzoff, 1991). These results suggest that the gaze preferences observed with intact auditory speech stimuli are not based on simple temporal or

amplitude commonalties between audio and visual streams. Rather, these authors suggest that the prior results were likely based on a matching of the spectral information contained in the auditory component. They claim that since spectral information (unlike temporal and amplitude envelope dimensions) is particularly dependent on articulatory changes, sensitivity to its relationship with visual speech information implicates a linkage to phonetic primitives.

Based on these observations, Kuhl and Meltzoff (1988) propose that perception of audiovisual speech correspondence involves underlying representations which specify both auditory and visual concomitants to phonetic units: speech is intermodally—or amodally—represented (Kuhl and Meltzoff, 1988; Meltzoff and Kuhl, 1994). Kuhl and Meltzoff (1984) propose two possible processes by which audiovisual information can be matched to amodal representations. First, an 'identity match' would involve input for which the exact same phonetic information is specified in both the audio and visual modalities. Thus, the process of sensing correspondence would involve detecting a match between identical information at either the segment or featural level. Alternatively, the two information streams need not specify identical phonetic information, but could be related through mutual contact with a common phonetic representation. This higher order, 'supramodal' representation could act as a mediator between nonidentical auditory and visual information. Supramodal representations could allow for multiplicitive audiovisual speech percepts such as is observed in the McGurk effect (Meltzoff and Kuhl, 1994).

In being used as an explanation for infant sensitivity to audiovisual correspondence, the supramodal account implies that infants should also be able to integrate nonidentical speech information. While there is substantial evidence that infants can sense correspondence, there is no current evidence that they actually integrate audiovisual speech. Clearly, a McGurk demonstration with infants would help in this regard. Observing a McGurk effect in infants would provide stronger evidence that infants represent speech amodally and that they perform a supramodal rather than identity match with audiovisual input.

The following experiments test whether a McGurk effect is evident in pre-linguistic infants. For these experiments, 5 month old infants are tested under an infant-control habituation-of-looking time procedure (e.g., Best, McRoberts, & Sithole, 1988; Horowitz, 1975; Horowitz, Paden, Bhana, and Self, 1972). This procedure tests the degree to which infants generalize to various test stimuli after habituation to an initial stimulus. For the first experiment, this procedure was used to test infant discrimination of audio /va/-visual /va/ from audio /ba/-visual/va/ and from audio /da/-visual /va/. Previous research in our laboratory has shown that an audio /ba/-visual /va/ is 'heard' as /va/ up to 98% of the time with adult observers (Saldaña and Rosenblum, 1993; 1994; Rosenblum and Saldaña, 1992; and 1996). In contrast, there is evidence that an audio /da/-visual /va/ rarely displays a visual influence and is heard by adults as /da/ over 88% of the time (Manuel, Repp, Liberman, and Studdert Kennedy, 1983). If infants also display the typical visual influence, then they should perceive the audio /ba/-visual /va/ as 'va' and audio /da/-visual /va/ as 'da'. This outcome would be reflected by infants generalizing habituation from an audiovisual /va/, to the audio /ba/-visual /va/ stimulus but not to the audio /da/-visual /va/ stimulus.

EXPERIMENT 1

Method

Participants

The participants consisted of twenty 5-month-old infants (10 male) with a mean age of 21.4 weeks, and a range from 20.5 to 22 weeks. All participants lived in a home where English was the primary language spoken (more than 50% of the time). Seventeen of these subjects were raised in a home where English was spoken 100% of the time. Two infants were raised in a home where French was the second language (which was spoken 5 and 25% of the time), and one infant was raised in a home where Philippino was the second language (spoken 20% of the time). Twelve other infants participated in this study, but their data were not considered usable due to fussiness by the infant during the study (7), experimenter error in running the study (2), equipment failure (2), and the parent making noise during the experiment (1). All infants were recruited through

local hospitals, daycares, or on the basis of other public records, and were drawn from the Scarborough, Ontario community. Detailed information concerning SES and ethnic background, other than the language(s) children heard on a regular basis, was not collected.

Stimuli

An American-English speaking Caucasian male actor was videotaped with a Panasonic PVS350 camcorder from a distance of 5 feet. The actor was told to maintain an even intonation, tempo, and vocal intensity while producing the stimulus syllable /va/ and was also told to avoid blinking during his articulations. The actor was recorded with lights focused on the face such that minimal shadowing occurred. His head was secured by a metal brace to inhibit movement. In the recorded image, the actor's entire head was visible against a black cloth background. However, in the image shown to infants, the top part of the actors head (down to his nose-tip) was covered by means of black cardboard placed over the top of the video monitor. This modification was performed because pilot experiments revealed that infants found the constant (unblicking) stare of the actor distracting and, in some cases, upsetting. A single clear visual exemplar of /va/ was selected for use in all audiovisual conditions. This edited token lasted 2 s and included 1496 ms of actual articulatory movement. The movement was initiated at the labiodental position and ended with an open mouth (the lips were never closed during the shown movement).

The auditory syllables /ba/, /va/, and /da/ were generated on an AMC 486 computer using a Klatt80 speech synthesis program. Parameters were largely based on those suggested by Klatt (1980). The three syllables were synthesized so that they shared the same duration (520 ms), fundamental frequency contour, amplitude of voicing, and vowel portion. The fundamental frequency of the syllables started at 106 Hz, rose to 116 Hz at 40 ms into the syllable, and then declined to 71 Hz by the syllable's end. The amplitude of voicing began at 41 dB, then rose to 60 dB at 70 ms, and then decreased to 10 dB by the end of the syllables. The steady state vowel portion of the syllables began at 85 ms after onset. The F1 frequency for the steady state portion was 700 Hz, while the F2 frequency was 1220 Hz and the F3 frequency was 2600. The F1 bandwidth (B1) for the steady state portion was 130 Hz, while the F2 bandwidth (B2) was 70 Hz, and the F3 bandwidth (B3) was 160 Hz. Other synthesis parameters that differed across these syllables are listed in Table 1. This table lists the initial values for the formant transitions which then progressed over the first 85 ms to the steady-state values listed above. Four additional 'variants' of each of the three syllables were also synthesized with F0 contours that were either 10 or 20 Hz above or below that of the original syllables. This resulted in a total of 15 different auditory tokens (5 frequency contours for each of the three syllable types).

Table 1. Initial values of synthesis parameters for auditory syllables (see text for details).

Syllable	F1 (Hz)	F2 (Hz)	F3 (Hz)	B1 (Hz)	B2 (Hz)	B3 (Hz)
/va/	220	1000	2080	60	90	120
/ba/	200	1000	2150	60	110	130
/da/	200	2000	2600	60	110	170

An AMC 486/33 computer and two Panasonic video recorders were used to dub the video and audio signals. The original videotape was played from one video recorder. The auditory files from the computer were output to a second videotape in the other video recorder whenever a voice activated circuit was closed. The output of the audio channel from the original videotape served as the input to the circuit such that when an audio signal from the original tape was produced, a stored audio file was immediately output to the second videotape, resulting in a new synchronous audiovisual token. The onset lag time for dubbing for all tokens was found to be no greater than 9.4 ms, well below the 80 ms range required for observers to detect an audio-visual asynchrony (McGrath and Summerfield, 1985). Because the same dubbing procedure was used for each audiovisual token, the onset and offset asynchrony for all stimuli were the same (within this 9.4 ms range). (Although it is acknowledged that the *perceived* audiovisual synchrony might still be different across the tokens, all of the stimuli seemed equally synchronous to the experimenters. Nevertheless, this issue is re-addressed in Experiment 2.)

The audiovisual stimuli were recorded in three different two-minute blocks. One block included two-minutes of the visual /va/ stimulus paired with a randomized ordering of the five different /va/ auditory tokens. A second block was generated in a similar manner except it involved the visual /va/ stimulus paired with the same randomized ordering of the five different auditory /ba/ syllables. The third block was comprised of the visual /va/ paired with the 5 auditory /da/ stimuli (again using the same random ordering of the different F0 stimuli). To us, as well as (native English speaking) adult pilot participants, the audio /va/-visual /va/ stimuli were perceived as /va/, while the audio /ba/-visual /va/ were perceived as /va/ and the audio /da/-visual /va/ were perceived as /da/. These pilot findings replicate previous results with similar stimuli (e.g., Manuel, Repp, Liberman, and Studdert Kennedy, 1983; Saldaña and Rosenblum, 1993; 1994; Rosenblum and Saldaña, 1992; and 1996). For all of the blocks, syllables were presented for a total of 2 seconds each. The syllables were recorded in immediate succession (with no black screen between them), so that the face was always visible throughout the entire two-minute block. The blocks were recorded onto the videotape in the order of 1) audio /va/-visual /va/; 2) audio /ba/-visual /va/; and 3) audio /da/-visual /va/ and with about 15 seconds in-between each block.

Design

An infant-control habituation-of-looking time procedure (e.g., Horowitz, et al, 1972) was used to test discrimination of the auditory-visual displays. In this procedure, a stimulus (called the habituation stimulus) is presented repeatedly until the infant becomes disinterested and visual attention to this stimulus drops. In this study, the point of disinterest was defined as the total looking time on two consecutive trials that did not amount to more than half the looking time on the first pair of trials summing to more than 12 seconds. Once infants reach this criterion level, they were shown the test—or dishabituation—stimuli. The amount of looking time at these test stimuli was the main dependent measure, and generally indexes the similarity perceived by infants between the habituation and test stimuli. If infants perceive habituation and test stimuli to be comparable on some level, little renewed interest is observed. If habituation and test stimuli differ in some way, infants show increased visual attention to the test stimuli. In this study, we paired a constant visual stimulus with different auditory stimuli for habituation and test trials; accordingly, we used change in looking time towards a visual display as indicative of auditory discrimination. While using an auditory-visual combination in an habituation paradigm is less common, the paradigm has been reliably used to study auditory discrimination in infancy (e.g., Best, McRoberts, & Sithole, 1988; Demany, McKenzie, and Vurpillot, 1977; and Horowitz, 1975).

In this experiment, the habituation stimulus consisted of the audio /va/-visual /va/. Two test stimuli were used: one consisted of the audio /ba/-visual /va/, while the other consisted of the audio /da/-visual /va/. The two test stimuli were alternated for two trials each, with half the infants presented with the test stimuli in the order audio /ba/-visual /va/; audio /da/-visual /va/; audio /ba/-visual /va/; audio /da/-visual /va/, with the remaining infants presented with the test stimuli in the order audio /da/-visual /va/; audio /ba/-visual /va/; audio /da/-visual /va/; audio /ba/-visual /va/.

Apparatus and procedure

Each infant was tested in a small, experimental room covered with acoustic paneling, and participated while seated on his or her parent's lap. Parents were asked not to interact with their infant during the experiment, nor to influence their looking behavior in any way. To avoid the parents noticing a change in the audio portion of the stimulus (and potentially biasing the infant's attention) all parents wore headphones playing masking music during the experimental session. The level of the headphones was set by the experimenters so that the stimuli could not be heard when the music was playing. None of the parents reported hearing the auditory stimuli.

While infants were seated on their parent's lap, they faced a Sony CVM-195 13 in video monitor, positioned approximately 12 in away from the parent/child, sitting atop of a table. This video monitor was used to present the video portion of each stimulus. A single Boss MA-12V micro-monitor loudspeaker was positioned on top of the video monitor, and was used to present the audio portion of each stimulus. All auditory stimuli were presented at comfortable listening levels which were 73 dB, 72 dB, and 72 dB SPL (A-weighted) for the /va/, /ba/, and /da/ tokens respectively. Both visual and audio portions of the stimuli were presented to infants using a JVC-BR8600U professional editing video recorder. A JVC GS-CD1U video camera was positioned

underneath the video monitor, allowing for a focus on and video taping of the infant's face. Stimulus presentations and on-line computations of looking time were controlled by an IBM-compatible 286 PC located in the adjacent control room.

An experimenter seated in the control room began each trial by turning on the visual portion of the display once the child appeared awake and alert. This observer coded the child's visual fixations by viewing a Sony CVM-950 video monitor located in the control room, which received a picture from the JVC camera located in the experimental room. When the infant fixated the (silent) moving face on the video monitor, the experimenter pressed a key on a computer mouse to record the time looking at the visual display. Pressing this mouse key automatically turned on the audio soundtrack, thereby providing the infant with a combined auditory-visual stimulus. When the child stopped looking at the visual display, the experimenter released the mouse button, which both ended the timing of visual fixations and turned off the audio soundtrack. The video portion of the stimulus was terminated (and the trial ended) two seconds after the infant stopped looking at the display. In other words, a 2-sec lookaway criterion was used to terminate the trial, although the audio soundtrack was stopped immediately when the infant looked away. This contingent toggling on and off of the auditory stimulus was performed based on the fact that the McGurk effect relies critically on the simultaneous presentation of both auditory and visual stimuli.

When the trial ended, the computer either rewound or forwarded the video tape to move to the next trial. During this time, the screen was dark and there was no sound other than any produced by the infant him/herself. Because of the rewinding or forwarding of the stimulus tape between trials, interstimulus intervals were about 5 to 10 seconds, depending on the length of the fixation for the preceding trial. The fact that the ISI varied was unavoidable given the nature of the equipment. The entire experimental session lasted approximately 20 minutes.

The observer in the control room coded all visual fixations during the experiment, based on the infant's direction of gaze. As mentioned, the camera was positioned in front of the infant and underneath the video monitor which presented the visual stimuli. Thus, the camera recorded a full-frontal view of the infant's face. Fixations directly forward and slightly up indicated that the infant was looking at the video monitor.

Although this first observer was blind to the order of test trials, he/she could have determined this order on the basis of the rewinding versus forwarding of the stimulus video tape. To ensure that looking times were not biased by this potential knowledge, a second observer (who was also unaware of the order of test trials) provided reliability codings of the looking times using the videotaped recordings of the infants' faces. This second observer knew when the video portion of the stimulus tape began (so as to know when there was a visual stimulus available to be looked at), although he/she was not aware of the onset of the audio portion, nor of the end of the trial. Averaging across both habituation and test trials, the mean absolute difference between the original measure and the reliability coding of visual fixations was 0.96 seconds, with a standard deviation of 1.59. Moreover, original and reliability codings were strongly correlated, with $r\ (184) = 0.996$, $p < .001$. Looking at the test trials only, the mean absolute difference between original and reliability codings was 0.69 seconds, with a standard deviation of 0.998. These two sets of codings were also strongly correlated, with $r\ (78) = 0.998$, $p < .001$.

Results and Discussion

The principal goal of the data analysis was to determine whether infants discriminated among the audio /va/-visual /va/, audio /ba/-visual /va/, and audio /da/-visual /va/ displays. Discrimination was assessed by comparing the looking time (in seconds) to the final two habituation trials (audio /va/-visual /va/), the two audio /ba/-visual /va/ test trials, and the two audio /da/-visual /va/ test trials; a significant increase in looking time to either of the test displays, relative to the habituation displays, indicates discrimination. The mean data are presented graphically in Figure 1. Infants looked at the final two audio /va/-visual /va/ habituation stimuli for 11.96 sec (SD = 7.08), at the two audio /ba/-visual /va/ test trials for 13.00 sec (SD = 8.47), and at the two audio /da/-visual /va/ test trials for 18.83 sec (SD = 11.99). Looking times were compared using a two-way analysis of variance (ANOVA), with the within-subject variable of trial type (audio /va/-visual /va/, audio /ba/-

visual /va/, audio /da/-visual /va/), and the between-subject variable of test order (audio /ba/-visual /va/; audio /da/-visual /va/ vs. audio /da/-visual /va/; audio /ba/-visual /va/). This analysis revealed a significant effect for trial type, $F(2, 36) = 3.71$, $p < .05$. There was no effect for order, $F(1, 18) = 0.002$, ns., and no interaction between the trial type and order variables, $F(2, 36) = 1.58$, ns.

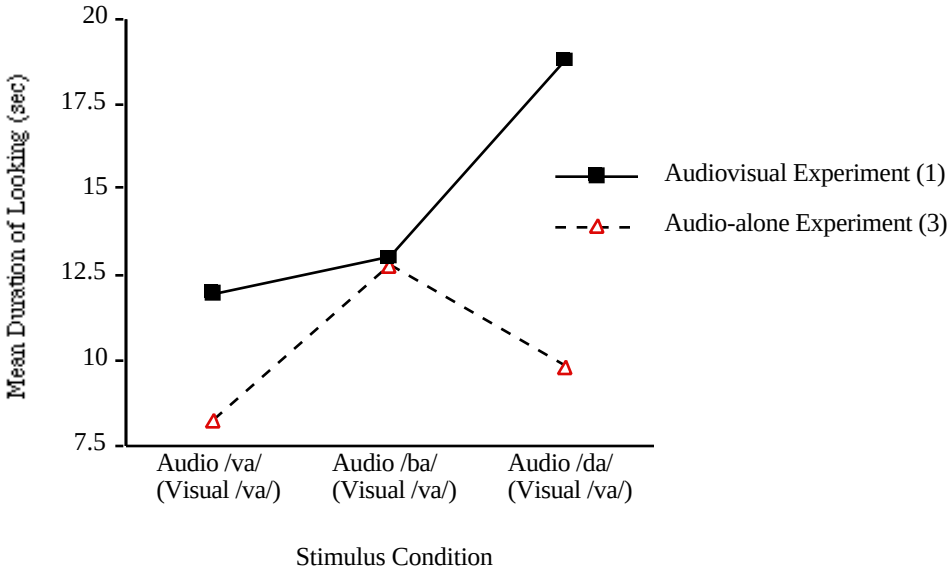

Figure 1. Mean duration of looking at the audiovisual stimuli of Experiment 1 and the audio-alone stimuli of Experiment 3 (see text for details).

Subsequent analyses compared the means for these trials, using Newman-Keuls tests for posteriori pairwise comparisons. These tests revealed that the audio /da/-visual /va/ test trials differed significantly from the final audio /va/-visual /va/ habituation trials, (mean difference = 6.86 sec, $p < .05$) and the audio /ba/-visual /va/ test trials (mean difference = 5.83 sec, $p < .05$). In contrast, there was no difference in looking times between the final audio /va/-visual /va/ habituation trials and the audio /ba/-visual /va/ test trials (mean difference = 1.04 sec, ns.).

These results indicate that the infants generalized to the audio /ba/-visual /va/ stimulus, but not to the audio /da/-visual /va/ stimulus. Potentially then, infants perceived both the audio /va/-visual /va/ and audio /ba/-visual /va/ stimuli as similar while they perceived the audio /da/-visual /va/ stimulus as different accounting for the observed dishabituation with this latter token. It could be then, that the infants perceived both the audio /va/-visual /va/ and audio /ba/-visual /va/ tokens as 'va' thereby displaying a McGurk-type visual influence. Before drawing this conclusion however, alternative explanations for these results must be entertained. For example, it could be that the infants gazed longer at the audio /da/-visual /va/ than audio /ba/-visual /va/ based on some general preference for the audio /da/-visual /va/ token. As an example, the infants might have found the audio /da/-visual /va/ to be more (or less) audiovisually compatible than the audio /ba/-visual /va/ leading them to gaze at this former token for longer durations[2]. It is possible that infants perceived the two syllable combinations as differentially compatible in, say, audiovisual asynchrony or phonetic concordance. If so, differential audiovisual compatibility could account for the longer audio /da/-visual /va/ stimulus gaze times observed in Experiment 1. While this explanation assumes that the infants were sensitive to audiovisual compatibility, it does not imply that they were visually influenced with the audio /ba/-visual /va/ token.

In order to determine whether infants have a general preference for the audio /da/-visual /va/ token, a gaze preference control experiment was conducted with the audio /da/-visual /va/ and audio /ba/-visual /va/ tokens. If the audio /da/-visual /va/ is more compelling because of differential audiovisual compatibility (or any other dimension), then infants should spend more time looking at this stimulus than at the audiovisual audio /ba/-visual /va/ stimulus when each is presented without a prior habituation phase. However, if the audio /da/-visual /va/ stimulus is not inherently more interesting, looking times to the two syllable combinations should be approximately equal.

EXPERIMENT 2

Methods

Participants

The participants consisted of twenty 5-month-old infants (11 male) with a mean age of 20.7 weeks, and a range from 20 to 21.7 weeks. All participants lived in a home where English was spoken at least 50% of the time. Of these infants, 12 were raised in a home where English was spoken 100% of the time. Two infants were raised in homes where French was spoken 5% of the time, one infant was raised in a home where German was spoken 30% of the time and one was raised in a home where Cantonese was spoken less than 5% of the time. The remaining four participating infants were raised in homes where the other language was spoken 50% of the time. These languages were Arabic, Jamaican Patois, Gujrati, and Tamil. Two additional infants participated in this study, but their data were not considered usable due to equipment failure while running the study. All infants were recruited via the same means as in the previous experiment. Detailed information concerning ethnic background and SES was not collected.

Stimuli, Design, Apparatus, and Procedure

The stimulus displays consisted of the audio /ba/-visual /va/ and the audio /da/-visual /va/ tokens used in Experiment 1. As in the previous study, the top half of the speaker's face was covered using a piece of black cardboard attached to the video monitor. All infants received three trials with the audio /ba/-visual /va/ stimulus and three trials with the audio /da/-visual /va/ stimulus; these trials were alternated. Half of the infants saw the auditory-visual stimuli in the order audio /ba/-visual /va/; audio /da/-visual /va/; audio /ba/-visual /va/; audio /da/-visual /va/; audio /ba/-visual /va/; audio /da/-visual /va/, and the remaining infants received the stimuli in the reverse order.

Each infant was tested using the same experimental apparatus as in the previous experiment. The presentation of the six trials (3 audio /ba/-visual /va/ stimuli and 3 audio /da/-visual /va/ stimuli) were identical to Experiment 1, with the experimenter toggling the auditory portion of the stimulus on and off dependent upon the infant's fixation of the visual display. When the infant looked away for 2 s, the trial ended and the visual stimulus was turned off. The time looking towards the two displays comprised the dependent measure of this study.

Once again, a second observer performed reliability measures on the infants' visual fixations from the videotape recordings. Across all 6 trials, the mean absolute difference between the original and reliability coding of visual fixations was 2.75 s, with a standard deviation of 4.9 s. Original and reliability codings were strongly correlated, with $r(118) = 0.98$, $p < .001$.

Results and Discussion

The principal goal of the data analysis was to determine whether infants looked preferentially at either the audio /ba/-visual /va/ or audio /da/-visual /va/ stimulus. On average, infants looked at the audio /ba/-visual /va/ for 20.88 sec (SD = 23.70) and at the audio /da/-visual /va/ for 22.81 sec (SD = 27.59). Comparisons were made using a three-way ANOVA, with the within-subject variables of trial type (audio /ba/-visual /va/ versus audio /da/-visual /va/) and repetition (trial 1 versus trial 2 versus trial 3), and the between-subject variable of order (audio /ba/-visual /va/ followed by audio /da/-visual /va/ versus audio /da/-visual /va/followed by audio /ba/-visual /va/). The only significant result produced by this analysis was a main effect for repetition, $F(2,26) = 9.86$, $p < .001$, with increased looking times on the first trial (M = 35.1 s) relative to the second and third trials (M's = 15.1 s and 15.3 s, respectively). Most importantly, there was no main effect for

stimulus type, $F(1,18) = 0.32$, $p = 0.58$ and no interaction of stimulus type and repetition, $F(2,36) = 0.68$, $p = 0.51$.[3]

The interpretation of this result is straightforward: Infants found the audio /ba/-visual /va/ and audio /da/-visual /va/ stimuli equally interesting. This finding supports the idea that infants' greater looking time to the audio /da/-visual /va/ display in Experiment 1 was not based on any inherent preference for that display. As such, these results provide further support that the results of Experiment 1 were based on a perceived similarity between audio /va/-visual /va/ and audio /ba/-visual /va/ displays, and a corresponding dissimilarity between audio /va/-visual /va/ and audio /da/-visual /va/ displays.

However, before drawing this general conclusion, one additional explanation should be entertained. It could be that the observed looking times in Experiment 1 were not a result of a visual influence, but instead a result of the relative auditory similarity among the tokens. It could be that infants showed generalization to the audio /ba/-visual /va/ stimuli and not the audio /da/-visual /va/ stimuli simply because audio /va/ is more similar to audio /ba/ than it is to audio /da/.

To determine if our results are a consequence of relative auditory similarity, a third experiment was conducted. In Experiment 3, the auditory stimuli were the same as in Experiment 1. The visual stimulus however, was comprised of a neutral, unmoving face which was dubbed with each of the auditory tokens. (Pilot studies with adults demonstrated that this visual stimulus had no effect on the perceived syllables.) If the effects observed in Experiment 1 were not from a visual influence, but were instead a consequence of relative auditory similarity, then we would expect to see the same patterning of results with the neutral face stimuli implemented in Experiment 3. In other words, generalization should occur from the auditory /va/ to the auditory /ba/, but not to the auditory /da/.

EXPERIMENT 3

Method

Participants

The final sample of participants consisted of twenty 5-month-old infants (13 male) with a mean age of 21.2 weeks, and a range from 20 to 22 weeks. All of the infants were raised in a home where English was spoken at least 50% of the time. For seventeen of these infants, English was spoken in the home 100% of the time. One infant was raised in a home where Dutch was spoken 2% of the time, and was one raised in a home where Philippino was spoken 20% of the time. Finally, one infant was raised in a home where Jamaican Patois was spoken less than 5% of the time. Nine other infants participated in this study, but their data were not considered usable due to fussiness by the infant during the study (5), equipment failure during the experiment (3), and a sibling who refused to be separated from his/her baby sister during the experiment (1). All infants were recruited via the same means as in Experiment 1.

Stimuli, Design, Apparatus, and Procedure

The experimental design, apparatus, stimuli, and procedure employed in this experiment were identical to that of Experiment 1 with the following exceptions. For the auditory stimuli, infants were again presented a /va/ during habituation, and a /ba/ and /da/ during dishabituation. In this experiment however, the visual stimulus consisted of a neutral, unmoving (slightly smiling) face. The face was that of the actor used in Experiment 1. This face was dubbed with each of the auditory stimuli and a presentation tape was produced in the same manner as for Experiment 1. The presentation of stimuli was identical to Experiment 1, with looking time again the dependent measure of interest. Two presentations of the test trials were alternated during dishabituation, with half the infants receiving the tests in the order /ba/ - /da/ - /ba/ - /da/, and the other half of the infants receiving the test trials in the order /da/ - /ba/ - /da/ - /ba/.

As in Experiments 1 and 2, a second observer performed reliability measures on the infants' visual fixations from the videotape recordings. Across both habituation and test trials, the mean absolute difference between the original and the reliability coding of visual fixations was 1.18 seconds, with a standard deviation of 2.15. Original and reliability codings were strongly correlated, with an $r(199) = 0.990$, $p < .001$. For the test trials only, the mean absolute difference

between original and reliability codings was 0.99 seconds, with a standard deviation of 1.38. These two sets of codings were also strongly correlated, with an $r(78) = 0.985$, $p < .001$.

Results and Discussion

As in Experiment 1, discrimination was assessed by comparing the looking times to the final two habituation trials (/va/), the two /ba/ test trials, and the two /da/ test trials. The mean data are presented graphically in Figure 1. Infants looked at the final two /va/ habituation stimuli for 8.25 sec (SD = 4.47), at the two /ba/ test trials for 12.79 sec (SD = 9.13), and at the two /da/ test trials for 9.84 sec (SD = 5.62). Comparisons were accomplished using a two-way ANOVA, with the within-subject variable of trial type (/va/, /ba/, /da/), and the between-subject variable of test order: 1) /ba/; 2) /da/ vs. 1) /da/; 2) /ba/. This analysis revealed a significant effect for trial type, $F(2, 36) = 4.13$, $p < .05$. There was no effect for order, $F(1, 18) = 0.72$, ns., and no interaction between the trial type and order variables, $F(2, 36) = 2.35$, ns. Subsequent Newman-Keuls tests revealed a significant difference in looking times between the /va/ habituation stimuli and the /ba/ test stimuli (mean difference = 4.54 sec, $p < .05$), but no difference between /va/ stimuli and the /da/ stimuli (mean difference = 1.59 sec, ns.), and no difference between /ba/ and /da/ test stimuli (mean difference = 2.95 sec, ns.).

These results indicate that the infants generalized from audio /va/ to the audio /da/ token, but not to the audio /ba/ token. This pattern of results is quite different from that obtained in Experiment 1 suggesting that the results of that experiment were not a consequence of relative auditory similarity among the tokens. If anything, the results of Experiment 3 are suggestive that infants found the audio /va/ more similar to the audio /da/ than the audio /ba/. In precluding an auditory similarity explanation, the results of Experiment 3 support the conclusion that infants in Experiment 1 were visually-influenced in a McGurk-type manner. In this sense, Experiment 3 was successful in its designed purpose.

However, a less critical component of the Experiment 3 findings was somewhat surprising. The data reveal that infants had trouble discriminating our audio /va/ and /da/ stimuli. Although there is no evidence in the literature that infants can make this specific discrimination, these syllables differ in both place and manner of articulation. A great deal of research has shown that infants can make discriminations along these dimensions (for reviews see Eimas & Tartter, 1979; Kuhl, 1979; 1987). Furthermore, evidence has shown that infants can make virtually all of the consonantal discriminations which have been tested (Kuhl, 1987) including /va/ from /sa/ (Eilers & Minifie, 1975; Eilers, Wilson, & Moore, 1977) and /va/ from / a/ (Levitt, Jusczyk, Murray, & Carden, 1988). (There is some evidence, however, that 5 and 6 year old children have some trouble discriminating /v/ and / / segments [Massaro, 1987; Snyder & Pope, 1970]). Thus, based on the infant speech literature, one would expect that our infants should have been able to make the /va/ - /da/ distinction. Although this issue is not central to the intended purpose of Experiment 3, we decided to explore this finding in a follow-up experiment.

In order to investigate why our auditory /va/ and /da/ were not discriminable by the infants, an experiment was conducted to test how the stimuli were perceived by adults. As mentioned, the selected stimuli were derived from informal adult pilot studies conducted prior to the infant experiments. Experiment 4 was designed as a more formal test with adult subjects. Adult subjects were asked to identify (in a free response task) the auditory, visual, and audiovisual tokens used in the infant experiments. Beyond testing the perceptual clarity of the auditory /da/ and /va/ tokens, this control experiment provides a way to check if the audiovisual stimuli were perceived in the assumed fashion.

EXPERIMENT 4

Methods

Participants

Fifteen undergraduates (7 male) at the University of California, Riverside, participated for partial fulfillment of a class requirement. All reported normal or corrected vision, and were native speakers of English.

Stimuli

The stimuli consisted of audiovisual, audio-alone and video-alone tokens. The audiovisual tokens involved the (15) audio /va/-visual /va/, audio /ba/-visual /va/, and audio /da/-visual /va/ stimuli used in Experiment 1 (comprised of the five different frequency contoured auditory components for each audiovisual combination). In addition, the 15 auditory /ba/, /va/, and /da/ stimuli (five different frequency contours for each) were presented in isolation (with the screen dark).[4] Also, the video /va/ image was presented in isolation (with no sound present) along with video-alone /ba/ and /da/. (Although the infants were never shown video /ba/ and /da/ images, they were included in the adult test to make the video-alone identification task sensible.) For the identification tests, subjects were presented with 10 repetitions of each of the audio-alone tokens and five repetitions of each of the video-alone tokens. As for the audiovisual tokens, subjects were presented with 10 repetitions of the audio /da/-visual /va/ tokens, and 5 repetitions each of the audio /ba/-visual /va/ and audio /va/-visual /va/ tokens. Thus, subjects were presented with a total of 265 stimuli: 10 repetitions x 5 audio /da/-visual /va/ tokens; 5 x 5 audio /ba/-visual /va/ tokens; 5 x 5 audio /va/-visual /va/ tokens; 10 x 5 audio/va/; 10 x 5 audio/ba/; 10 x 5 audio/da/; 5 x 3 video-alone tokens (/va/, /ba/, /da/).

The presentation tape was set up in the following order: The first presentations were the 150 audio-alone tokens completely randomized together in six blocks of 25 presentations each. The second set of presentations were the 100 audiovisual tokens randomized together in four blocks. The final presentation set included the 15 visual-alone tokens randomized together in a single block. There was a 2.5 sec ISI between tokens and 15 sec between each block.

Procedure

Subjects were run in groups of two or three. They were seated at a table 5 feet in front of a video monitor. The audio stimuli were presented through a loudspeaker positioned directly beneath the monitor. All auditory stimuli were presented at comfortable listening levels which were 74 dB, 73 dB, and 73 dB SPL (A-weighted) for the /va/, /ba/, and /da/ tokens respectively. The lights were turned off in the presentation room. The only sources of illumination were the television monitor and a small light positioned near the table so that subjects could see their response sheets.

Subjects were told that they would be required to watch and listen to speech syllables. They were told that after each token was presented, they were to write down whatever syllable they heard and then look back up to the monitor for the next presentation. For the audiovisual blocks, subjects were told that it was important to watch each presentation; however they were to write down only what they heard (e.g., McGurk & MacDonald, 1976). For the audio-alone blocks, the video monitor was switched off and subjects were asked to write down what they heard. For the video alone blocks, the loudspeaker was shut off and subjects were asked to write down what they thought they might hear if the articulation they saw was producing some sound. The entire experimental session took less than one hour for each subject.

Results and Discussion

Mean percentage correct responses for the audiovisual, audio-alone, and video-alone tokens are listed in Table 2. (Since subjects were asked to base their judgments on what they heard, the audiovisual scores are listed as percentage correct based on the auditory component.) Means for the audiovisual and audio-alone tokens are pooled over their five frequency contour stimuli. The primary motivating issue for this experiment was the result of Experiment 3 that infants did not show dishabituation between the audio /va/ and /da/ stimuli suggesting that they these tokens were not discriminated. In surveying the means for the audio-alone presentations, it is clear that adults could easily identify the /da/ and /va/, and to a slightly lesser extent, the /ba/ token as well. Thus, the /da/ and /va/ tokens were identified as they had been designed to be, suggesting that the inability of infants to discriminate these tokens in Experiment 3 was not due to ambiguous stimuli.

Table 2. Pooled percentage of correct responses for each token type for Experiment 4.

Token		Percentage Correct (based on audio)
Audiovisual	/va-va/	98.40
	/ba-va/	1.60
	/da-va/	96.53
Audio-Alone	/va/	92.93
	/da/	99.87
	/ba/	83.47
Video-Alone	/va/	98.67
	/da/	97.33
	/ba/	100

Surveying the means for the visual alone and audiovisual stimuli, it is clear that these tokens also were effective in conveying the intended information. The visual /va/ component used in the infant experiments was identified correctly by adults 99% of the time. As for the audiovisual stimuli, identification accuracy for the audio /va/-visual /va/ token was also very high. The discrepant audiovisual stimuli were also judged in the expected manner. Thus, audio /ba/-visual /va/ was reportedly heard as something other than 'ba' over 98% of the time. In fact, 98.6% of the incorrect responses to audio /ba/-visual /va/ were 'va' which replicates results from prior experiments (e.g., Saldaña and Rosenblum, 1993; 1994; Rosenblum and Saldaña, 1992; and 1996). A t-test revealed that the percentage of correct (auditory-based) responses for the audio /ba/-visual /va/ token was significantly lower than the percentage of correct responses for the audio alone /ba/ token, $t(14)=29.61$, $p<.0001$, indicating a significant visual influence. In contrast, the audiovisual audio /da/-visual /va/ showed very little visual influence in eliciting 96.5% correct (audio-based) 'da' responses. These findings replicate previous results with similar stimuli (e.g., Manuel, Repp, Liberman, and Studdert Kennedy, 1983). Of the 3.5% incorrect responses, 3.2% were 'va'. A t-test revealed that the percentage of correct (auditory-based) responses for the audio /da/-visual /va/ token was not significantly different from the percentage of correct responses for the audio alone /da/ token, $t(14)=1.31$, $p=.21$, indicating no significant visual influence.

In summary, Experiment 4 revealed that the stimuli used in the infant experiments were identified by adults in the expected manner. With regard to the audiovisual stimuli, the audio /ba/-visual /va/ did show a substantial McGurk effect, while the audio /da/-visual /va/ did not. Thus, if the infant results of Experiment 1 are attributable to a visual influence, then it is a visual influence similar to that found with adults. The results of the auditory alone identification show that the adults had no difficulty in identifying the /va/ and /da/ stimuli. Thus, the fact that the infants in Experiment 3 did not show dishabituation between these tokens is not attributable to any obvious ambiguity in these stimuli. It is unclear why infants in Experiment 3 might have had trouble discriminating these tokens. Still, the intended purpose of Experiment 3 was served in showing that the findings of Experiment 1 were not based on the auditory similarity of /va/ and /ba/ relative to /va/ and /da/.

General Discussion

Taken together, these experiments suggest that infants generalize across two different audiovisual stimuli which are perceived as the same by adults, and that this generalization is based on a visual influence. Experiment 2 showed that the habituation differences of Experiment 1 were not based on some general preference for the audio /da/-visual /va/ token. Experiment 3 showed that the habituation differences were not based on the relative auditory similarity between the stimuli. Thus, we have observed evidence for a McGurk-type effect in 5-month-old infants.

In finding evidence that infants can integrate audiovisual speech, these results are relevant to issues of infant sensitivity to audiovisual speech correspondences. As mentioned, Kuhl and Meltzoff (1984) propose two ways in which audiovisual information can be matched to amodal

representations. While an identity match requires input which is phonetically identical across audio and visual modalities, a supramodal account requires only that the input mutually contact a common phonetic representation. As stated, the supramodal account allows for a percept to integrate the different, and potentially conflicting, information available in auditory and visual streams. As support for supramodal representations, these authors cite the evidence that adults can integrate discrepant audiovisual speech as in the McGurk effect. The current findings are the first to suggest that infants can also integrate audiovisual speech in a McGurk-type manner. In this sense, our results are supportive of the existence of supramodal representations in infants. Accordingly, these results are also evidence for a supramodal vs. identity match strategy for correspondence sensitivity. More generally, our integration results are supportive for Kuhl and Meltzoff's (1984) thesis that speech is represented to infants amodally.

In finding evidence for a McGurk effect in young, pre-linguistic infants, our results bear on the question of the developmental basis of the effect. Along these lines, theories of the McGurk effect can be broken into two classes: those that suggest the effect is a result of experience in associating audio and visual speech, and those that suggest that the effect is based on something other than experience (Fowler and Dekle, 1991). Regarding the former, the auditory enhancement theory of Diehl and Kluender (1989) proposes that the McGurk effect is based on a perceived association between audio and visual properties which is established through perceptual learning and experience with audiovisual speech.

In contrast, the motor theory of speech (Liberman and Mattingly, 1985) proffers that the McGurk effect occurs through the processes of an 'innately specified' speech module (Fodor, 1983; Liberman and Mattingly, 1985). Similarly, Meltzoff and Kuhl (1994) have suggested that the ability to implement supramodal representations is a perceptual function which is present at birth. They suggest that this ability could have a strong neural basis. An alternative extra-experiential account is offered by the ecological approach (e.g., Fowler and Rosenblum, 1991). From this perspective, the McGurk effect is based on the recovery of stimulus information which is (ostensibly) lawfully generated by, and therefore fully specificational to its source events (Fowler & Dekle, 1991; Gibson, 1979). This stimulus information in turn, structures receptor surfaces and the activities of perceptual systems (Fowler and Dekle, 1991; Shaw, Turvey, and Mace, 1982).

On its own, the current results can not provide direct support for either experiential or extra-experiential accounts of the McGurk effect. Although we did find evidence for the effect in young infants, it could be that our infants had sufficient experience with audiovisual speech to provide a basis for the observed effects. It is known that experience with auditory speech plays a significant role in developing perceptual sensitivities during the first year of life (e.g., Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Werker, 1989; Werker and Tees, 1984). There is no reason to believe that audiovisual speech would not play a similarly important role (e.g., Meltzoff & Kuhl, 1994). Potentially, evidence for an extra-experiential basis could be gained by testing the effect with younger infants who have had even more restricted audiovisual experience. However, testing younger infants is hampered by the fact that there are immaturaties in both the auditory and (particularly) visual systems (e.g., Aslin, 1987; Massaro, et al, 1986). Thus, even if younger infants failed to evidence a McGurk effect, ostensibly supporting an experiential account, such results would be suspect on the basis of a potential failure of either auditory or visual perception by itself, and not audiovisual integration.

Thus, an infant McGurk demonstration on its own cannot definitively support an extra-experiential basis to the effect. However, our evidence could be used to buttress other recent findings which are supportive of an extra-experiential account. For example, Fowler and Dekle (1991) have shown that heard speech can be influenced by a type of information which observers have not previously experienced. In their experiments, haptically conveyed syllables (felt with the hand using a Todoma technique) were found to affect perception of discrepant auditory syllables in a way similar to visual influences.[5] Since their subjects had no reported experience with haptic speech, Fowler and Dekle (1991) conclude that their findings are inconsistent with experience-based explanations of McGurk-type effects.

There is also evidence for the McGurk effect's linguistic universality. Initial reports on this topic indicated that the effect differed in strength across English, German (Mills and Theim, 1980, as discussed in Massaro, Cohen, Gesi, Heredia, & Tsuzaki, 1993), and Japanese (Sekiyama and Tohkura, 1991; 1993). However, more recent evidence suggests that these differences are based on a languages' specific phoneme inventories, phonetic realizations, and phonological constraints as they act to limit possible subject responses (Massaro, et al., 1993). Using a forced-choice response paradigm, Massaro, et al (1993) found no differences in visual influence strength among English, Spanish, and Japanese speaking subjects. Additionally, they found evidence for a similar bimodal processing strategy (described by the Fuzzy Logical Model of Perception) across the language groups. Thus, there is support that the general strategies underlying the McGurk effect are independent of specific language experience which could implicate a non-experiential basis.

Finally, indirect support for an extra-experiential account is provided by a series of findings on infant sensitivity to audiovisual speech correspondences. These findings show that sensitivity occurs: a) in 1-month old babies (Walton and Bower, 1993); b) for segments which are unfamiliar to infants (Walton and Bower, 1993); and c) with a left-hemisphere advantage implicating a specialized mechanism (MacKain, et al., 1983). Thus, there is evidence that very little—if any—experience is needed to detect correspondences and that specialized mechanisms might be involved. Although this evidence does not bear on the McGurk effect directly, parsimony would dictate that whatever basis underlies correspondence sensitivity is likely also be involved with audiovisual integration.

To summarize, while our infant results on their own do not provide unequivocal support for an extra-experiential basis, they do add to the evidence for linguistic universality (Massaro, et al, 1993), haptic influences (Fowler and Dekle, 1991), and early sensitivity to audiovisual speech correspondence (e.g., Walton and Bower, 1993) towards these ends. At the same time however, the influence of learning cannot be ignored. The developmental research suggests that for many stimuli, the strength of the McGurk effect increases through childhood (e.g., McGurk and MacDonald, 1976; Massaro, 1987; Boliek, et al., submitted). This might reflect an increased attunement to visual information (Massaro, et al, 1986). In fact, there is evidence that young children might not have the lipreading skills of adults (Massaro, et al, 1986). If so, then developmental experience could improve the perceiver's sensitivity to visual speech information so that it increases its influence on heard speech. Also, the cross-language research on the McGurk effect has revealed different response patterns in listeners who have experience with different native languages. This makes a great deal of sense if, as Meltzoff and Kuhl (1994) suggest, the representations which unite multi-modal information are heavily influenced by early linguistic environment. At the very least, native language experience bears on the response inventories subjects use in McGurk-type experiments (Massaro, et al, 1993). Future research will determine the relative degree that experience and extra-experiential factors play a role in the McGurk effect.

References

Aronson, E. & Rosenbloom, S. (1971). Space perception in early infancy: Perception within a common auditory-visual space. Science, 172, 1161-1163.

Aslin, R. N. (1987). Perceptual development. Annual Review of Psychology, 39, 435-473.

Best, C., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual organization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. Journal of Experimental Psychology: Human Perception and Performance, 14, 345-360.

Boliek, C., Green, K., Fohr, K, & Obrzut, J. (submitted). Auditory-visual perception of speech in children with learning disabilities: The McGurk effect. Paper submitted to the "International Neuropsychological Society" 24th Annual meeting, 1996.

Desjardins, R.N. & Werker, J.F. (1995, June-July). 4-month-old infants notice both auditory and visual components of speech. Poster presented at the annual meeting of the American Psychological Society, New York.

Demany, L., McKenzie, B., & Vurpillot, E. (1977). Rhythm perception in early infancy. In J. Oates & S. Sheldon (Eds.), Cognitive development in infancy (pp. 105-109). Hove, England: Lawrence Erlbaum Associates, Inc.

Dodd, B. (1979). Lipreading in infants: Attention to speech presented in and out of synchrony. Cognitive Psychology, 11, 478-484.

Diehl, R. L. & Kluender, K. R. (1989). On the objects of speech perception. Ecological Psychology, 1, 121-144.

Eilers, R. E., & Minifie, F. D. (1975). Fricative discrimination in early infancy. Journal of Speech and Hearing Research, 18, 158-167.

Eilers, R. E., Wilson, W. R., & Moore, J. M. (1977). Developmental changes in speech discrimination in three-, six-, and twelve-month-old infants. Journal of Speech and Hearing Research, 20, 766-780.

Eimas, P. D., & Tartter, V. C. (1979). On the development of speech perception: Mechanisms and analogies. In H. W. Reese & L. P. Lipsitt (Eds.), Advances in child development and behavior, vol. 13. NY: Academic Press.

Fodor, J. A. (1983). Modularity of mind. Cambridge, MA: Bradford Books.

Fowler, C. A. & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. Journal of Experimental Psychology: Human Perception and Performance, 17, 816-828.

Fowler, C. A. & Rosenblum, L. D. (1991). Perception of the phonetic gesture. In I. G. Mattingly and M. Studdert-Kennedy (Eds.), Modularity and the motor theory. Hillsdale, NJ: Lawrence Erlbaum and Assoc.

Gibson, J. J. (1979). The ecological approach to visual perception. Boston: Houghton-Mifflin.

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

Horowitz, F. D. (1975). Infant attention and discrimination: Methodological and substantive issues. Monographs of the society for research in child development, 5, 1-15.

Horowitz, F. D., Paden, L., Bhana, K., & Self, P. (1972). An infant-control procedure for studying infant visual fixations. Developmental Psychology, 7, 90.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America, 67, 971-994.

Kuhl, P. K. (1979). The perception of speech in early infancy. In N. J. Lass (Ed.), Speech and language: Advances in basic research and practice. NY: Academic Press.

Kuhl, P. K. (1987). Perception of speech and sound in early infancy. In P. Salapatek and L. Cohen (Eds.), Handbook of infant perception, Vol. 2: From perception to cognition. (p. 275-382). New York: Academic Press.

Kuhl, P. K. & Meltzoff, A. N. (1982). The bimodal development of speech in infancy. Science, 218, 1138-1141.

Kuhl, P. K. & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. Infant Behavior and Development, 7, 361-381.

Kuhl, P. K. & Meltzoff, A. N. (1988). Speech as an intermodal object of perception. In A. Yonas (Ed.), Perceptual development in infancy: The Minnesota symposia on child psychology, Vol. 20. Hillsdale, NJ: Erlbaum.

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

Kuhl, P. K., Williams, K. A., & Meltzoff, A. N. (1991). Cross-modal speech perception in adults and infants using nonspeech auditory stimuli. Journal of Experimental Psychology: Human Perception and Performance, 17, 829-840.

Legerstee, M. (1990). Infants use multimodal information to imitate speech sounds. Infant Behavior and Development, 13, 343-354.

Levitt, A., Jusczyk, P. W., Murray, J., & Carden, G. (1988). Context-effects in two-month-old infants' perception of labiodental/interdental fricative contrasts. Journal of Experimental Psychology: Human Perception and Performance, 14, 361-368.

Liberman, A. M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. Cognition, 21, 1-36.

MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left hemisphere function. Science, 219, 1347-1349.

Manuel, S. Y., Repp, B. H., Liberman, A. M., & Studdert-Kennedy, M. (1983). Exploring the "McGurk Effect." Paper presented at the 24th Annual Meeting of the Psychonomic Society, San Diego, CA.

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 Assoc, Inc.

Massaro, D. W., & Cohen, M. M. (1990). Perception of synthesized audible and visible speech. Psychological Science, 1, 55-63.

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

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.

McGrath, M. & Summerfield, A.Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. Journal of the Acoustical Society of America, 77, 678-685.

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

Meltzoff, A. N., & Kuhl, P. K. (1994). Faces and speech: Intermodal processing of biologically relevant signals in infants and adults. In D. J. Lewkowitz & R. Lickliter (Eds.), The development of intersensory perception: Comparative perspectives. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Mills, A. E. (1987). The development of phonology in the blind child. In B. Dodd & R. Campbell (Eds.), Hearing by eye: The psychology of lip-reading. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Mills, A. E. & Theim, R. (1980). Auditory visual fusions and illusions in speech perception. Linguistische Berichte, 6. 85-106.

Pickens, J., Field, T., Nawrocki, T., Martinez, A., Soutollo, & Gonzalez (1994). Full-term and preterm infants' perception of face-voice synchrony. Infant Behavior and Development, 17, 447-455.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lipreading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), Hearing by eye: The psychology of lip-reading. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (under review). Point-light facial displays enhance comprehension of speech in noise. Journal of Speech and Hearing Research.

Rosenblum, L. D. & Saldaña, H. M. (1992). Discrimination tests of visually influenced syllables. Perception and Psychophysics, 52, 461-473.

Rosenblum, L. D. & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. Journal of Experimental Psychology: Human Perception and Performance, 22(2), 318-331.

Saldaña, H. M. & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. Perception and Psychophysics, 54, 406-416.

Saldaña, H. M. & Rosenblum, L. D. (1994). Selective adaptation in speech perception using a compelling audiovisual adaptor. Journal of the Acoustical Society of America, 95, 3658-3661.

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

Shaw, R., Turvey, M. T., & Mace, W. (1982). Ecological psychology: The consequence of a commitment to realism. In W. Weimer & D. Palmero (Eds.), Cognition and the symbolic processes, 2 (pp. 3-51). London: Erlbaum.

Snyder, R. T., & Pope, P. (1970). New norms for and an item analysis of the Wepman Test at the first grade, six-year-level. Perceptual and Motor Skills, 31, 1007-1010.

Spelke, E. S., & Cortelyou, A. (1980). Perceptual aspects of social knowing: Looking and listening in Infancy. In M. E. Lamb & L. R. Sherrod (Eds.), Infant social cognition. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Spelke, E. S. & Owsley, C. (1979). Intermodal exploration and knowledge in infancy. Infant Behavior and Development, 2, 13-27.

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

Summerfield, Q., MacLeod, P., McGrath, M., & Brooke, N. M. (1989). Lips, teeth, and the benefits of lipreading. In A. W. Young & H. D. Ellis (Eds.), Handbook of research on face processing (pp. 223-233). New York: Elsevier Science Pub. Co.

Walker, A. S. (1982). Intermodal perception of expression behaviors by human infants. Journal of Experimental Child Psychology, 33, 514-535.

Walton, G. E. & Bower, T. G. R. (1993). Amodal representation of speech in infants. Infant Behavior and Development, 16, 233-243.

Werker, J. (1989). Becoming a native listener. American Scientist, 77, 54-59.

Werker, J., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behavior and Development, 7, 49-63.

Author Notes

We gratefully acknowledge the assistance of Chad Audet, Chantel Bosely, David Kim, and Sunny Moore as well as the helpful comments of Richard Aslin, Kerry Green, Dominic Massaro, Joanne Miller, an anonymous reviewer and the UCR Cognitive Science group on an earlier version of the manuscript.

This research was supported by NSF Grant DBS-9212225 awarded to the first author and a grant from the Natural Sciences and Engineering Research Council of Canada to the second author.

Requests for reprints should be sent to Lawrence D. Rosenblum, Department of Psychology, University of California, Riverside, Riverside, California, 92521, rosenblu@citrus.ucr.edu

Footnotes

[1]Preliminary—and inconclusive—results of an infant McGurk study were recently reported as a conference poster by Desjardins and Werker (1995).

[2]If relative audiovisual compatibility were to account for the Experiment 1 results, it would more likely be because infants found the audio /da/-visual /va/ more compatible than the audio /ba/-visual /va/. Infants looked longer at the audio /da/-visual /va/ than audio /ba/-visual /va/ token and as reviewed above, the gaze preference literature has overwhelmingly shown that infants are more attentive to audiovisual compatible tokens. This has been demonstrated with many types of speech stimuli (see Kuhl and Meltzoff, 1984, for a review) and in different methodological contexts (e.g., Legerstee, 1990; Walton and Bower, 1993).

[3]It was somewhat of a concern to us that four infants in this study were raised in a 50% English speaking home. For this reason, an additional analysis was conducted on the data from the 16 subjects who were raised in primarily English-speaking homes. The results of this analysis were essentially the same as the analysis including all 20 subjects. The mean looking times for the

audio /ba/-visual /va/ and audio /ba/-visual /da/ were 18.50 (SD=21.30) and 21.60 (SD=28.00) respectively. The three-way ANOVA (trial type x repetition x order) revealed a significant effect only for trial repetition, $F(2,28) = 7.53$, $p = .002$. Again, there was no effect of stimulus type, $F(1,14) = 0.68$, $p = 0.43$, and no interaction between the two variables, $F(2,28) = 0.901$, $p = 0.418$ (see Experiment 2 results section).

[4]Based on convenience, the audio-alone stimuli were presented without a neutral face image for the adult control experiment. While it is acknowledged that this presentation procedure is different from that used in the infant auditory control experiment (Experiment 3), there is much previous research showing that a neutral face image does not influence adult identification of similar syllables (e.g., Massaro, 1987).

[5]However, Massaro (personal communication) has suggested that the integration of audio-haptic speech is of a different nature than the integration processes used for audiovisual speech. More specifically, Massaro claims that while the integration of audiovisual speech is captured by the Fuzzy Logical Model of Perception, the integration of audio-haptic speech is better described by an additive or averaging model.