Perception & Psychophysics
2006, 68 (4), 582-592

A sex difference in visual influence on heard speech

JULIA R. IRWIN and D. H. WHALEN
Haskins Laboratories, New Haven, Connecticut

and

CAROL A. FOWLER
Haskins Laboratories, New Haven, Connecticut,
University of Connecticut, Storrs, Connecticut,
and Yale University, New Haven, Connecticut

Reports of sex differences in language processing are inconsistent and are thought to vary by task type and difficulty. In two experiments, we investigated a sex difference in visual influence on heard speech (the McGurk effect). First, incongruent consonant–vowel stimuli were presented where the visual portion of the signal was brief (100 msec) or full (temporally equivalent to the auditory). Second, to determine whether men and women differed in their ability to extract visual speech information from these brief stimuli, the same stimuli were presented to new participants with an additional visual-only (lipread) condition. In both experiments, women showed a significantly greater visual influence on heard speech than did men for the brief visual stimuli. No sex differences for the full stimuli or in the ability to lipread were found. These findings indicate that the more challenging brief visual stimuli elicit sex differences in the processing of audiovisual speech.

Visible speech influences what listeners hear. Visual information assists in the recognition of speech in noise (Sumby & Pollack, 1954) and is also used in the perception of unambiguous, acoustically specified speech (Desjardins, Rogers, & Werker, 1997; McGurk & MacDonald, 1976). McGurk and MacDonald provided a compelling demonstration of this by presenting incongruent audio consonant–vowel (CV) and video CV syllables to perceivers. Perceivers watching stimuli manipulated in this manner sometimes reported hearing consonants that combined the places of articulation of the visual and auditory tokens (e.g., visual /ba/ auditory /ga/ heard as /bga/), or they heard fusions (e.g., visual /ga/ auditory /ba/ heard as /da/), or the visual place information dominated (e.g., visual /va/ auditory /ba/ heard as /va/). This finding, called the *McGurk effect*, has become a focus of study in speech perception (e.g., Campbell, 1994; Desjardins et al., 1997; Fowler, Brown, & Mann, 2000; Fowler & Dekle, 1991; Green, 1998; Green, Kuhl, Meltzoff, & Stevens, 1991; Green & Norrix, 1997; MacDonald, Andersen, & Bachmann, 2000; Massaro, 1987; Paré, Richler, ten Hove, & Munhall, 2003; Rosenblum & Saldaña, 1996, 1998; Rosenblum, Schmuckler, & Johnson, 1997; van Wassenhove, Grant, & Poeppel, 2005).

Perceivers who experience the McGurk effect describe it as compelling, occurring even when they are aware of how the stimuli have been manipulated (Massaro, 1987). However, the effect does not always occur, even for those tokens that have previously shown a visual influence (Brancazio & Miller, 2005). This variability in visual influence on what is heard has been relatively unexplored in studies of audiovisual speech perception. One potential source of variability that has been reported in the literature is sex differences in visual influence on heard speech. A few studies have indicated that women are more influenced by visual information than are men when presented with incongruent auditory and visual McGurk stimuli (Aloufy, Lapidot, & Myslobodsky, 1996, in the context of American English; Öhrström & Traunmüller, 2004, in the context of Swedish). However, sex differences do not always obtain (for Hebrew, Aloufy et al., 1996). Desjardins and Werker (2004) also recently reported sex differences in the integration of seen and heard speech in infants. However, no clear pattern was observed, with either male or female infants demonstrating greater rates of integration, depending on the type of stimuli presented.

Additional indirect evidence that there may be differences between men and women in the processing of audiovisual speech has come from neuroimaging studies, where there has been overlap in the areas of the brain in which audiovisual speech has been processed and where sex differences in language processing have been found. Functional magnetic resonance imaging (fMRI) technology indicates that audiovisual speech is processed in the superior temporal sulcus (STS; e.g., Calvert, 2001; Calvert et al., 1999; Calvert & Campbell, 2003; Calvert, Campbell, & Brammer, 2000; Olson, Gatenby, & Gore, 2002; Sekiyama, Kanno, Miura, & Sugita, 2003) and the inferior

Correspondence concerning this article should be addressed to J. R. Irwin, Haskins Laboratories, 300 George Street, Suite 900, New Haven, CT 06511 (e-mail: julia.irwin@haskins.yale.edu).

frontal gyrus (IFG; Calvert & Campbell, 2003; Jones & Callan, 2003). A number of fMRI studies have shown sex differences in the pattern of activation for auditory tasks that require phonological processing in the IFG, with males showing significantly greater left- than right-side activation in this area, in comparison with females, who are more bilateral (Pugh, Shaywitz, Shaywitz, Constable, et al., 1996; Pugh, Shaywitz, Shaywitz, Fulbright, et al., 1996; Shaywitz et al., 1995). Females also show bilateral activation in the IFG, the superior temporal gyrus (STG), and cingulate regions for semantic tasks (Baxter et al., 2003) and for lexical visual field tasks (Rossell, Bullmore, Williams, & David, 2002). Similar findings have been reported using positron emission tomography. Jaeger et al. (1998) demonstrated greater bilateral activation in the perisylvian cortex (which includes the STG, the IFG, and the adjacent premotor cortex; Riecker, Wildgruber, Dogil, Grodd, & Ackermann, 2002) for females than for males in the production of past tense verb forms. Differential patterns of processing are thought to underlie these reported sex differences (Majeres, 1999), with bilaterality affording faster and/or more efficient phonological processing for women (Coney, 2002).

The evidence that there are sex differences in language processing is not unequivocal, however. A number of studies have failed to show sex differences in language processing for word generation tasks (Knecht et al., 2000), language comprehension tasks (Frost et al., 1999), or verbal reasoning (Gur et al., 2000). To date, no studies have shown males to be more bilateral, indicating that issues of experimental design and power may explain the lack of an effect in some reports.

To further complicate our understanding of sex differences in language processing, findings may vary as a function of task type (Josse & Tzourio-Mazoyer, 2004; Kansaku & Kitazawa, 2001) or task difficulty (Jaeger et al., 1998). Jaeger et al. hypothesized that task difficulty is a critical factor in finding sex differences in language processing, with differences emerging as task demands increase.

To this point, the available evidence from neuroimaging and behavioral studies is contradictory as to whether there are sex differences in the processing of language. Previous research has indicated that there may be sex differences in the processing of audiovisual (AV) speech. Thus, to more closely examine sex differences in the perception of AV speech, we compared males' and females' responses to incongruent AV stimuli that were varied for perceptual difficulty.

To vary perceptual difficulty, we manipulated the amount of visual information available to the perceiver. Incongruent AV CV syllables were presented to perceivers either with the audio and the video temporally equivalent or with a brief segment of the speaker's face paired with the audio. The temporally equivalent condition provides a rich visual signal, with articulatory information preceding and following the consonantal burst. In contrast, in the brief condition, the visual signal is significantly attenuated, providing only visual information about the consonantal closure. Thus, incongruent AV stimuli with brief visual signals should be more challenging perceptually for perceivers.

Furthermore, the brief visual signals were manipulated so as to vary motion in the speaker's face. The brief visual stimuli were either static, with a repeated single image of a speaker at the clearest indication of consonant closure, or dynamic, where the moving visible gestures produced by the speaker for the consonantal closure and release were presented. Rosenblum and Saldaña (1996) compared static and dynamic visual signals in the visual influence of seen on heard speech, reporting that static visual signals are impoverished in comparison with dynamic visual signals, which provide kinematic information unavailable from static stimuli. If this is the case, there should be more of a visual influence from dynamic than from static visual signals.

Finally, the visual syllables produced by the speaker were varied for ease of discriminability. The first visual syllable, /va/, is a voiced labiodental fricative produced in the front of the mouth, making it particularly easy to detect visually. This viseme leads to high levels of visual influence when paired with incongruent auditory signals (e.g., visual /va/ paired with auditory /ba/ results in a visually influenced percept of /va/ more than 96% of the time; Rosenblum & Saldaña, 1996; Rosenblum et al., 1997). In contrast, the other visemes that were paired with incongruent auditory stimuli, visual /da/ (a voiced alveolar) and /ða/ (a voiced dental fricative; the consonant is the initial consonant of "the"), are produced farther back in the speaker's mouth than is /va/, making them less visually accessible. Accordingly, visual /da/ and /ða/ lead to levels of visual influence in the context of /ba/ that are lower than those reported for /va/ (Brancazio, 2004). The syllables that contain the more visually discriminable consonants (i.e., /va/) should be easier to detect and should lead to a greater visual influence on what is heard than the syllables that contain the less visually discriminable consonants (/da/ and /ða/).

The present comparison of the influence of visual information on heard speech in men and women can lead to a number of possible outcomes. The first possibility is that there are no sex differences in AV speech perception, regardless of task difficulty. Alternatively, if there is a sex difference, women are likely to show more visual influence than are men for the incongruent AV stimuli, as has been reported in prior research (Aloufy et al., 1996; Öhrström & Traunmüller, 2004). However, if task difficulty is an important variable in eliciting sex differences, men and women may show behavioral differences only for the brief, more challenging task conditions. The present findings will provide new evidence about the perceptual processing of men and women in the domain of AV speech.

EXPERIMENT 1

This experiment was designed to examine the influence of visual information on heard speech in male and female perceivers. Participants were presented with incongruent

brief visual and temporally equivalent visual and auditory stimuli. In addition, motion in the brief visual stimuli was manipulated.

Method

Participants

The participants were 30 male and 36 female native English speakers. Nine male and 12 female participants, ranging in age from 25 to 49 years, were colleagues at Haskins Laboratories and were naive as to the purposes of the experiment. The remaining participants were undergraduate students at the University of Connecticut, ranging in age from 18 to 22 years, who received course credit for participating. All were right-handed, with normal or corrected vision, and reported no hearing or speech difficulties.

Stimuli

The speech stimuli were recorded on videotape in an Industrial Acoustics Company booth to reduce ambient noise. A female native speaker of English repeated the syllables /ba/, /da/, /ða/, and /va/ several times in random order. From these, a single token of each videotaped syllable was selected. To create the test stimuli, the auditory and visual syllables were derived from the videotaped recording.

The videotaped syllables were digitized into a Macintosh computer. Video editing was done using the Adobe Premiere software program (Adobe Systems, 1995). Two versions of the AV stimuli were created: The audio and video were either congruent or incongruent. The incongruent trials had the video syllable /da/, /ða/, or /va/ with the audio /ba/. (There could be no AV mismatch for the /ba/ token, so no incongruent trials of this type were presented.) The congruent trials had the original, unedited video and audio signals.

Brief audiovisual stimuli. The visual portion of the congruent and incongruent AV stimuli was manipulated to be either full or brief (for an example of the full visual stimuli, see Figure 1A). The brief video segments were digitally edited to present an attenuated visual signal (approximately 100 msec, three video frames). These tokens were further manipulated to vary motion in the speaker's face. To vary motion, three versions of the brief visual stimuli were created: three-frame static, three-frame dynamic, and two-frame dynamic (see Figures 1B, 1C, and 1D, respectively). The static stimuli were created by repeating the visual frame representing the clearest indication of the place of articulation for three frames (100 msec). The remainder of the video image was deleted and replaced with a solid black background. Three-frame dynamic stimuli consisted of three continuous frames showing consonant closure and release. Deleting the middle frame from the three-frame dynamic stimuli created a discontinuous visual image for the two-frame dynamic stimuli. In general, we anticipated that the dynamic stimuli would lead to more visual influence than would the static stimuli. Within the dynamic types, the two-frame version was expected to lead to less influence than would the three-frame version.

Examination of the video segments indicated that for /ða/ and /da/, the frame that contained the clearest indication of place of articulation was one frame (33 msec) prior to the release of the consonant constriction, whereas for /ba/ and /va/, it was two frames prior to the acoustic indications of release. To avoid conflict between the openness of the vocal tract implied by the acoustics and the closure visible in the video, the static and dynamic conditions were designed so as to have slightly different timing. In particular, this was done to avoid the problem of an image of a closed vocal tract co-occurring with a vowel sound (as in Rosenblum & Saldaña, 1996). To do this, the final frame of the three-frame static segment was temporally aligned at its original location with the consonantal release (see Fig-

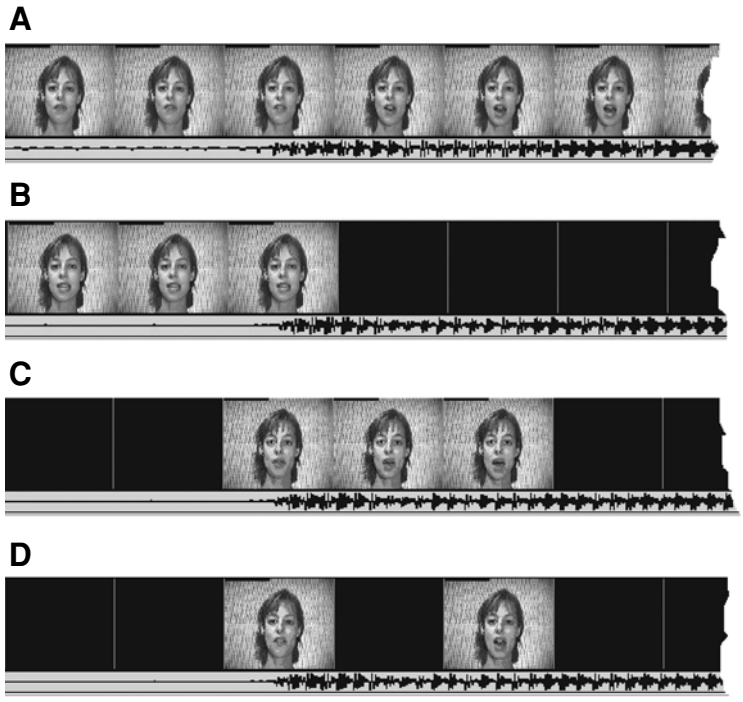

Figure 1. Composition of the stimuli. (A) The first 6 of the 20 frames for the full video condition for /va/. (B) Static (one frame repeated an additional two times) configuration for /da/. (C) Three-frame dynamic configuration for /ða/. (D) Two-frame dynamic configuration for /va/. (Note that the two frames used are in the same temporal relation to the audio signal.)

ure 1B). Therefore, the static stimuli always occurred two frames (66 msec) earlier with reference to the onset of the vocalic segment than did the dynamic stimuli, to preserve the compatibility of the visual and the acoustic signals as described above. The audio signal was always aligned so that the onset of the vocalic segment was in its original relationship to the video frames.

After editing was complete, the brief audiovisual stimuli were compiled in random order and recorded to videotape so that each trial could be individually checked for accuracy of reproduction. The timing demands occasionally exceeded system capability, so that three brief trials did not contain clear visual face information for the specified number of frames (e.g., only half of the face was shown in a frame). To find these trials, the stimulus tape was reviewed frame by frame in its entirety. Responses to the three trials were not scored; they were replaced with responses made to the same stimuli during the warm-up trials that preceded each stimulus block.

Full visual stimuli. In the full video stimuli condition, the visual syllable lasted as long as the acoustic syllable (667 msec, or 20 video frames; see Figure 1A). The full visual stimuli consisted of the congruent (/ba/, /da/, /ða/, and /va/) and incongruent (/da/, /ða/, and /va/ with audio /ba/) CV tokens. For the incongruent trials, the visual and auditory frames that specified consonant closure were aligned. As with the brief visual stimuli, the full audiovisual tokens were compiled in random order and recorded to videotape for presentation.

For both the brief and the full visual stimuli, the congruent trials allowed an assessment of overall attention to the task and ability to identify the syllables. The degree to which the visible signal influenced the heard response was determined from the responses to the incongruent stimuli. The audio signal was always /ba/ for the incongruent condition; therefore, a response corresponding to the visual token (e.g., /va/) indicated visual influence. Specifically, "v" in response to visual /va/ stimuli and "d" or "th" in response to /da/ or /ða/ visual stimuli were considered evidence of visual influence.

Procedure

Each trial consisted of a warning tone accompanying a white screen (for 1 sec), followed by the printed word "Ready" (for 500 msec) and the AV token. After the AV token, a black screen was presented for 3 sec to allow the participants to respond. Each participant was tested separately, seated in front of a television screen at a distance of approximately 1 m. Observation of numerous trials per participant indicated that the warning tone served its purpose of reorienting visual attention to the screen at the appropriate time. The acoustic signal was delivered over the monitor's speaker. The participants indicated their response by circling one of four printed choices ("b v th d") on an answer sheet. The participants were asked to pay attention to the video image but to record only what they heard. They were to guess if necessary and to look up at the monitor once they heard the warning tone, to ensure that their visual attention was on the face as the syllable was presented.

Brief visual stimuli. The brief visual stimuli were presented first. Five blocks were presented to the perceivers. Four warm-up tokens, one of each video consonant, began each block. A total of five tokens of each type of brief visual stimuli were presented in random order, with one repetition of each of the congruent and incongruent brief versions per block. These stimuli varied by movement (static and dynamic, two and three frames) and viseme (both incongruent visual /va/, /da/, and /ða/ with audio /ba/ and congruent /va/, /da/, and /ða/), for a total of 90 trials. To avoid a majority of /ba/ responses for the participants who did not show a visual influence on heard speech in the more difficult, brief visual condition, congruent /ba/ trials were not presented. The same order of blocks was used with all the participants.

Full visual stimuli. The participants were presented with the full visual stimuli after the brief versions. Ten repetitions of each of the incongruent (visual /va/, /da/, and /ða/ with audio /ba/) and congruent (/va/, /ba/, /da/, and /ða/) stimuli were presented in random order, for a total of 70 trials. Again, the order of blocks was the same for all the participants.

Results and Discussion

Congruent Audiovisual Stimuli

Responses to the full and brief congruent stimuli were accurate (averaging 98% in both tests), indicating that the male and female participants were able to make the phonetic judgments easily in both conditions.

Incongruent Audiovisual Stimuli

For the incongruent stimuli, visually influenced responses were defined as the percentage of responses in which the phonetic category matched the visual information. Because /d/ and /ð/ have been found to be highly confusable for English listeners in the context of /a/ as McGurk stimuli (Green et al., 1991), these two categories (as both stimulus and response) were collapsed. For visual /va/, only "v" responses counted as visually influenced. Since our primary interest was in responses to the more challenging brief visual stimuli, planned comparisons (ANOVAs) were undertaken for these stimuli. Responses to the full visual stimuli were analyzed separately.

Brief Visual Stimuli

An ANOVA with the factors of sex, phone type (/d/–/ð/ or /v/), and motion (static, dynamic two-frame, or dynamic three-frame) was performed on responses to the brief video stimuli. The sex factor was statistically marginal [$F(1,64) = 3.11$, $p < .10$], with females reporting visually influenced percepts in 8.8% more instances than did males for the brief visual stimuli. The main effect of phone type was significant [$F(1,64) = 107.91$, $p < .001$], with /d/–/ð/ eliciting a smaller rate of visually influenced percepts (38.3%) than did /v/ (74.5%). There was also a main effect of motion [$F(2,128) = 13.57$, $p < .001$]. Visually influenced responses for the static images were 7.9 percentage points higher than the average for the two dynamic conditions, and a Scheffé post hoc test revealed significant differences between the static and the dynamic two- and three-frame stimuli [$F(1,64) = 14.62$ and 21.66, respectively; $p < .01$]. There were no differences between the two dynamic conditions.

The interaction of sex and phone type was significant [$F(1,64) = 4.46$, $p < .05$]. Separate main effects analyses for the /d/–/ð/ stimuli and the /v/ stimuli showed a significant effect of sex [$F(1,64) = 6.13$, $p < .05$]. Female perceivers showed more visually influenced responses than males in the context of the /d/–/ð/ visemes, but not in the context of the /v/ viseme ($F < 1$). None of the interactions of sex or phone type with motion was significant.

Full Video Stimuli

A separate ANOVA with two factors (sex and phone type) was performed on the full video results. There was a significant main effect of phone type [$F(1,64) = 35.22$, $p < .001$], with less visual influence for the /d/–/ð/ than for the /v/ stimuli. No significant main effects or interactions were found with sex.

A comparison of the responses by males and females to the incongruent brief and full AV stimuli can be seen in Figure 2. Furthermore, histograms showing group differences in the "v" and "d"–"th" responses for males and females can be seen in Figures 3 and 4. Although the pattern of responses is similar for the "v" responses, which did not differ significantly, the histogram depicting the distribution of the "d"–"th" responses differs for the male and the female perceivers.

The major finding in Experiment 1 was that women reported significantly more visually influenced percepts than did men in the most challenging perceptual condition, with brief visual stimuli in the context of the /da/–/ða/ viseme. In comparison, there were no sex differences for the full McGurk AV stimuli. This appears to provide support for Jaeger et al.'s (1998) hypothesis that women and men differ in AV integration of speech in the context of more difficult task conditions. Previous studies have indicated that females show a greater degree of bilaterality in processing certain types of linguistic stimuli (Baxter et al., 2003; Coney, 2002; Jaeger et al., 1998; Pugh, Shaywitz, Shaywitz, Constable, et al., 1996; Pugh, Shaywitz, Shaywitz, Fulbright, et al., 1996; Rossell et al., 2002; Shaywitz et al., 1995). This bilaterality in women may lead to greater processing speed or efficiency for AV speech, yielding more visual influence for the brief visual stimuli than in men. Alternatively, it is possible that the sex difference is due to a differential ability to *perceive* the brief visual speech stimuli. That is, perhaps women extract phonetic information from the brief displays more successfully than do men. In order to test for this possibility, in the next experiment, males' and females' identifications of brief visual-only (lipread) stimuli were compared.

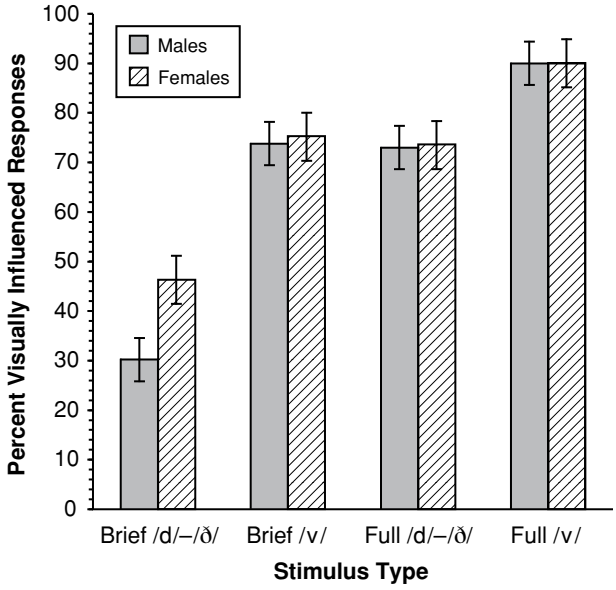

Figure 2. Percentages of visually influenced responses for males and females for incongruent brief and full /d/–/ð/ and /v/ visual stimuli in Experiment 1.

With regard to motion in the speaker's face, the results of Experiment 1 indicate that presentation of brief static images yields significantly more visually influenced percepts than do brief dynamic signals. This appears to contradict Rosenblum and Saldaña's (1996) claim that dynamic signals provide richer information for the perceiver. However, this difference may be a function of the timing differences between the static and the dynamic brief stimuli. In order to avoid showing a visual image of the speaker with a closed mouth paired with acoustics that specify an open vocal tract, the brief visual stimuli in the static condition were presented two frames (66 msec) earlier with reference to the consonantal burst. Information for place of articulation is present in the visual signal prior to the auditory signal in natural speech, and perceivers appear to be sensitive to this information (Munhall & Tokhura, 1998). For example, perceivers are better at detecting AV asynchrony when the auditory leads the visual signal (Conrey, 2004; Conrey & Pisoni, 2004). The visual lead for the static stimuli may have provided an advantage in this condition, yielding an increase in visual influence.

Thus, a second experiment was designed for two reasons: to see whether men and women differ in their ability to perceive brief visual information from a speaker's face and to address the possibility that greater visual influence for the static stimuli was a result of earlier presentation of the visual signal in this condition.

EXPERIMENT 2

This experiment was designed to examine whether the sex differences in Experiment 1 reflected differences in visual influence on heard speech or in the ability to detect brief visual signals. To assess whether males and females differ in their ability to identify consonantal information from brief visual stimuli, a new group of participants were presented with both the brief and the full visual stimuli from Experiment 1 and with visual-only (lipread) versions of these stimuli.

In addition, to examine whether the greater visual influence for the brief static stimuli in Experiment 1 was due to timing differences between the static and the dynamic conditions, in this experiment the visual signals were edited to have the same timing, relative to the audio signal.

Method

Participants

The participants were 27 male and 30 female native English-speaking undergraduates at the University of Connecticut (18 to 22 years of age) who received course credit for participating. All were right-handed, with normal or corrected vision, and reported no hearing or speech difficulties. None had been a participant in Experiment 1.

Stimuli

Brief and full visual stimuli. The brief and full visual stimuli were modified versions of the brief stimuli in Experiment 1. In Experiment 1, the static and dynamic conditions were designed to have slightly different timing in order to avoid an image of a closed vocal tract co-occurring with a vowel sound. The static stimuli always oc-

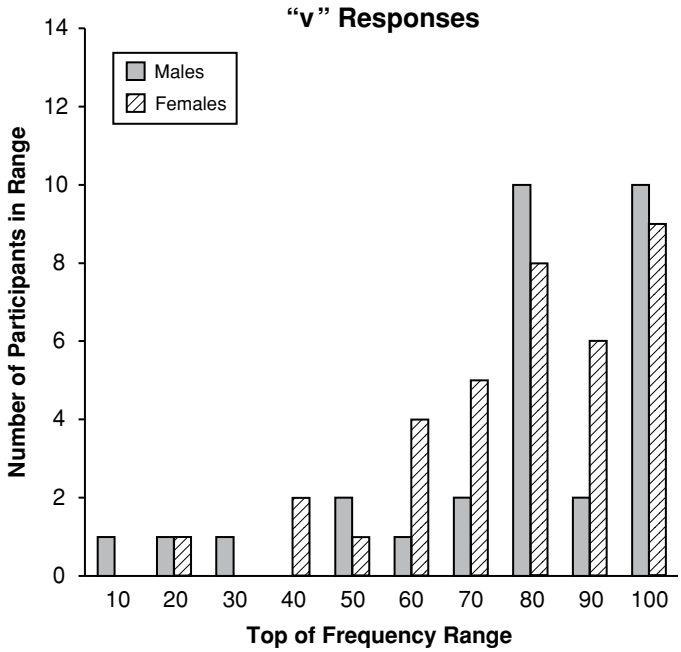

Figure 3. Number of "v" responses to incongruent audiovisual stimuli by sex in Experiment 1.

curred earlier (with reference to the onset of the vocalic segment) than the dynamic stimuli in order to preserve the compatibility of the visual and acoustic signals. In this experiment, the timing relation between the video and the audio was equated in static and dynamic conditions. To do this, the static images were shifted one frame later (relative to the audio) and the dynamic video one frame earlier than their locations in Experiment 1. The middle frame of the video display then aligned with the release of the consonant in the audio signal, whether the visual display was static or dynamic. The consonantal place of articulation in the two signals was congruent or incongruent, as in Experiment 1. As in Experiment 1, the three versions of brief visual stimuli (three-frame static, two-frame dynamic, and three-frame dynamic) were presented to the perceivers. To create an analogous two-frame static stimulus, the middle frame

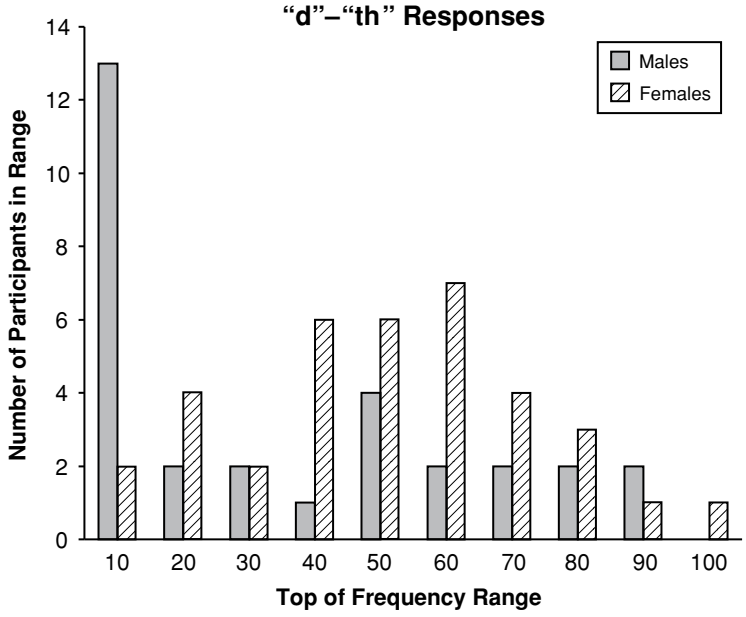

Figure 4. Number of "d"–"th" responses to incongruent audiovisual stimuli by sex in Experiment 1.

of the three-frame static stimulus was replaced with a solid black video frame. Thus, there were 5 trials of each type of movement (static two- and three-frame and dynamic two- and three-frame) by viseme (both incongruent visual /va/, /da/, and /ða/ with audio /ba/ and congruent /va/, /da/, and /ða/), for a total of 120 brief trials. As in Experiment 1, there were 70 full trials. Within a block, the brief and full stimuli were presented in random order. Again, the order of blocks was the same for all the participants.

Visual-only stimuli. The visual-only stimuli were silent versions of the brief stimuli described above.

Procedure

The same procedure as that in Experiment 1 was used. Five repetitions of each of the brief video stimuli were recorded with the same presentation structure as that in the previous experiment. The AV condition was presented first, with the acoustic signal played at a comfortable listening level. The visual-only condition was presented next, with the same stimuli as those in the AV condition, but with the volume on the television monitor turned to its minimum. Finally, the full video condition was presented with sound to the participants.

In the AV and full video conditions, the participants were asked to pay attention to the video image but to record only what they heard. In the visual-only condition, they were to report which of the four consonants they thought the speaker had said. In all the conditions, they were to guess, if necessary, and to look up at the monitor once they heard the warning tone.

Results and Discussion

Congruent Audiovisual Stimuli

In the AV condition, responses to the congruent brief and full video stimuli were accurate (averaging 87.7% for both), indicating that the participants were able to make these phonetic judgments easily, as was the case in Experiment 1. However, the participants in this experiment were approximately 10% less accurate than those in Experiment 1. This reduction in accuracy may have been due to the one-frame shift in the video signal, relative to the audio signal. The shift could have reduced accuracy by lowering the overall plausibility of the stimuli, since the visual signal now sometimes conflicted with the audio (e.g., when a closed mouth occurred with a vowel sound).

As in Experiment 1, for the incongruent AV stimuli, visually influenced responses were defined as the percentage of responses in which the phonetic category matched the visual information.

Incongruent Audiovisual Stimuli

As in Experiment 1, visually influenced responses were defined as the percentage of responses in which the phonetic category matched the visual information. Again, the /d/ and /ð/ categories (both as stimulus and as response) were collapsed. For visual /va/, only "v" responses counted as visually influenced. Again, our primary interest was in responses to the more challenging brief visual stimuli. Therefore, planned comparisons were conducted, with the brief, full, and visual-only conditions analyzed separately.

Brief Visual Stimuli

An ANOVA was performed with the between-subjects factor of sex and the within-subjects factors of motion (dynamic vs. static), number of frames (two vs. three), and phone type (/d/–/ð/ vs. /v/). There was a significant main effect of sex, with females reporting visually influenced percepts in 13.2% more instances than did males for the brief visual stimuli [$F(1,55) = 6.70$, $p < .02$]. The main effect of phone type was also significant [$F(1,55) = 47.63$, $p < .001$], with /v/ eliciting more visual influence (62.5%) than did /d/–/ð/ (34.1%). There was no significant interaction of sex and phone type. Furthermore, there was no main effect or significant interaction with the motion or number of frames variables.

Full Visual Stimuli

A separate analysis was performed for the full visual stimuli with the factors of sex and phone type. Females reported slightly more visually influenced percepts than did males (3.2%), but this difference was not significant. As in the brief visual stimuli, there was a significant main effect of phone type [$F(1,55 = 178.9$, $p < .001$], with /v/ eliciting more visual influence (82.1%) than did /d/–/ð/ (66.4%).

Visual-Only Stimuli

An ANOVA with the between-subjects factor of sex and the within-subjects factors of motion, number of frames, and phone type was performed for the visual-only stimuli. There were no main effects or significant interactions involving sex ($F < 1$) for the visual-only stimuli. Phone type was not significant in this analysis [$F(1,55) = 2.57$, $p < .20$]. With regard to motion, the static stimuli were more accurately identified (81.7%) than were the dynamic stimuli (68.1%) [$F(1,55) = 81.54$, $p < .001$]. Surprisingly, the two-frame dynamic stimuli were more accurately identified than were the three-frame dynamic stimuli [$F(1,55) = 5.97$, $p < .05$] by 7.2%, but the corresponding difference was negligible for the static stimuli (0.7%).

Responses to the incongruent brief, full, and visual-only conditions for males and females can be seen in Figure 5. Histograms showing group differences in "v" and "d"–"th" responses by sex can be seen in Figures 6 and 7, respectively. Reflecting the significant sex differences for both the /va/ and the /da/–/ða/ stimuli, the histograms depicting the "v" and "d"–"th" responses indicate a different distribution of responses for the male and the female perceivers.

To further examine the relationship between accuracy in lipreading and visual influence in the brief visual condition, correlations were run for both the /va/ and the /d/–/ð/ incongruent stimuli. Neither correlation was significant ($R^2 = .08$ for /v/; $R^2 = .16$ for /d/–/ð/), suggesting that the differences in visual influence for the brief stimuli in the AV condition were not due to the ability to identify the visual information.

The results of Experiment 2 replicate the major findings of Experiment 1, with females showing more visual influence than did males for the brief visual stimuli. Importantly, there was no sex difference in the visual-only condition, indicating that this finding was not due to an overall difference in the ability of males and females to

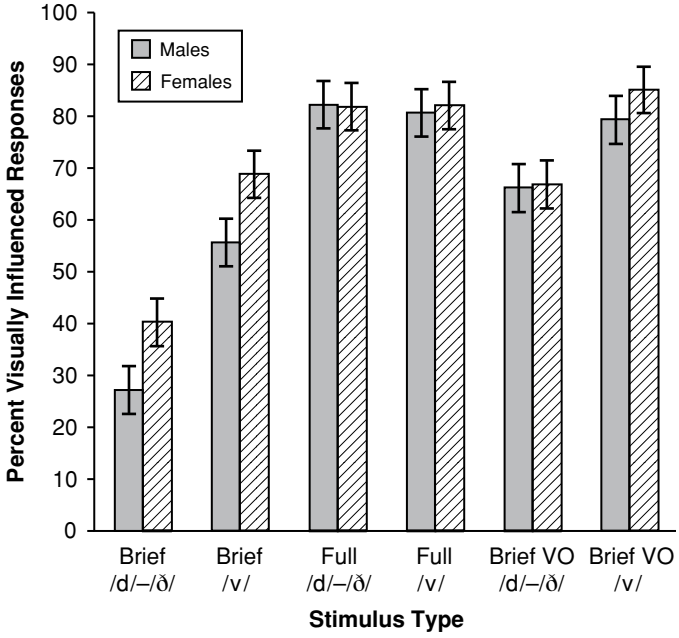

Figure 5. Percentages of visually influenced responses for males and females for incongruent brief and full /d/–/ð/ and /v/ visual stimuli and visual-only (VO) stimuli in Experiment 2.

extract the visual information from brief visual stimuli. The sex difference occurred in the context of incongruent brief visual and audio stimuli. As in Experiment 1, men and women showed more visual influence in the context of visual /va/ than in the context of /da/–/ða/. In Experiment 1, a sex difference was found in the context of visual /da/–/ða/ only; in this experiment, women showed greater audiovisual integration than did men for both the /va/ and

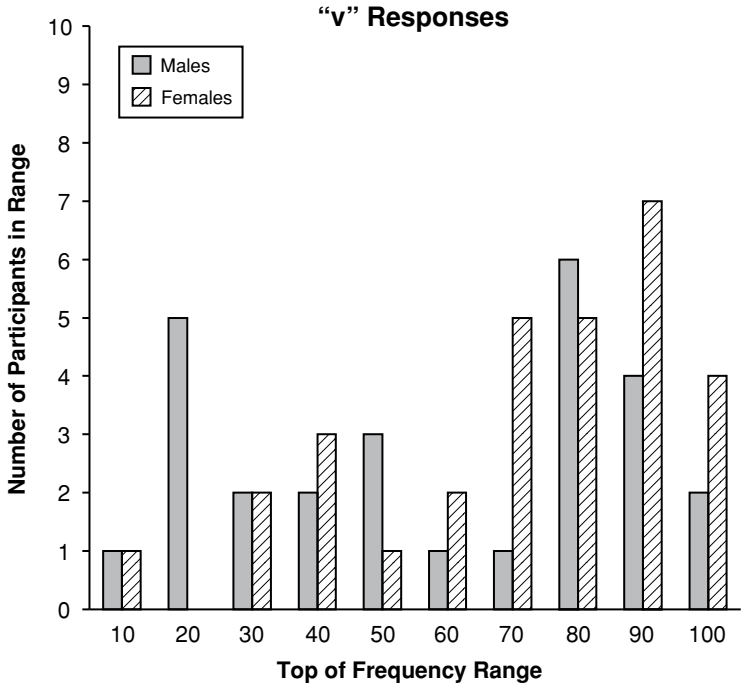

Figure 6. Number of "v" responses to incongruent audiovisual stimuli by sex in Experiment 2.

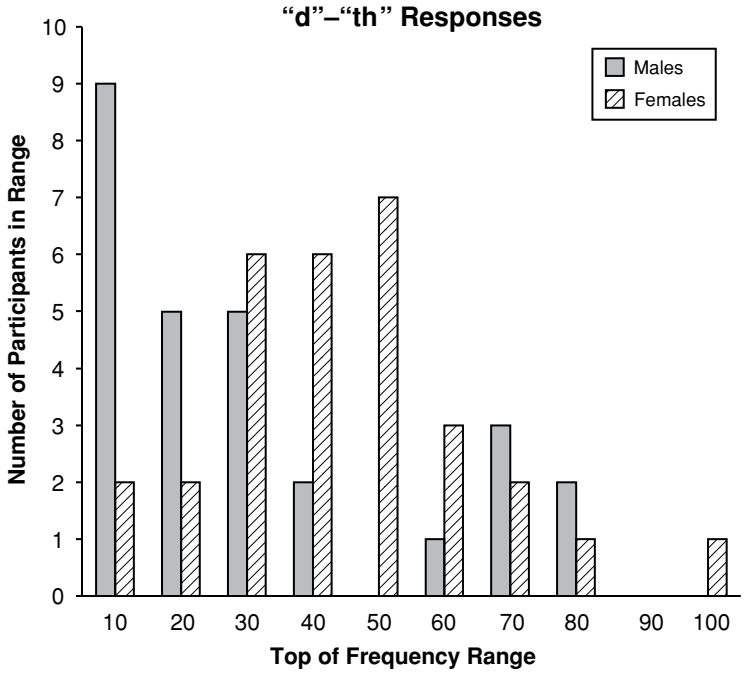

Figure 7. Number of "d"–"th" responses to incongruent audiovisual stimuli by sex in Experiment 2.

the /da/–/ða/ stimuli. A sex difference for both the /va/ and the /da/–/ða/ stimuli in Experiment 2 might have been a function of the change in timing from Experiment 1 to 2. Since the visual signal now sometimes conflicted with the audio, the easier-to-see /va/ tokens might have been less compelling, allowing for sex differences to emerge for /va/, as well as for /da/–/ða/, stimuli.

Greater visual influence for the brief AV static stimuli was not found in this experiment, suggesting that the earlier presentation of brief visual stimuli in Experiment 1 led to an increased visual influence for the static stimuli. In the context of lipreading, the static stimuli appeared to be particularly informative. In the visual-only (lipreading) condition, the static brief visual stimuli were more accurately identified than were the dynamic brief visual stimuli. Unexpectedly, the two-frame visual-only dynamic stimuli were more accurately identified than were the three-frame visual-only dynamic stimuli in this experiment. The intervening black frame may have made the visual information for articulation more prominent in this condition.

GENERAL DISCUSSION

This research explored whether males and females differ in influence of visual on heard speech. To examine this, male and female participants were presented with congruent and incongruent auditory and visual speech syllables, varied for perceptual difficulty. The manipulation of difficulty included presentation of dynamic and static visual signals of varying duration. For both men and women, static faces had a greater influence on phonetic judgments than did dynamic stimuli in Experiment 1. However, when differences in the timing of the static and the dynamic stimuli, present in Experiment 1, were eliminated, no differences in motion were found. For the visual-only (lipread) stimuli, the static stimuli were more informative for the perceivers than were the dynamic stimuli. This suggests that in the context of lipreading, the static visual information about place of articulation is particularly salient. Importantly, the present results indicate that phonetic information can be extracted from a speaker's face and can influence perception whether the visual information is static or dynamic. Perceivers appear to get information about motion from static stimuli extracted from a dynamic event (Freyd, 1983a, 1983b), as well as from dynamic stimuli. For example, using fMRI methodology, Calvert and Campbell (2003) showed that images of both silent still and moving facial speech engage areas associated with the perception of dynamic speech.

In terms of the primary question, whether females were more influenced by visual information in the context of the brief visual stimuli than were males, the expected difference was shown for the incongruent /d/–/ð/ visemes in Experiment 1. In Experiment 2, the female perceivers showed greater visual influence than did the males for both the /d/–/ð/ and the /v/ visual visemes. Notably, men and women did not differ on visual influence for the full video condition (typical for McGurk experiments) or for the visual-only (lipread) stimuli. In both experiments, sex differences were found specifically in the context of the incongruent, brief AV stimuli.

Previous research has shown that women are better at identifying lipread speech than are men for sentences (e.g.,

Johnson, Hicks, Goldberg, & Myslobodsky, 1988; Watson, Qiu, Chamberlain, & Li, 1996). However, shorter speech stimuli, such as the syllables used in the present study, have not been examined previously. In our study, we observed no sex difference in lipreading. Furthermore, there was no significant correlation between accuracy in lipreading and visual influence in the brief AV condition, suggesting that the sex differences in the perception of AV speech are not a function of differences in the ability to lipread.

One possible account for the present findings is that the sex of the speaker influenced perception differently for the male and the female perceivers. In both experiments, a single female speaker produced the CV stimuli. Thus, a comparison of speaker sex was not possible. However, Daly, Bench, and Chappell (1996, 1997) reported that normal-hearing perceivers found female speakers easier to lipread than male speakers. Daly et al. (1996) reported that this effect was equal for both males and females, suggesting that the sex of the speaker in the present experiments is not likely to be the source of the observed sex differences in perception.

A second account of the present results is that there are attentional differences between men and women in the processing of AV speech. Attention has been shown to modulate visual influence on heard speech (Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Massaro & Cohen, 1983; Tiippana, Andersen, & Sams, 2004). Alsius et al. recently reported a reduction in AV integration in a dual-task paradigm in which perceivers were asked to attend to a separate auditory or visual token, placing particularly high attentional demands on the perceiver. Alsius et al. did not examine sex differences, and the present findings do not address the issue of attention directly, because attentional load was not explicitly manipulated. However, in the present experiments, men and women performed equally well in the lipreading (visual-only) task for the attentionally demanding brief visual stimuli, which were the same as those in the brief AV condition, where sex differences were found. Future research will be needed to better explain the role of attention for sex differences in the perception of AV speech.

Another account for the present findings is that differences in language processing between men and women underlie the sex differences in performance. A number of functional neuroimaging studies have shown that females engage the right perisylvian cortex more than do males for phonologically based language tasks (Jaeger et al., 1998; Pugh, Shaywitz, Shaywitz, Constable, et al., 1996; Pugh, Shaywitz, Shaywitz, Fulbright, et al., 1996; Rossell et al., 2002; Shaywitz et al., 1995). In addition, previous research has indicated that there is a right-hemisphere (RH) processing advantage for audiovisually presented speech (Baynes, Funnell, & Fowler, 1994; Diesch, 1995). Bilaterality in female perceivers' neural organization has been proposed to lead to faster or more efficient processing of incongruent auditory and visual speech (Coney, 2002). In the present experiments, sex differences were found for the AV brief speech tokens, which (according to the studies cited above) engage the RH. Thus, we hypothesize that greater bilaterality in language processing may allow for more visual influence on what is heard by female perceivers in the context of the brief visual signals, where the influence of visual information for place of articulation must occur within a very brief window of time. The processing demands involved in the perception of these brief, incongruent AV stimuli appear to elicit differences between male and female perceivers, consistent with Jaeger et al.'s (1998) assertion that sex differences emerge in the context of challenging stimuli. Because sex differences in the processing of AV speech have not been explicitly examined using functional neuroimaging, future research will be needed to assess this tentative account of the present findings.

A closer examination of the neuroanatomical substrates that underlie sex differences during the processing of AV speech is needed. Future study, such as that with functional neuroimaging, can provide new information about how males and females process information about seen and heard speech.

REFERENCES

Adobe Systems (1995). Adobe Premiere Version 5.1 [Computer software]. San Jose, CA: Author.

Aloufy, S., Lapidot, M., & Myslobodsky, M. (1996). Differences in susceptibility to the "blending illusion" among native Hebrew and English speakers. *Brain & Language*, **53**, 51-57.

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

Baxter, L. C., Saykin, A. J., Flashman, L. A., Johnson, S. C., Guerin, S. J., Babcock, D. R., & Wishart, H. A. (2003). Sex differences in semantic language processing: A functional MRI study. *Brain & Language*, **84**, 264-272.

Baynes, K., Funnell, M. G., & Fowler, C. A. (1994). Hemispheric contributions to the integration of visual and auditory information in speech perception. *Perception & Psychophysics*, **55**, 633-641.

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **30**, 445-463.

Brancazio, L., & Miller, J. (2005). Use of visual information in speech perception: Evidence for a visual rate effect both with and without a McGurk effect. *Perception & Psychophysics*, **67**, 759-769.

Calvert, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex*, **11**, 1110-1123.

Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D., & David, A. S. (1999). Response amplification in sensory-specific cortices during cross-modal binding. *NeuroReport*, **10**, 2619-2623.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience*, **15**, 57-70.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, **10**, 649-657.

Campbell, R. (1994). Audiovisual speech: Where, what, when, how? *Current Psychology of Cognition*, **13**, 76-80.

Coney, J. (2002). Lateral asymmetry in phonological processing: Relating behavioral measures to neuroimaged structures. *Brain & Language*, **80**, 355-365.

Conrey, B. L. (2004). Multimodal sentence intelligibility and the detection of auditory–visual asynchrony in speech and nonspeech signals: A first report. *Research on Spoken Language Processing: Progress Report No. 26* (pp. 345-355). Bloomington: Indiana University, Department of Psychology, Speech Research Laboratory.

Conrey, B. L., & Pisoni, D. B. (2004). Detection of auditory–visual asynchrony in speech and nonspeech signals. *Research on Spoken Language Processing: Progress Report No. 26* (pp. 71-94). Bloomington: Indiana University, Department of Psychology, Speech Research Laboratory.

Daly, N., Bench, R. J., & Chappell, H. (1996). Gender differences in speech readability. *Journal of the Academy of Rehabilitative Audiology*, **29**, 1-14.

Daly, N., Bench, [R.] J., & Chappell, H. (1997). Gender differences in visual speech variables. *Journal of the Academy of Rehabilitative Audiology*, **30**, 63-76.

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

Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology*, **45**, 187-203.

Diesch, E. (1995). Left and right hemifield advantages of fusions and combinations in audiovisual speech perception. *Quarterly Journal of Experimental Psychology*, **48A**, 320-333.

Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 1-12.

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

Freyd, J. J. (1983a). The mental representation of movement when static stimuli are viewed. *Perception & Psychophysics*, **33**, 575-581.

Freyd, J. J. (1983b). Representing the dynamics of a static form. *Memory & Cognition*, **11**, 342-346.

Frost, J. A., Binder, J. R., Springer, J. A., Hammeke, T. A., Bellgowan, P. S. F., Rao, S. M., & Cox, R. W. (1999). Language processing is strongly left-lateralized in both sexes: Evidence from functional MRI. *Brain*, **122**, 199-208.

Green, K. P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In R. Campbell, B. Dodd, & D. K. Burnham (Eds.), *Hearing by eye: II. Advances in the psychology of speechreading and auditory–visual speech* (pp. 3-25). Hove, U.K.: Psychology Press.

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.

Green, K. P., & Norrix, L. W. (1997). Acoustic cues to place of articulation and the McGurk effect: The role of release bursts, aspiration, and formant transitions. *Journal of Speech & Hearing Research*, **40**, 646-665.

Gur, R. C., Alsop, D., Glahn, D., Petty, R., Swanson, C. L., Maldjian, J. A., et al. (2000). An fMRI study of sex differences in regional activation to a verbal and spatial task. *Brain & Language*, **74**, 157-170.

Jaeger, J. J., Lockwood, A. H., Van Valin, R. D., Kemmerer, D. L., Murphy, B. W., & Wack, D. S. (1998). Sex differences in brain regions activated by grammatical and reading tasks. *NeuroReport*, **9**, 2803-2807.

Johnson, F. M., Hicks, L. H., Goldberg, T., & Myslobodsky, M. S. (1988). Sex differences in lipreading. *Bulletin of the Psychonomic Society*, **26**, 106-108.

Jones, J. A., & Callan, D. E. (2003). Brain activity during audiovisual speech perception: An fMRI study of the McGurk effect. *NeuroReport*, **14**, 1129-1133.

Josse, G., & Tzourio-Mazoyer, N. (2004). Hemispheric specialization for language. *Brain Research Reviews*, **44**, 1-12.

Kansaku, K., & Kitazawa, S. (2001). Imaging studies on sex differences in lateralization of language. *Neuroscience Research*, **41**, 333-337.

Knecht, S., Drager, B., Deppe, M., Bobe, L., Lohmann, H., Floel, A., et al. (2000). Handedness and hemispheric dominance in healthy humans. *Brain*, **123**, 2512-2518.

MacDonald, J., Andersen, S., & Bachmann, T. (2000). Hearing by eye: How much spatial degradation can be tolerated? *Perception*, **29**, 1155-1168.

Majeres, R. L. (1999). Sex differences in phonological processes: Speeded matching and word reading. *Memory & Cognition*, **27**, 246-253.

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

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

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

Munhall, K. G., & Tokhura, Y. (1998). Audiovisual gating and the time course of speech perception. *Journal of the Acoustical Society of America*, **104**, 530-539.

Öhrström, N., & Traunmüller, H. (2004). Audiovisual perception of Swedish vowels with and without conflicting cues. In *Proceedings, FONETIK 2004* (pp. 40-43). Stockholm: Stockholm University, Department of Linguistics.

Olson, I. R., Gatenby, J., & Gore, J. C. (2002). A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Cognitive Brain Research*, **14**, 129-138.

Paré, M., Richler, R. C., ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Perception & Psychophysics*, **65**, 553-567.

Pugh, K. R., Shaywitz, B. A., Shaywitz, S. E., Constable, R. T., Skudlarski, P., Fulbright, R. K., et al. (1996). Cerebral organization of component processes in reading. *Brain*, **119**, 1221-1238.

Pugh, K. R., Shaywitz, B. A., Shaywitz, S. E., Fulbright, R. K., Byrd, D., Skudlarski, P., et al. (1996). Auditory selective attention: An fMRI investigation. *NeuroImage*, **4**, 159-173.

Riecker, A., Wildgruber, D., Dogil, G., Grodd, W., & Ackermann, H. (2002).Hemispheric lateralization effects of rhythm implementation during syllable repetitions: An fMRI study. *NeuroImage*, **16**, 169-176.

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

Rosenblum, L. D., & Saldaña, H. M. (1998). Time-varying information for visual speech perception. In R. Campbell, B. Dodd, & D. K. Burnham (Eds.), *Hearing by eye: II. Advances in the psychology of speechreading and auditory–visual speech* (pp. 61-81). Hove, U.K.: Psychology Press.

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

Rossell, S. L., Bullmore, E. T., Williams, S. C. R., & David, A. S. (2002). Sex differences in functional brain activation during a lexical visual field task. *Brain & Language*, **80**, 97-105.

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

Shaywitz, B. A., Shaywitz, S. E., Pugh, K. R., Constable, R. T., Skudlarski, P., Fulbright, R. K., et al. (1995). Sex differences in the functional organization of the brain for language. *Nature*, **373**, 607-609.

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

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology*, **16**, 457-472.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, **102**, 1181-1186.

Watson, C. S., Qiu, W. W., Chamberlain, M. M., & Li, X. (1996). Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition. *Journal of the Acoustical Society of America*, **100**, 1153-1162.

(Manuscript received June 22, 2004;
revision accepted for publication July 8, 2005.)