Atten Percept Psychophys (2012) 74:1761–1781
DOI 10.3758/s13414-012-0375-z

# Seeing facial motion affects auditory processing in noise

**Jaimie L. Gilbert · Charissa R. Lansing · Susan M. Garnsey**

Published online: 16 October 2012


**Abstract** Speech perception, especially in noise, may be maximized if the perceiver observes the naturally occurring visual-plus-auditory cues inherent in the production of spoken language. Evidence is conflicting, however, about which aspects of visual information mediate enhanced speech perception in noise. For this reason, we investigated the relative contributions of audibility and the type of visual cue in three experiments in young adults with normal hearing and vision. Relative to static visual cues, access to the talker's phonetic gestures in speech production, especially in noise, was associated with (a) faster response times and sensitivity for speech understanding in noise, and (b) shorter latencies and reduced amplitudes of auditory N1 event-related potentials. Dynamic chewing facial motion also decreased the N1 latency, but only meaningful linguistic motions reduced the N1 amplitude. The hypothesis that auditory–visual facilitation is distinct to properties of natural, dynamic speech gestures was partially supported.

**Keywords** Evoked potentials · Multisensory processing · Speech perception

Everyone experiences difficulty understanding speech at some point, especially in noisy environments or in the presence of a hearing loss. The natural setting for speech communication is in a face-to-face environment where people can both hear the speech and see the face of the talker. Considerable evidence has shown that performance on speech perception tasks improves when information from both auditory and visual modalities is available (Bergeson, Pisoni & Davis 2003; Bernstein, Auer & Takayanagi 2004; Grant & Seitz, 1998; Kaiser, Kirk, Lachs, & Pisoni, 2003; Kim & Davis, 2004; Schwartz, Berthommier, & Savariaux, 2004; Sumby & Pollack, 1954). However, evidence is conflicting about what aspects of the visual speech signal contribute to the auditory–visual facilitation.

From the classic McGurk paradigm (McGurk & MacDonald, 1976), it is well known that visual information influences auditory perception. Facial perception may differ from visual perception of other kinds of objects (see, e.g., Bentin, Allison, Puce, Perez & McCarthy 1996). For instance, a smile on the face may influence how a talker is perceived (Otta, Lira, Delevati, Cesar, & Pires, 1994), and perceivers may look to parts of the face other than the mouth, such as the eyes, before and after a speech utterance (Lansing & McConkie, 2003). In addition, visual motion may elicit special processing (see, e.g., Ahlfors et al., 1999; Bavelier et al., 2001). Results from Schwartz et al. (2004) have suggested that a dynamic nonspeech visual cue (a dynamic rectangle) did not facilitate auditory–visual speech, and that facilitation required a dynamic visual speech gesture. However, results from Bernstein et al. (2004) suggested that, on the contrary, a variety of visual stimuli—including the full face of the talker, a static rectangle, a dynamic rectangle, and a dynamic *Lissajous* figure—all facilitated auditory–visual speech perception.

Another important consideration for auditory–visual speech perception is the quality of the acoustic environment, which affects speech audibility, and may thus influence the use of observable visual information from the face of the talker. For individuals with normal hearing, auditory–visual facilitation may not be measurable in ideal listening conditions, due to near-ceiling-level performance when auditory information is all that is available. However, recent studies have demonstrated that auditory–visual speech in

J. L. Gilbert · C. R. Lansing · S. M. Garnsey
University of Illinois,
Urbana-Champaign, IL, USA

J. L. Gilbert (✉)
School of Communicative Disorders, University of Wisconsin,
1901 Fourth Avenue,
Stevens Point, WI 54481, USA
e-mail: jaimie.gilbert@gmail.com

quiet listening conditions is processed differently than auditory-only speech. The amplitude of the N1 component of the event-related brain potential (ERP) was reduced in response to matching, congruent, and synchronous auditory–visual speech syllables, in comparison to responses to auditory-only speech syllables (Pilling, 2009; van Wassenhove, Grant, & Poeppel, 2005), and also in comparison to the sum of responses to auditory-only and visual-only speech syllables (Besle, Fort, Delpuech, & Giard, 2004). There is also some evidence that in quiet conditions, congruent auditory–visual speech can be processed faster than auditory-only speech (Pilling, 2009; van Wassenhove et al., 2005).

When the audibility of the speech signal is reduced, the N1 response to auditory-only speech syllables is slower and reduced in amplitude (Martin, Kurtzberg, & Stapells, 1999; Martin, Sigal, Kurtzberg, & Stapells, 1997). Environments with reduced audibility, such as speech in noise, provide an opportunity for the visual speech signal to contribute to perception (Sumby & Pollack, 1954). Questions remain, though, about how auditory–visual speech is processed in noise. Modifications to an acoustic signal presented in noise have been demonstrated to alter the neural response pattern in aggregate neural responses recorded in guinea pig auditory cortex (Cunningham, Nicol, King, Zecker, & Kraus, 2002). Unrelated competing stimuli have also been shown to influence a click-evoked neural response (Weihing, Daniels, & Musiek, 2009). In the present experiments, we employed ERPs measured in humans in response to auditory–visual stimuli in order to investigate the influence of different types of visual signals on the processing of acoustic speech syllables in white noise.

The goals of the present experiments were (1) to investigate the neural correlates of auditory–visual facilitation in noise and (2) to test whether neural processing of auditory–visual cues would vary with the type of visual cue. Detailed information about when and how auditory–visual integration occurs in individuals with normal hearing sensitivity is necessary to the development of theories and applications for those whose perceptual processes may be different because of sensory deprivation from deafness during a critical period or because their sensory stimulation is delivered by a cochlear implant.

Three separate experiments were conducted. In Experiment 1, we tested the hypothesis that audibility and type of visual cue affect the identification of /ba/ and /ga/ in a speeded-response task. Experiments 2 and 3 investigated the hypotheses that audibility and type of visual cue, respectively, influenced the neural correlates of auditory–visual facilitation, as measured by the N1 component of the evoked auditory response. Behavioral data (accuracy and response time measures) were collected for all three experiments. The behavioral findings from Experiments 2 and 3 replicated the results observed in Experiment 1, so those are presented in detail only for the first experiment.

## Experiment 1

### Method

#### *Participants*

A group of 16 participants (eight female, eight male) between the ages of 18 and 25 were recruited. All spoke American English as their native language, reported that they were right-handed, and had normal or corrected-to-normal visual acuity. On the basis of self-report, the participants had no significant history of hearing, language, or reading problems or neuropathology. In addition, all of the participants completed an automated hearing test procedure with calibrated headphones within the laboratory sound booth. This automatic procedure (Home Audiometer Hearing Test software, Version 1.83; www.audiometer.co.uk) calculates a hearing threshold level, and it was tested on lab personnel to verify its accuracy as compared to a traditional hearing test conducted in a hearing clinic. Using the automated, calibrated procedure, all of the participants had pure-tone thresholds of 30 dB HL or better at 500, 1000, 2000, 3000, 4000, and 6000 Hz. Participants demonstrated a range of lipreading proficiency (2 %–44 %) on the basis of words-correct scoring (Bernstein, Demorest, & Eberhardt, 1994; Demorest & Bernstein, 1991, 1992), with average performance ($M$ = 27 %, $SD$ = 12 %) similar to the previously reported mean of 20.8 % of words correct (Demorest & Bernstein, 1991). The lipreading screening results indicated that the participants represented a diverse group in terms of their proficiency to extract words from visual-only speech gestures. Written informed consent was obtained, and the participants received monetary compensation or fulfilled a course requirement as partial compensation for their time.

#### *Stimuli*

Audio–video recordings were made of two different talkers (one male and one female) producing five exemplars of the speech syllables /ba/ and /ga/. These syllables were selected to contrast the perception of a gesture with a high degree of visibility (/ba/) with the perception of one with limited visibility (/ga/). Recordings were made in a quiet room with a Canon XL1 HD digital camcorder at 30 frames per second, with an MS condenser microphone. The talkers' faces were illuminated with a Lowell Caselight 5 5400 K. Video recordings were also made of the same individuals chewing gum or displaying a fixed natural smile. The productions of speech and nonspeech facial movements began and ended with a neutral face position.

*Auditory stimuli* Speech stimuli (16-bit sample size, 44.1-kHz sampling rate) were extracted from the audio–video

recording in quiet and equated in average root-mean square (RMS) level. The acoustic characteristics of the speech syllables are detailed in Table 1 (/ba/) and Table 2 (/ga/). In addition, the stimuli were mixed with white noise (bandwidth of 0–22050 Hz), with the level of the noise adjusted to an average RMS value that was equal to that of the speech stimuli, 9 dB more intense than the speech stimuli, or 18 dB more intense than the speech stimuli. Thus, two auditory signals were presented overlapping in time—the speech signal and the white noise, with the white noise preceding the onset of the speech stimulus and continuing after the speech offset. As this was an initial foray into investigations of the neural correlates of auditory–visual speech perception in noise, specifically by measurement of the N1 ERP, we decided to start by observing effects that occurred in broadband white noise. Broadband white noise was selected to ensure that differences in responses were related to the properties of the syllable and not to differential characteristics of the masker. This selection also had the advantage of allowing for comparisons between the present findings and previous research that had also employed broadband noise, in investigations of auditory–visual speech perception (Bernstein et al., 2004; Kim & Davis, 2004; Sumby & Pollack, 1954) and neural activity (Cunningham et al., 2002), and importantly, in a study differentiating activity related to two separate acoustic onsets (Kaplan-Neeman, Kishon-Rabin, Henkin, & Muchnik, 2006). The stimuli were presented at signal-to-noise ratios (SNRs) of 0, –9, or –18 dB. These particular SNRs were chosen on the basis of the results of Sumby and Pollack (1954); the 0- to –18-dB SNR range encompasses differences seen on a function of SNR for the percent difference (between auditory-only and bisensory presentation) of speech intelligibility scores (see Sumby & Pollack, 1954, Fig. 3). For the present experiments, SNRs were determined by holding the level of the speech constant at a presentation level of approximately 60 dB SPL and adjusting the level of the noise.

*Visual stimuli* Four types of visual stimuli were presented: (a) a static rectangle, (b) a static smiling image of the talker's face, (c) a full-motion video of the talker's face producing a chewing motion, and (d) a full-motion video of the talker's face producing the dynamic speech gestures to vocalize /ba/ or /ga/. The video display (600 × 800 pixels [h × w] in the center of the computer screen) presented the complete face of the talker, in frontal view from above the neck to the top of the head, at a rate of 29.97 frames per second (fps). The first frame of each video was presented for 1,000 ms before the video played, and the last frame was presented for 500 ms. Dynamic video motion (talking or chewing) began after the initial 1,000-ms presentation of the first frame of a neutral face.

*Experimental stimuli* Auditory and visual stimuli were paired to create the experimental test stimulus conditions. Auditory speech (in quiet or mixed with noise) was presented with a static rectangle to form the auditory-with-rectangle (AR) condition (Fig. 1), with a static smiling face to form the auditory-with-static-face (ASF) condition (Fig. 2), with dynamic chewing motion on the face to form the auditory-with-dynamic-nonspeech-facial-motion (ADF) condition (Fig. 3), and with visual speech gestures to form the auditory–visual speech (AVS) condition (Fig. 4). The visual-only (VO) condition (Fig. 5) consisted of visual speech gestures presented in the absence of an auditory stimulus or presented with white noise only. In the AVS condition, auditory speech was always paired with the speech gesture that had produced that particular exemplar. The stimuli for each talker consisted of five exemplars per syllable. The initial video frame (or the static image) was presented for 1,000 ms prior to motion onset. The onset of

**Table 1** Acoustic characteristics of /ba/ stimuli

| | Talker | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Female /ba/ | | | | | Male /ba/ | | | | |
| Exemplar | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Voice onset time (ms) | 6 | 6 | 9 | 16 | 7 | 15 | 10 | 8 | 12 | 7 |
| Speech duration (ms) | 322 | 286 | 342 | 281 | 290 | 457 | 362 | 413 | 498 | 461 |
| Stimulus duration (ms) | 1,535 | 1,034 | 1,068 | 1,335 | 1,168 | 934 | 934 | 834 | 1,301 | 1,568 |
| Consonant *F*1 (Hz) | 671 | 666 | 638 | 728 | 593 | 612 | 645 | 623 | 628 | 593 |
| Consonant *F*2 (Hz) | 1,191 | 1,246 | 1,416 | 1,386 | 1,164 | 1,055 | 1,188 | 1,148 | 1,076 | 1,154 |
| Consonant *F*3 (Hz) | 2,506 | 2,344 | 2,472 | 2,430 | 2,451 | 2,477 | 2,484 | 2,498 | 2,385 | 2,501 |
| Vowel *F*1 (Hz) | 860 | 911 | 804 | 872 | 852 | 763 | 744 | 744 | 746 | 755 |
| Vowel *F*2 (Hz) | 1,293 | 1,282 | 1,260 | 1,338 | 1,227 | 1,183 | 1,162 | 1,176 | 1,169 | 1,200 |
| Vowel *F*3 (Hz) | 2,483 | 2,379 | 2,435 | 2,534 | 2,396 | 2,538 | 2,450 | 2,496 | 2,401 | 2,499 |

**Table 2** Acoustic characteristics of /ga/ stimuli

| | Talker | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Female /ga/ | | | | | Male /ga/ | | | | |
| Exemplar | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Voice onset time (ms) | 15 | 18 | 13 | 18 | 20 | 14 | 16 | 14 | 16 | 17 |
| Speech duration (ms) | 362 | 387 | 341 | 345 | 351 | 483 | 498 | 466 | 530 | 514 |
| Stimulus duration (ms) | 1,401 | 1,602 | 1,602 | 1,502 | 1,068 | 1,502 | 1,768 | 1,835 | 1,468 | 1,502 |
| Consonant *F*1 (Hz) | 587 | 743 | 573 | 509 | 615 | 507 | 502 | 562 | 491 | 389 |
| Consonant *F*2 (Hz) | 1,755 | 1,551 | 1,629 | 1,569 | 1,878 | 1,804 | 1,675 | 1,706 | 1,678 | 1,826 |
| Consonant *F*3 (Hz) | 2,671 | 2,405 | 2,396 | 2,123 | 2,528 | 2,535 | 2,739 | 2,586 | 2,451 | 2,714 |
| Vowel *F*1 (Hz) | 761 | 948 | 850 | 819 | 771 | 708 | 735 | 753 | 753 | 753 |
| Vowel *F*2 (Hz) | 1,348 | 1,402 | 1,350 | 1,420 | 1,387 | 1,156 | 1,273 | 1,235 | 1,267 | 1,265 |
| Vowel *F*3 (Hz) | 2,572 | 2,468 | 2,381 | 2,440 | 2,390 | 2,437 | 2,456 | 2,464 | 2,391 | 2,485 |

the initial image coincided with the onset of the white noise, such that at least 1,000 ms separated the onset of the noise and the onset of speech sounds (Kaplan-Neeman et al., 2006). Motion for the video stimuli began following this 1,000-ms static presentation. However, as speech requires prearticulatory motion, the onset of auditory speech did not occur exactly 1,000 ms following the beginning of the trial. For all of the experimental stimuli, the auditory speech signal began at the same time as it would have, had the participant been viewing the combined auditory-plus-visual-speech condition.

*Experimental design*

Experimental trials were organized into ten blocks (two per visual cue condition), each consisting of 80 trials and lasting approximately 8 min. In each block, one of the possible visual cues was paired with auditory stimuli in each of the four possible acoustic environments (quiet or 0, –9, or –18 dB SNR), with equal numbers of presentations of /ba/ and /ga/ in each acoustic environment. The orders of trials in each acoustic environment and of the two syllables were randomized within blocks. Across blocks, all possible combinations of visual cue and acoustic environment (quiet and noise levels) were presented. The order of blocks using different visual cues was presented in a new randomized order for each participant, with the exception that the visual-only (without sound) condition was always presented last, to mitigate the development and use of visual-only strategies for auditory–visual conditions.

Productions of the stimuli from two talkers were tested (one male, one female). Each participant received stimuli in all experimental conditions, but saw and heard stimuli produced by only one of the talkers, with half seeing the female and half the male talker. The participants were asked to press

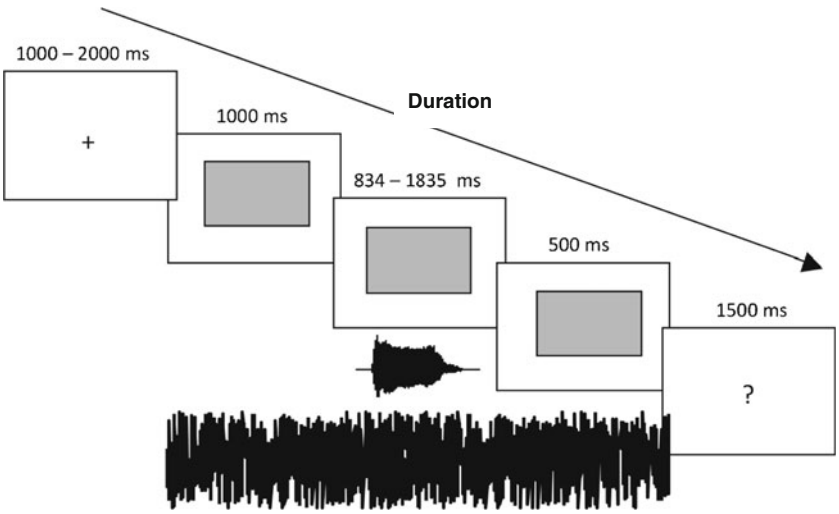


**Fig. 1** Schematic of a trial in the auditory–rectangle (AR) visual cue condition

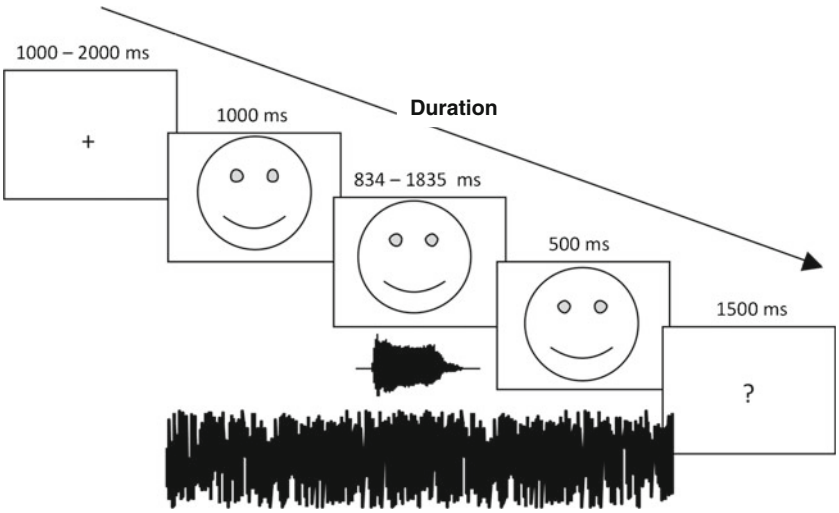


**Fig. 2** Schematic of a trial in the auditory–static face (ASF) visual cue condition

a button to respond to a yes/no question after each trial. The yes/no design allowed for the application of signal detection theory. To control for the possibility of variability in the task demands for perceiving the different syllables, half of the participants were asked, "Do you think it was /ba/?" and the other half were asked, "Do you think it was /ga/?" Participants were instructed to respond as quickly and as accurately as possible. Responses were measured for 2 syllables (/ba/, /ga/) × 4 acoustic environments (quiet, 0 dB SNR, –9 dB SNR, –18 dB SNR) × 5 visual cues (AR, ASF, ADF, AVS, VO), with 20 trials per cell, for a total of 800 trials for each participant.

*Procedure*

Written instructions and practice items were provided prior to the experimental trials. The instructions and the task ("Do you think it was . . .") were designed to avoid biasing attention toward a single modality. Practice items allowed participants to ask any questions about the task prior to beginning data collection. Auditory stimuli were presented via an insert earphone (ER-3A, Etymotic Research) to the right ear. Visual stimuli were presented on a cathode ray tube (CRT) computer monitor with 1,280 × 1,024 pixel resolution, using 32-bit color quality with a 75-Hz refresh rate. Participants were seated such that the distance between their eyes and the monitor was approximately 85 cm, with chair height adjusted so that their eye heights were approximately level with the center of the monitor.

Each trial within a block was followed by a screen reminding the participant to respond to the trial (1,500-ms duration). Participants were instructed to respond as quickly and as accurately as possible, by pressing buttons for either a "yes" or a "no" answer on a Cedrus response pad (Model RB-830). Subsequently, a screen appeared with a fixation cross centered in the middle of the screen to indicate that the participant should prepare for the next trial. The interval of the fixation cross ranged between 1,000 and 2,000 ms and was randomly varied by Presentation software (Version 10.0, Build 07.03.06; www.neurobs.com). Participants were encouraged to take breaks to reduce fatigue. Following the experimental blocks, participants completed a lipreading

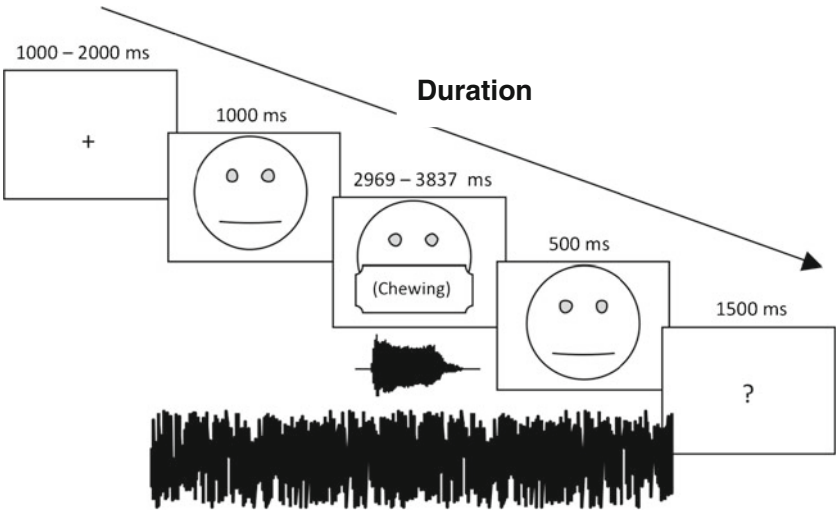


**Fig. 3** Schematic of a trial in the auditory–dynamic face (ADF) visual cue condition

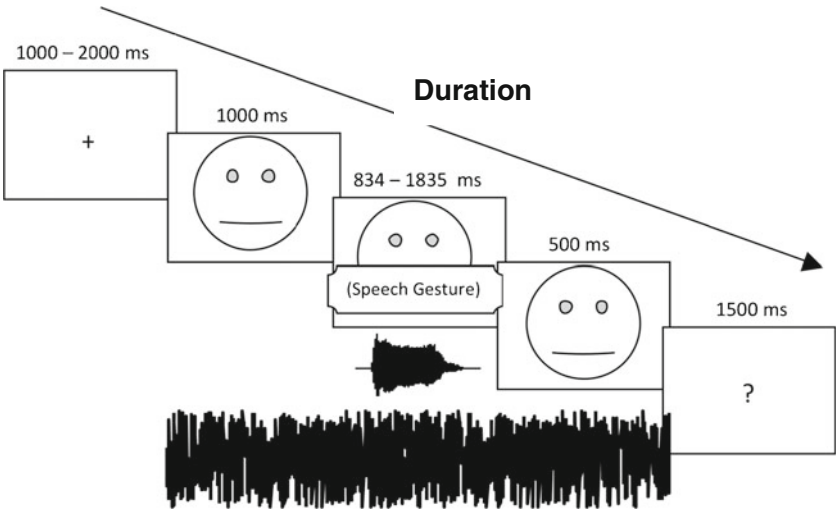


**Fig. 4** Schematic of a trial in the auditory–visual speech (AVS) visual cue condition

screening measure (Bernstein, Demorest, Coulter & O'Connell 1991; Eberhardt, Bernstein, Demorest, & Goldstein, 1990) consisting of 20 Central Institute for the Deaf (CID) everyday sentences (Davis & Silverman, 1978) spoken by a male talker, digitized from the Bernstein and Eberhardt (1986) corpus. The purpose of screening lipreading performance was to ascertain that the participants' lipreading proficiencies were consistent with that of the general population on average (Demorest & Bernstein, 1991), and that a range of proficiency on extracting words from visual-only speech information was represented.

*Data analyses*

Signal detection theory (Green & Swets, 1966) was applied in order to analyze response sensitivity and bias across experimental conditions. A hit was defined as a correct response, indicating that the target syllable was identified when the target syllable was in fact presented (e.g., the task "Do you think it was /ba/?", acoustic stimulus /ba/, and response "yes"). A false alarm was defined as an incorrect response, indicating that the target syllable was identified when the other syllable was presented (e.g., the task "Do you think it was /ba/?", acoustic stimulus /ga/, and response "yes"). Repeated measures analyses of variance (ANOVAs) were performed for each measure. In this and all following experiments, significance for omnibus ANOVAs was determined at the $p < .01$ level with Huynh–Feldt corrections (uncorrected degrees of freedom will be reported), and significance for follow-up testing was defined at the $p < .05$ level with Bonferroni corrections (with the alpha level set to the value of .05 divided by the number of comparisons). The between-subjects factors Talker (male, female), Target Syllable Task (Do you think it was /ba/?, Do you think it was /ga/?), and Sex of the Participant were initially included in this and all following analyses for all experiments. However, since these factors failed to reach significance, they are excluded from all results reported in all experiments. Effect sizes are reported in terms of *f* according to Cohen's (1988) power tables, and are categorized as small ($f = .10$), medium ($f = .25$), or large ($f = .40$). All significant main effects in all three experiments had large effect sizes.

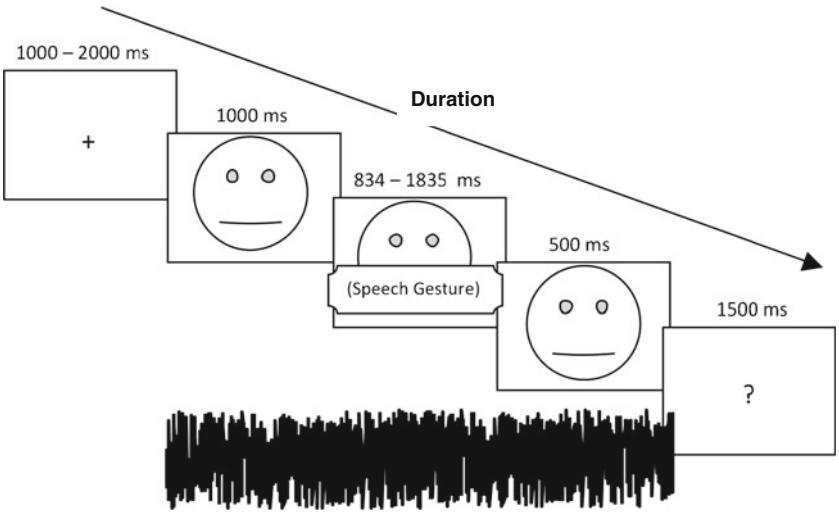


**Fig. 5** Schematic of a trial in the visual-only (VO) visual cue condition

## Results

Response sensitivity was better overall in noise when visible speech gestures were present than when nonspeech visual cues were present. The participants demonstrated a bias to answer "no" in response to /ba/ presented in noise at both 0 and –9 dB SNR, and also when it was presented with nonspeech visual cues. Response times were slower overall in the most intense noise environment tested in this experiment (–18 dB SNR), but the effect of noise was significant only with nonspeech visual cues. When acoustic syllables were presented with visual speech gestures, the –18-dB SNR did not slow down response times.

### *Accuracy: Signal detection analysis*

Sensitivity as measured by $d'$ was reliably affected by both visual cue type [$F(4, 56) = 99.87$, $p < .001$, $f = .70$] and acoustic environment [$F(3, 42) = 196.09$, $p < .001$, $f = .80$]. The participants' ability to distinguish the signal from the noise decreased as the noise level increased when no visible speech gestures were available. A significant Visual Cue × Acoustic Environment interaction [$F(12, 168) = 78.11$, $p < .001$] revealed that sensitivity decreased especially at the most challenging noise level in the absence of visible speech gestures (Fig. 6). Sensitivity decreased significantly at –18 dB SNR as compared to –9 dB SNR only in the AR, ASF, and ADF visual cue conditions. In these nonspeech visual cue conditions, sensitivity was also decreased significantly at –9 dB SNR, as compared to the quiet listening conditions.

The analysis of response bias revealed a difference between the two syllables [$F(1, 14) = 18.46$, $p = .001$, $f = .70$], with participants being more likely to respond "no" for /ba/ trials [/ba/, $M = 1.44$, $SD = 0.77$; /ga/, $M = 0.92$, $SD = 0.37$]. A significant Syllable × Visual Cue interaction [$F(4, 56) = 8.31$, $p < .001$] revealed a significant bias to answer "no" for /ba/ in the AR, ASF, and ADF presentation conditions. We also found a reliable Syllable × Acoustic Environment interaction [$F(3, 42) = 7.08$, $p = .004$], because participants were biased to respond "no" to /ba/ relative to /ga/ at 0 and –9 dB SNR. The three-way Visual Cue × Acoustic Environment × Syllable interaction was also significant [$F(12, 168) = 3.72$, $p = .005$]. Follow-up testing revealed similar response bias $\beta$ statistics across syllables in the 0- and –9-dB SNR acoustic environments in the nonspeech visual cue presentation conditions. This pattern of results suggests that participants were more likely to respond "yes" and "no" equally when visible speech gestures were presented or when the task could be completed using information from one sensory modality (auditory-only in quiet, or visual-only at –18 dB SNR). With nonspeech visual cues at 0 and –9 dB SNR, perhaps when the information from both modalities was salient, there was a bias to respond "no" to /ba/.

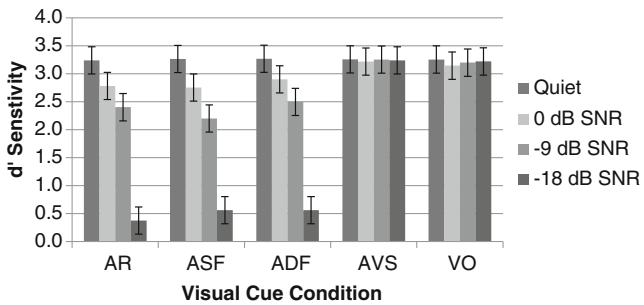


**Fig. 6** Experiment 1 mean $d'$ sensitivity as a function of visual cue and acoustic environment. Error bars represent standard errors of the means. AR, auditory–rectangle; ASF, auditory–static face; ADF, auditory–dynamic face; AVS, auditory–visual speech; VO, visual-only

### *Response time*

The results for response times were similar when analyzing all responses and when analyzing only correct (hit) responses. The statistics from all responses are reported to provide an analysis with more-similar *n*s across conditions. We found significant main effects of both visual cue [$F(4, 60) = 38.84$, $p < .0001$, $f = .70$] and acoustic environment [$F(3, 45) = 112.74$, $p < .0001$, $f = .80$] on response times. Response times were significantly faster when visual speech gestures were presented (for AVS, $M = 823.14$ ms, $SD = 260.47$ ms; for VO, $M = 789.29$ ms, $SD = 266.56$ ms) than when they were not (for AR, $M = 1{,}141.28$ ms, $SD = 313.88$ ms; for ASF, $M = 1{,}109.45$ ms, $SD = 316.29$ ms; for ADF, $M = 1{,}118.13$ ms, $SD = 425.90$ ms).

Across acoustic environments, responses were faster in both the quiet condition ($M = 879.01$ ms, $SD = 250.49$ ms) and the 0-dB SNR condition ($M = 907.70$ ms, $SD = 261.84$ ms) than in the two highest-noise-intensity conditions. Responses were also faster at –9 dB SNR ($M = 946.38$ ms, $SD = 272.36$ ms) than at –18 dB SNR ($M = 1{,}251.95$ ms, $SD = 465.55$ ms).

A significant Visual Cue × Acoustic Environment interaction [$F(12, 180) = 54.22$, $p < .0001$] revealed that responses were slower in the –18-dB than in the –9-dB SNR environment only when visual speech cues were absent (Fig. 7). Acoustic environment also interacted with syllable [$F(3, 45) = 7.99$, $p = .0007$], but differences in the response times to syllables at the same SNR did not reach significance. A faster response time to /ga/ ($M = 1{,}215.69$ ms, $SD = 444.42$ ms) than to /ba/ ($M = 1{,}288.21$ ms, $SD = 485.84$ ms) at –18 dB SNR approached significance, but the difference did not survive the Bonferroni corrections. The significance of the Acoustic Environment × Syllable interaction in the omnibus ANOVA may have been confounded by a marginally significant Visual Cue × Acoustic Environment × Syllable interaction in the omnibus ANOVA [$F(12, 180) =$

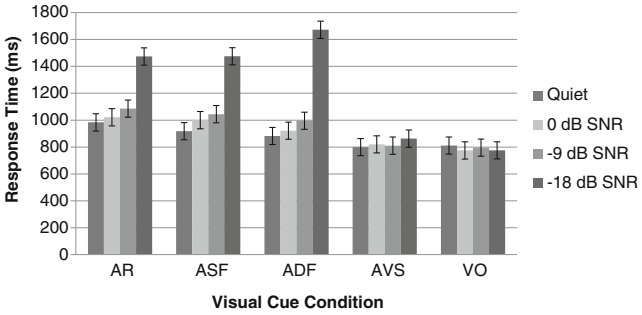


**Fig. 7** Experiment 1 mean response times as a function of visual cue and acoustic environment. Error bars represent standard errors of the means. AR, auditory–rectangle; ASF, auditory–static face; ADF, auditory–dynamic face; AVS, auditory–visual speech; VO, visual-only

2.25, $p = .0115$ (significance set at $p < .01$ for the omnibus ANOVA)]. In testing comparisons, a marginally significant difference was found, with responses being faster to /ga/ than to /ba/ at –18 dB SNR with static smiling face (ASF) visual cues [$F(1, 15) = 10.72$, $p = .0051$ (significance set at $p < .05$ with ten comparisons tested; therefore, $p < .005$ with Bonferroni corrections)].

## Discussion

The measures of response sensitivity and response time were improved in noisy acoustic environments when participants saw visible speech gestures, as compared to seeing nonspeech visual cues. In the most intense noise environment tested (noise level at 78 dBA SPL, or a –18-dB SNR), response times were somewhat faster to /ga/ than to /ba/, mainly when a nonspeech visual cue, the static smiling face, was presented. Since place of articulation is not a robust acoustic cue for perceiving consonants in noise (Miller & Nicely, 1955), people may need help from visual cues when determining consonant identity in noisy environments. The slower responses for /ba/ when accompanied by a static face suggest that listeners rely on visual cues especially for sounds that are produced with articulatory motion that is visible on the face, such as /ba/, and do so less for sounds with less visible articulation.

An alternative explanation for syllable differences is the role of an upward spread of masking with broadband white noise. Formant frequencies, especially *F*2, were generally higher for /ga/ than for /ba/ (see Tables 1 and 2), suggesting that in greater levels of noise, more masking would occur for /ga/. However, it is unlikely that an upward spread of masking impacted our results, as response times were generally faster for the syllable that would have been more masked. In addition, no syllable differences were evident in the *d′* sensitivity analyses; if more masking had occurred for /ga/, there should have been differences in both sensitivity and response time measures. Individuals also differ in their susceptibility to masking effects (Neff & Dethlefs, 1995), such that differential effects of the upward spread of masking on the syllables might not be consistent across participants. The data suggest instead that the manipulation of the type of visual cue was a key factor in the observed participant behaviors.

No consensus exists in the literature regarding the influence of nonspeech visual cues on behavioral measures of speech perception such as detection and identification. Bernstein et al. (2004) found lower detection thresholds with auditory–visual presentations, regardless of whether the visual stimuli consisted of speech gestures or static/dynamic geometric shapes. In contrast, Schwartz et al. (2004) found better identification of voiced plosives only when the visual stimuli consisted of the visible speech gestures, not when they consisted of a dynamic geometric shape. In the present experiment, behavioral measures of response sensitivity and response time were also improved only by visible speech gestures, not by the nonspeech visual cues of a chewing motion, a static face, or a static rectangle.

The production of some speech sounds involves preparatory motion as a talker moves from a resting state to articulation of an isolated phoneme, and also during coarticulation in natural, continuous speech. When visual precues such as this were included in Bernstein et al.'s (2004) experiments, there was an advantage for speech gestures over nonspeech visual cues, consistent with the results obtained here and by Schwartz et al. (2004). Kim and Davis (2004) reported that dynamic visual speech gestural movement was necessary for an auditory–visual advantage in a two-interval forced choice task, but the advantage went away when the speech gesture was reversed temporally, suggesting that the local temporal correlation between visual and auditory speech cues was also important.

The results from Experiment 1 indicated that only visual speech gestures led to an improvement in behavioral measures of response sensitivity and response time. The influence of visible speech gestures on measures of behavioral performance was most noticeable in the –9- and –18-dB SNR acoustic environments. However, performance did not differ between the AVS and VO presentation conditions, suggesting that visual information alone was sufficient to accomplish this yes/no behavioral task in these acoustic environments, with simple speech stimuli that were highly contrastive in visual saliency. The influence of visual information at earlier stages in the processing of these stimuli, as reflected in the ERPs, was investigated in the following experiments.

# Experiment 2

The aim of Experiment 2 was to investigate the processing of auditory–visual stimuli, as measured by the N1 ERP

response, in quiet and at different SNRs. ERP responses to auditory syllables (presented in quiet and in white noise) were compared when accompanied by either a dynamic chewing face or a visual speech gesture. The selection of the ADF and AVS visual cue conditions provided the opportunity to evaluate the effect of audibility with speech as compared with nonspeech visual motion.

## Method

### *Participants*

A new group of 16 participants (eight female, eight male) participated in Experiment 2. The participant eligibility criteria were the same as those for Experiment 1. As in Experiment 1, participants denied any hearing problems and had pure-tone thresholds of 30 dB HL or better at 500, 1000, 2000, 3000, 4000, and 6000 Hz, as measured in the laboratory by the automated, calibrated hearing test procedure. Participants again varied in their ability to extract words from visual-only speech, with a range of 8 %–40 % words correct ($M$ = 22 %, $SD$ = 10 %) on a lipreading test (Demorest & Bernstein, 1991).

### *Stimuli*

The auditory and visual stimuli were a subset of those used in Experiment 1. In particular, only the ADF and AVS visual cue conditions were used.

### *Experimental design*

The amplitude and latency of the N1 potential, along with the accuracy and response times for behavioral responses, were measured. Sixteen conditions were created in a 2 visual cues (ADF, AVS) × 4 acoustic environments (quiet, 0 dB SNR, –9 dB SNR, –18 dB SNR) × 2 syllables (/ba/, /ga/) design with 100 trials per condition, for a total of 1,600 trials per participant. The experimental trials were organized into blocks lasting approximately 10 min, with eight blocks per experimental session. Each block consisted of one visual cue condition accompanying randomized acoustic environments and syllables, with equal numbers of presentations of /ba/ and /ga/. The between-subjects factors [Talker (male, female), Target Syllable Task ("Do you think it was /ba/?", "Do you think it was /ga/?"), and Sex of the Participant] were identical to those of Experiment 1. Eight participants were presented with stimuli produced by the male talker, of which four responded to the question "Do you think it was /ba/?", and of these four participants, two were male, one of whom received the ADF condition first, and the other received the AVS condition first.

### *Procedure*

The procedures were similar to those of Experiment 1, with the addition of electroencephalograph (EEG) recording, a lab visit, and a second experimental session. ERPs were recorded via Ag–AgCl electrodes snapped onto a cap (Easy Cap Modular EEG-Recording Caps, EASYCAP; www.easycap.de/easycap/), using InstEP IWave software (Version 5.21) with Grass Model 12 Neurodata Acquisition System amplifiers (Grass Instrument Co., West Warwick, RI). The scalp recording locations (referred to as the Channel factor in the analyses), based on the 10–10 international electrode system (Chatrian, Lettich, & Nelson, 1985, 1988), were Fz, Cz, Pz, F3, F4, C3, C4, C5, C6, CP5, CP6, FC3, FC4, CP3, CP4, P3, and P4. The recording parameters were set to a bandpass of 0.01–30 Hz; the data sampling rate was 200 Hz. Recordings were referenced to the left mastoid online and later re-referenced to the averaged mastoid locations (A1, A2). Bipolar electrode pairs above and below the right eye and just lateral to each eye were used to monitor eye blinks and saccades. Recording electrode impedances were below 5 kΩ, and the ground electrode was below 10 kΩ.

Participants completed a lab visit to verify eligibility and to become familiarized with the setup procedures required for recording the ERPs. Data were collected across two sessions, with all conditions (ADF and AVS at each SNR) presented at both sessions in order to reduce the chances that the results would differ across sessions for spurious reasons. All other procedures were identical to those for Experiment 1.

### *Data analyses*

*Waveform analysis* ERP waveforms were time-locked to the acoustic stimulus onset (i.e., burst release). Single-trial epochs, with a 100-ms baseline before the burst and 500 ms after it, were extracted and first submitted to an eye movement correction program (Gratton, Coles, & Donchin, 1983), and then to an artifact rejection procedure. Trials were rejected and excluded from the analysis if the amplitude in any recording (EEG) channel exceeded a criterion value of ±100 $\mu$V. N1 waveforms were analyzed for peak amplitude and latency within an 80- to 200-ms time window. Peaks were identified via an automatic detection program that selected the time and value of the largest amplitude—for instance, in the case of a double peak. After the data analysis, a digital smoothing filter (bandpass 0–15 Hz) was applied to the averaged waveforms for plotting.

*Statistical analyses* N1 amplitudes and latencies were submitted to separate repeated measures ANOVAs: 2 visual cues (ADF, AVS) × 4 acoustic environments (quiet, 0 dB

SNR, –9 dB SNR, –18 dB SNR) × 2 syllables (/ba/, /ga/) × 17 electrodes (Fz, Cz, Pz, F3, F4, C3, C4, C5, C6, CP5, CP6, FC3, FC4, CP3, CP4, P3, P4). The results were similar for all trials and for only correct (hit) trials. The results reported here reflect the analyses performed on all trials.

## Results

### *Behavioral responses*

Response sensitivity and response times showed the same pattern as in Experiment 1, with the presence of visual speech cues mitigating the effects of noise in the acoustic environment (see Figs. 8 and 9). As this subset of conditions replicated the findings from Experiment 1, these will not be discussed in further detail. Instead, the focus of this discussion will be the new information revealed in the analysis of the ERPs.

### *Event-related potentials*

Visual inspection of grand mean waveforms revealed that N1 responses were absent at –18 dB SNR, so this condition was excluded from further analysis. The probable reason for the absence of these waveforms in the most difficult listening condition is that the speech was not reliably audible in this level of noise (e.g., Martin et al., 1999; Martin et al., 1997). The artifact rejection criteria led to excluding 2.29 % of all trials (1 %–4 % within each condition, approximately equally distributed across conditions) from the analysis. The between-subjects factors Talker, Target Syllable Task, and Sex of the Participant failed to reach significance. Thus, data were collapsed over those factors and compared for each syllable, in three acoustic environments, with two types of visual cues. Plots of the N1 response for /ga/ in the ADF (chewing) condition across scalp locations are shown in Fig. 10; plots of the N1 response in different conditions recorded at the Cz electrode are shown in Fig. 11.

*N1 amplitude* Significant main effects occurred for visual cue [$F(1, 15) = 52.59$, $p < .0001$, $f = .80$], acoustic environment [$F(2, 30) = 36.36$, $p < .0001$, $f = .80$], and channel [$F(16, 240) = 16.12$, $p < .0001$, $f = .40$]. N1 amplitudes were smaller in the AVS condition ($M = –1.80\ \mu V$, $SD = 2.29\ \mu V$) than in the ADF condition ($M = –4.33\ \mu V$, $SD = 2.16\ \mu V$). With respect to the acoustic environment, amplitudes were largest in quiet conditions ($M = –4.05\ \mu V$, $SD = 2.63\ \mu V$), were reliably smaller at 0 dB SNR ($M = –2.93\ \mu V$, $SD = 2.44\ \mu V$), and were smallest at –9 dB SNR ($M = –2.21\ \mu V$, $SD = 2.26\ \mu V$). The amplitudes of the effects of both visual cue and noise level were largest over central and left hemisphere scalp locations, with the greatest amplitudes at electrode C3. At each electrode location, amplitudes were smaller in the AVS condition than in the ADF condition, but the Visual Cue × Channel interaction was significant [$F(16, 240) = 9.75$, $p < .0001$], likely due to a decrease in the amplitude difference across visual cue conditions at more frontal and right locations—for instance, F4 and C6. Similarly, although the amplitude at each of the 17 electrodes decreased as noise level increased, a significant Acoustic Environment × Channel interaction occurred [$F(32, 480) = 4.33$, $p = .0005$]. Channel amplitude differences were greater between quiet and 0 dB SNR than between 0 and –9 dB SNR; amplitude differences across acoustic environments were decreased at posterior and lateral electrode locations—for instance, Pz and C6. In summary, N1 amplitudes decreased when the visual cue was a speaking face relative to when it was a chewing face, and also decreased as noise was introduced to the acoustic environment and increased in intensity, regardless of visual cue type.

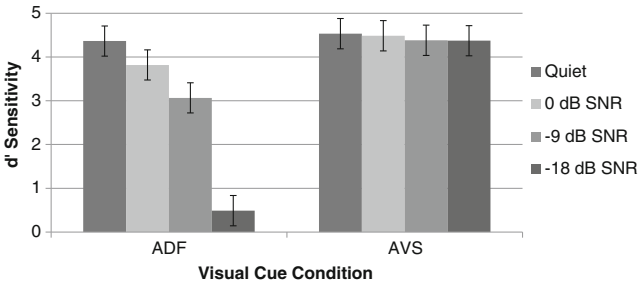


**Fig. 8** Experiment 2 mean *d′* sensitivity as a function of visual cue and acoustic environment. Error bars represent standard errors of the means. ADF, auditory–dynamic face; AVS, auditory–visual speech

*N1 latency* Significant main effects of visual cue [$F(1, 15) = 44.31$, $p < .0001$, $f = .80$], acoustic environment [$F(2, 30) = 74.17$, $p < .0001$, $f = .80$], syllable [$F(1, 15) = 30.78$, $p = .0001$, $f = .80$], and channel [$F(16, 240) = 8.96$, $p < .0001$, $f = .40$] on N1 latencies emerged. Latencies were faster in the AVS condition ($M = 127.56$ ms, $SD = 28.83$ ms) than in the ADF condition ($M = 144.89$ ms, $SD = 25.20$ ms), as well as fastest in quiet ($M = 119.38$ ms, $SD = 18.52$ ms), slower at 0 dB SNR ($M = 138.30$ ms, $SD = 23.19$ ms), and slowest at –9 dB SNR

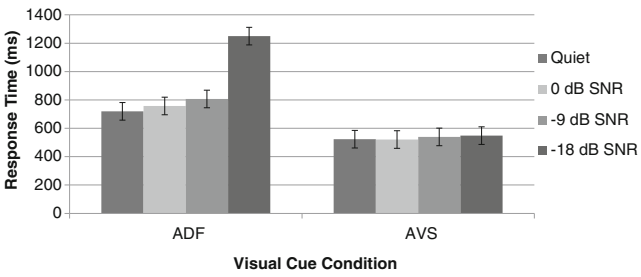


**Fig. 9** Experiment 2 mean response times as a function of visual cue and acoustic environment. Error bars represent standard errors of the means. ADF, auditory–dynamic face; AVS, auditory–visual speech

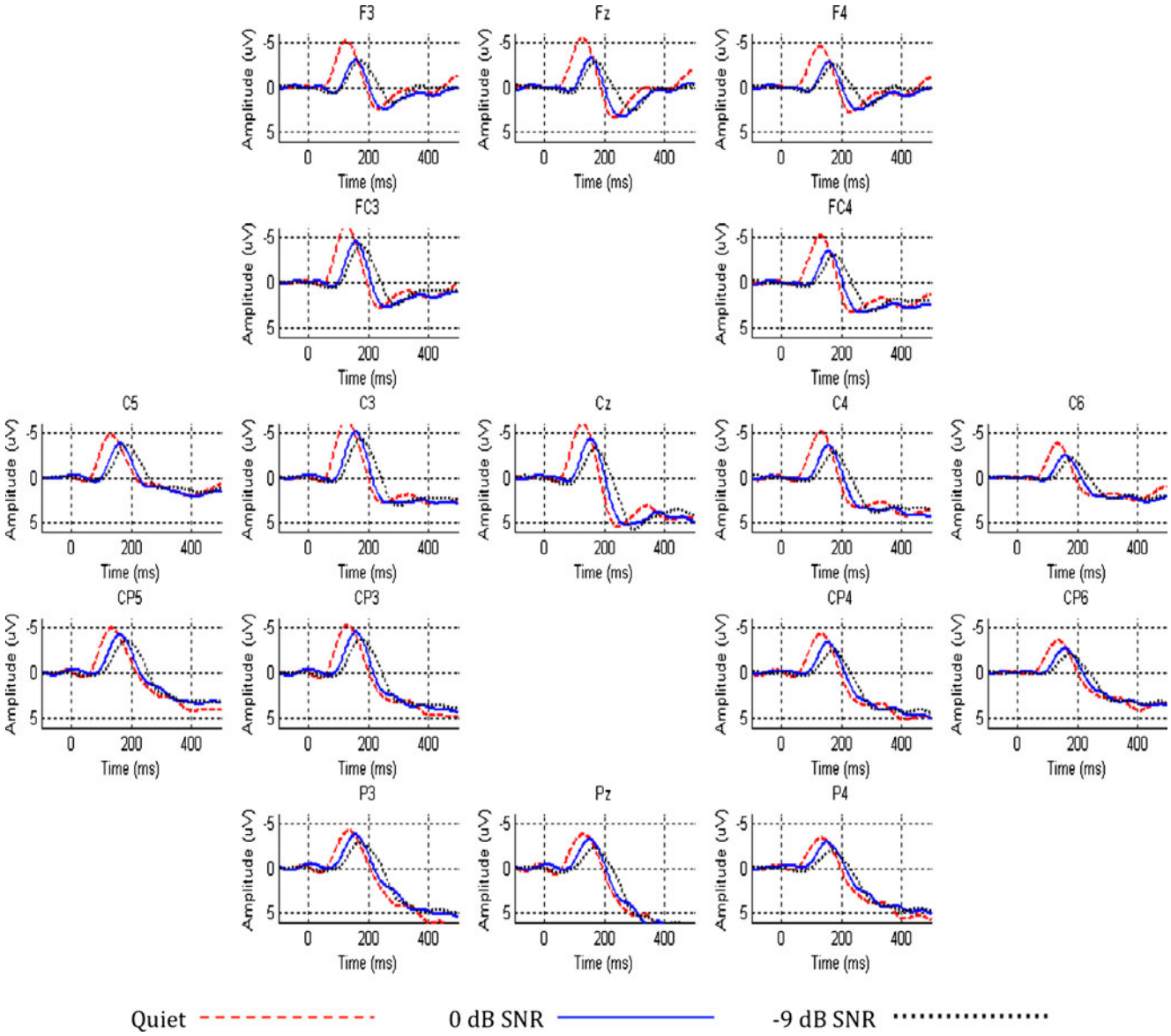


**Fig. 10** Experiment 2 N1 plots for /ga/ in the auditory–dynamic face (ADF) visual cue condition, time-locked to acoustic onset; negative is plotted up

(*M* = 151.01 ms, *SD* = 32.21 ms). Latencies for /ba/ (*M* = 129.20 ms, *SD* = 25.76 ms) were 14 ms faster on average than latencies for /ga/ (*M* = 143.25 ms, *SD* = 29.23 ms). Latencies were faster at posterior electrode locations, with the fastest latency at midline Pz.

Several significant interactions occurred for N1 latencies. Visual cue interacted with both acoustic environment [$F(2, 30) = 5.98$, $p = .0081$] and channel [$F(16, 240) = 5.38$, $p < .0001$]. Acoustic environment also interacted with syllable [$F(2, 30) = 6.81$, $p = .0052$]. In the ADF presentation condition, all three acoustic environments were significantly different from each other, with the fastest latencies in quiet and the slowest at –9 dB SNR. In the AVS presentation condition, the quiet condition was significantly faster than both noise conditions, with no difference between the noisy environments. Latencies were faster in the AVS condition than in the ADF condition across all channels, but the latency difference between conditions was reduced at frontal locations—for instance, Fz, F3, and F4. For each syllable, latencies were significantly slower in progressively poorer acoustic environments. Latencies for /ba/ were significantly shorter than latencies for /ga/ in both noise conditions, but not in quiet.

In sum, N1 latencies were faster when visual speech gestures accompanied the auditory stimuli, and also faster when the quality of the acoustic environment improved. However, when speech gestures were visible, the presence but not the intensity level of the noise affected latencies,

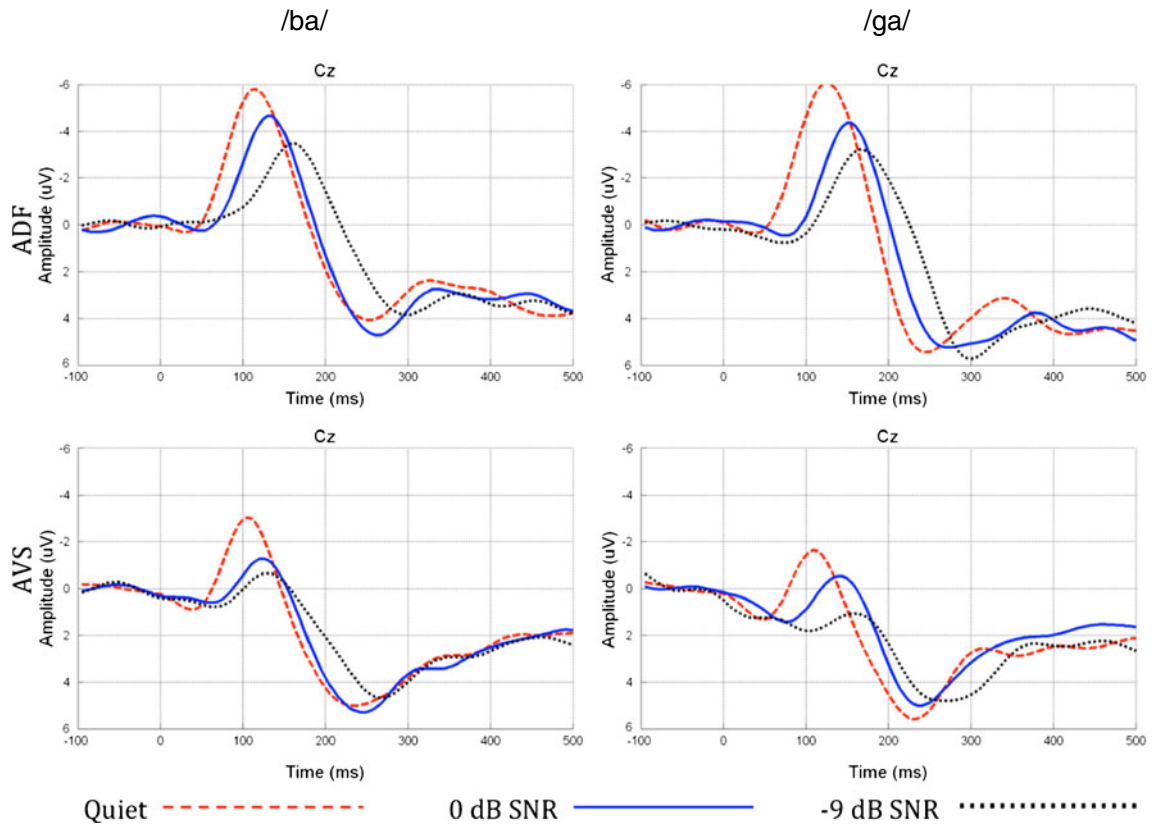


**Fig. 11** Experiment 2 N1 plots at the Cz electrode, time-locked to acoustic onset; negative is plotted up. ADF, auditory–dynamic face; AVS, auditory–visual speech

particularly for the /ba/ syllable. Overall, latencies were faster for /ba/ than for /ga/, but only in noise. In the quiet condition, we found no difference in latencies between the syllables.

## Discussion

The literature reports (e.g., Martin et al., 1999; Martin et al., 1997) that the N1 response is present in response to consonant–vowel syllables, provided that acoustic speech energy is audible. In the present experiment, no N1 waveforms were measurable at –18 dB SNR, suggesting that this level of noise masked the acoustic speech information. This was true in both the chewing (ADF) and AVS conditions, suggesting that the N1 waveform reflects auditory processing of audible speech. Thus, the changes in N1 morphology to each individual acoustic speech syllable with different visual cues could be attributed to the influence of visual information on auditory processing. However, the effect of the quality of the auditory environment, which in this experiment was degraded with white noise, also interacted with the influence of visual information. N1 latencies were significantly slower at –9 than at 0 dB SNR with the dynamic chewing gesture; the N1 latencies for auditory–visual speech did not differ significantly between these two environments. This finding suggests that the processing speed for auditory–visual speech (acoustics with the matching articulatory gesture) may be less sensitive to effects of increasing the noise level. This finding was based on data collapsed across syllables, and syllable differences for the latency measures occurred in the direction predicted by an upward spread of masking (i.e., comparatively reduced audibility of /ga/ predicting smaller amplitudes and slower latencies than were found for /ba/). However, voice onset times were also later for /ga/ than for /ba/, which may have contributed to the relatively slower N1 latencies for /ga/. However, pertinent to our experimental motivation, the results indicated differential effects of the visual cue on processing of the acoustic stimuli in quiet and in noise.

Modifications to a signal presented in noise can be related to a change in underlying neural activity that occurs during processing. For example, previous research (Cunningham et al., 2002) has demonstrated that modifications to the acoustic signal (increasing the stop gap duration and increasing the burst intensity) for syllables presented in noise led to a change in the neural substrates (an increase in the onset amplitude measures of the aggregate neural responses). In the present Experiment 2, we modified (reduced) the audibility of

auditory–visual stimuli that contained facial motion. The results showed that the morphology of the far-field scalp-recorded brain potential N1 in response to auditory–visual stimuli changed in relation to the quality of the acoustic environment, and that these changes were observable in quiet and in noise at a –9-dB SNR. In Experiment 3, we investigated changes in the underlying neural substrates, as measured by the N1 response that occurred with modifications to the type of visual cue—that is, with and without facial motion.

## Experiment 3

The goal of Experiment 3 was to identify changes in the processing of auditory speech syllables that are associated with different visual inputs. The response associated with visual-only speech was also recorded, as a control condition to verify that responses in the other presentation conditions reflected auditory processing.

### Method

#### *Participants*

A group of 16 new participants (eight female, eight male) were recruited for Experiment 3. The eligibility criteria for participation were the same as for Experiments 1 and 2. Again, no participant reported any difficulty hearing, and when tested on the in-laboratory, automated, calibrated hearing test, all had pure-tone thresholds of 30 dB HL or better at 500, 1000, 2000, 3000, 4000, and 6000 Hz. The participants represented a diversity in their ability to recognize visual-only words, with a range in performance of 9 %–58 % of words correct ($M$ = 28 %, $SD$ = 12 %) on a lipreading screening (Demorest & Bernstein, 1991).

#### *Stimuli*

The auditory, visual, and experimental (paired) stimuli were identical to those for Experiment 1, but only two acoustic environments were tested: quiet and –9 dB SNR.

#### *Experimental design*

The N1 potential, response time, sensitivity, and bias were recorded for 5 visual cues (AR, ASF, ADF, AVS, VO) × 2 acoustic environments (quiet, –9 dB SNR) × 2 syllables (/ba/, /ga/), with 100 trials per cell, for a total of 2,000 trials per participant. Ten experimental blocks (two per visual cue) were presented per session, each of approximately 10 min duration, for two sessions, with each block presenting stimuli in one visual cue condition. The order of presentation of the trials with different acoustic environments was randomized within each block. The order of the visual cue conditions was randomized, such that each participant received a different order, with the exception that the visual-only condition was always presented last, as in Experiment 1.

#### *Procedure and data analyses*

The procedures and data analyses were the same as were outlined for Experiment 2.

### Results

Behavioral responses were again recorded and analyzed. The effects of the visual cue condition replicated the findings from Experiment 1 in the quiet and –9-dB SNR acoustic environments (see Figs. 12 and 13). Due to the consistency of these findings with prior experiments, only the ERP results will be highlighted.

#### Event-related potentials

Waveform analyses were performed as for Experiment 2. The artifact rejection criteria (detailed in the Exp. 2 Method) led to the exclusion of 2.23 % of the trials in Experiment 3 (1 %–3 % of trials per condition, approximately equally distributed across conditions). Grand mean waveforms time-locked to the acoustic burst were extracted; the plots of the N1 response in all conditions across scalp locations are shown in Fig. 14. As expected, no measurable peaks were observed in the VO condition; that is, EEG activity time-locked to the release of the acoustic burst did not result in an N1 potential when only visual information was presented. This substantiates that the N1 peaks observed with the visual and auditory pairings reflect processing of the auditory stimulus. The AR and ASF conditions produced virtually identical results, so to simplify the analyses reported here, the AR condition was excluded from the ANOVAs, so that discussion will focus on the ASF, ADF, and AVS results. Plots of the N1 response to the AVS, ADF,

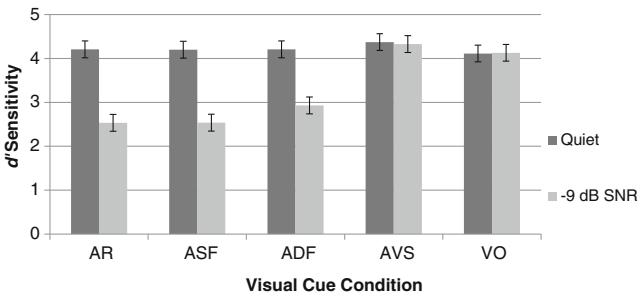


**Fig. 12** Experiment 3 mean $d'$ sensitivity as a function of visual cue and acoustic environment. Error bars represent standard errors of the means. AR, auditory–rectangle; ASF, auditory–static face; ADF, auditory–dynamic face; AVS, auditory–visual speech; VO, visual-only

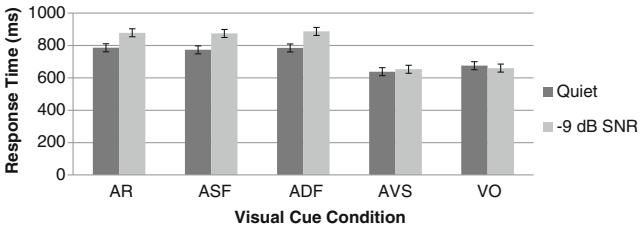


**Fig. 13** Experiment 3 mean response times as a function of visual cue and acoustic environment. Error bars represent standard errors of the means. AR, auditory–rectangle; ASF, auditory–static face; ADF, auditory–dynamic face; AVS, auditory–visual speech; VO, visual-only

and ASF presentation conditions, as recorded at the Cz electrode, are shown in Fig. 15.

*N1 amplitude* Again we found significant main effects of visual cue [$F(2, 30) = 14.91$, $p = .0001$, $f = .70$], acoustic environment [$F(1, 15) = 69.66$, $p < .0001$, $f = .80$], and channel [$F(16, 240) = 22.34$, $p < .0001$, $f = .40$]. N1 amplitudes were again reduced when visual speech gestures were available. For nonspeech facial cues, the amplitudes were on average –4.06 $\mu$V ($SD$ = 2.31 $\mu$V), with a static smiling face, and –4.21 $\mu$V ($SD$ = 1.97 $\mu$V), with a chewing face, and these two conditions did not differ from each other

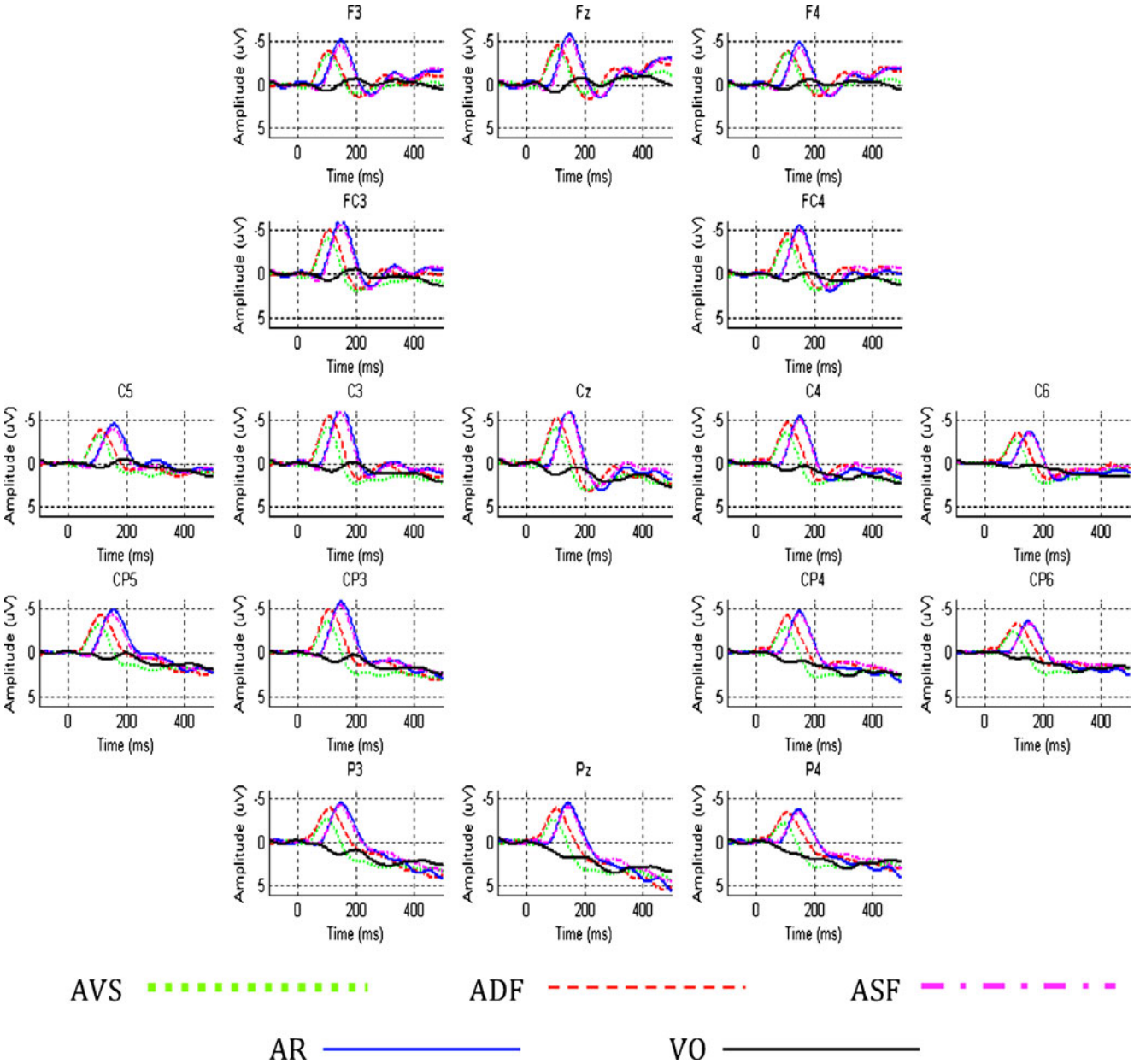


**Fig. 14** Experiment 3 N1 plots for /ba/in the quiet acoustic environment, time-locked to acoustic onset; negative is plotted up. AVS, auditory–visual speech; ADF, auditory–dynamic face; ASF, auditory–static face; AR, auditory–rectangle; VO, visual-only

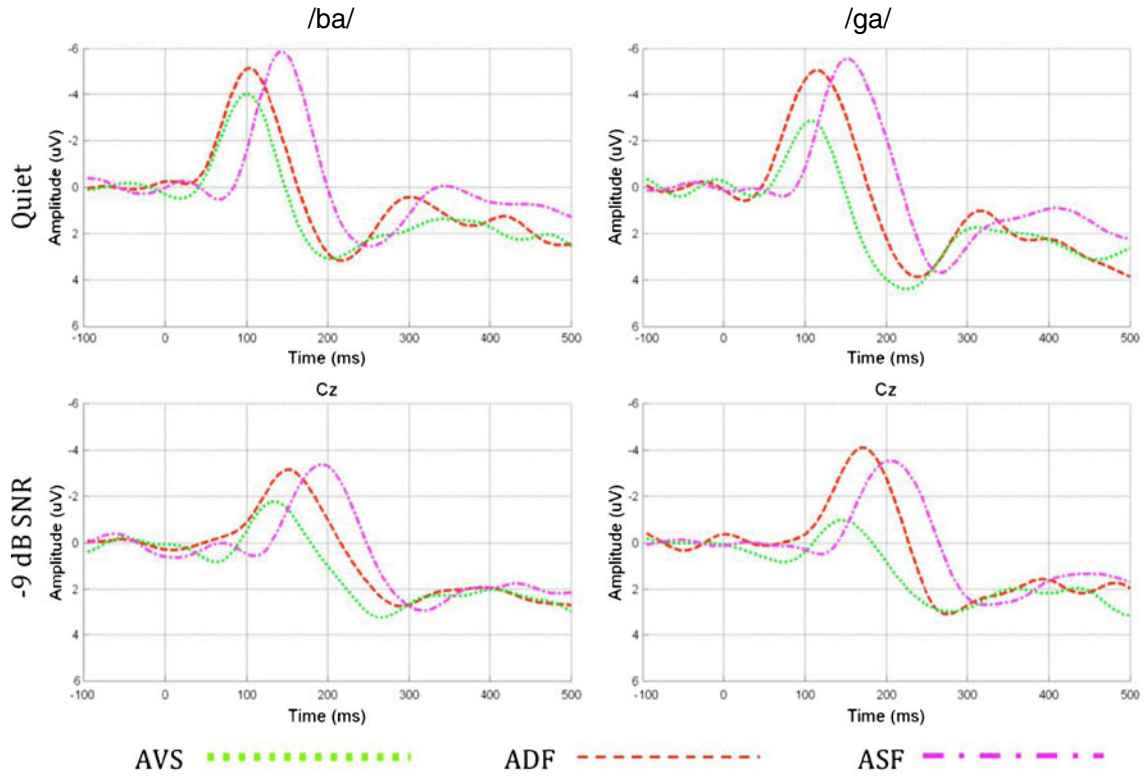


**Fig. 15** Experiment 3 N1 plots at the Cz electrode, time-locked to acoustic onset; negative is plotted up. AVS, auditory–visual speech; ADF, auditory–dynamic face; ASF, auditory–static face

[$F(1, 15) = 0.40$, $p = .5391$]. In contrast, the average amplitude in the AVS condition was –2.56 $\mu$V ($SD$ = 1.76 $\mu$V), which was smaller than those for both the static smiling face [$F(1, 15) = 17.44$, $p = .0008$] and the chewing face [$F(1, 15) = 18.89$, $p = .0006$]. Thus, amplitude was reduced when speech gestures were visible. As expected, amplitude was also reduced in noise ($M$ = –2.84 $\mu$V, $SD$ = 1.83 $\mu$V) as compared to quiet ($M$ = –4.38 $\mu$V, $SD$ = 2.19 $\mu$V). Overall, the amplitude was greatest over left central electrode locations—for instance, Cz, C3, and FC3. The effects of visual cue and acoustic environment described above were evident at each electrode location, but amplitudes were particularly reduced at frontal lateral electrode locations (e.g., F3, F4, C5, C6) for visual cues [$F(32, 480) = 7.40$, $p < .0001$], and at posterior electrode locations (e.g., Pz, P3, P4) for acoustic environments [$F(16, 240) = 11.24$, $p < .0001$]. To summarize the main points of the amplitude results, N1 peak amplitudes were significantly reduced with visual speech gestures, and also with noise.

*N1 latency* We found significant main effects of visual cue [$F(2, 30) = 199.00$, $p < .0001$, $f = .80$], acoustic environment [$F(1, 15) = 269.57$, $p < .0001$, $f = .80$], syllable [$F(1, 15) = 49.27$, $p < .0001$, $f = .80$], and channel [$F(16, 240) = 4.46$, $p = .0054$, $f = .40$]. Latencies were shortest in the AVS condition ($M$ = 126.83 ms, $SD$ = 29.51 ms), next shortest in the ADF condition ($M$ = 139.61 ms, $SD$ = 32.36 ms), and longest in the ASF condition ($M$ = 169.92 ms, $SD$ = 25.01 ms). Follow-up testing revealed the latencies in all three visual cue conditions to be significantly different from each other. Latencies were delayed by approximately 40 ms in noise as compared to quiet (quiet, $M$ = 125.14 ms, $SD$ = 27.06; noise, $M$ = 165.77 ms, $SD$ = 28.12). The latency of the N1 for /ba/ was approximately 10 ms faster ($M$ = 140.72 ms, $SD$ = 34.38 ms) than the latency for /ga/ ($M$ = 150.19 ms, $SD$ = 33.49). The fastest latencies occurred at right posterior electrode locations Pz and P4.

Significant Visual Cue × Acoustic Environment [$F(2, 30) = 10.25$, $p = .0006$], Syllable × Channel [$F(16, 240) = 3.94$, $p = .0008$], and Visual Cue × Acoustic Environment × Channel [$F(32, 480) = 3.23$, $p = .0005$] interactions also emerged. Within each visual cue type, latencies in quiet were significantly faster than those in noise. In quiet, the latency in the ASF condition was significantly slower than the latencies in both the ADF and AVS conditions, which did not differ from one another. The same was true in noise, except that we also found a significant difference between the latencies for ADF and AVS. Latencies were slower for /ga/ than for /ba/ across all channels, but this difference was minimal at frontal locations Fz, F3, and F4.

In sum, N1 latencies were slowest with a static smiling face, faster with a chewing face, and fastest with visible

speech gestures. Latencies were also faster in quiet than in noise and for /ba/ than for /ga/. In quiet, there was no significant difference in latencies for the two dynamic visual cues. However, in noise, latencies were significantly faster with visual speech gestures than with a chewing gesture.

### Discussion

Previous arguments in the literature have suggested that the magnitude of the neural response can be used to determine whether the relationship between visual and auditory speech involves an alerting mechanism (Baier, Kleinschmidt & Müller 2006; Besle et al., 2004). Baier et al. argued that neural activity as measured in an fMRI procedure increased when there was a learned association between the auditory and visual stimuli, and decreased when the stimuli were not associated. Besle et al. argued that the auditory–visual facilitation associated with an alerting mechanism should result in an N1 amplitude increase. Therefore, they interpreted their findings of decreased N1 amplitudes with auditory–visual speech as evidence against an alerting mechanism.

The results from the present experiment suggest an alternative interpretation. In Experiment 3, a decrease in N1 amplitudes occurred only when both the auditory and visual stimuli presented linguistic information (the AVS condition). However, faster N1 latencies occurred with either the chewing gesture or the visible speech gesture—that is, with dynamic facial cues. Therefore, we hypothesized that the speed of the neural response, not the magnitude, may be related to a learned association between facial movement and acoustic speech, such that facial motion serves as an alerting mechanism for acoustic speech. In complex environments, humans have difficulty suppressing the visual information in a bimodal auditory–visual stimulus, demonstrating visual dominance (Sinnett, Spence, & Soto-Faraco, 2007). Due to the high ecological significance of speech, seeing facial motion may prime, or alert, the auditory system to be ready to respond. This alert would be activated prior to determining whether the facial motion is actually informative about temporally co-occurring auditory stimuli. This would explain why both uninformative chewing and informative speech gestures speeded up the N1 response to an auditory syllable. The meaningfulness of the visual cue appears to influence the magnitude of the neural response rather than its timing, since N1 amplitudes were decreased only with a linguistically informative congruent visual gesture. A possible alternative interpretation may be that the visible speech gesture introduced another channel of information, leading to the reduction in amplitude. Paulmann, Jessen, and Kotz (2009) proposed that the magnitude of an ERP response is related to how many channels present information. In discussing results for the P200 and N300 (in a gender/talker identification task), they reported reduced amplitudes for three channels of information (a visual static picture of a face conveying an emotional valence, an auditory semantically meaningful sentence, and congruent prosody) as compared to two channels of information (semantic content removed from the multimodal stimulus), which in turn was reduced as compared to responses to the emotional valence of the face alone. Combined with the results from the present experiment, meaningfulness, or linguistic significance, appears to be a factor requiring further investigation for its role in multisensory processing.

## General discussion

Perhaps the most important of the new findings obtained in the experiments reported here is the observed dissociation between the amplitude and latency of N1 responses to spoken syllables paired with different kinds of visual cues. Latency was sensitive to any kind of facial motion (at least, to both of the two different kinds of facial motion used in these experiments) and not to its meaningfulness, while amplitude was sensitive to the meaningfulness of the facial motion. Latency also appeared to be more sensitive to effects of audibility, evidenced in syllable differences, with slower latencies for /ga/, which would be more susceptible to upward spread of masking than would /ba/. The present study shows that auditory N1 morphology is affected by visual cues in noisy environments, whereas previous studies had examined these effects only in quiet (e.g., Pilling, 2009; van Wassenhove et al., 2005). Behavioral measures of response sensitivity and response time were improved with auditory–visual speech in the most challenging SNR tested in this experiment, as compared to responses to acoustic speech paired with three types of nonspeech visual cues. Linguistically familiar gestures, when presented with acoustic speech, influenced auditory processing by speeding up the N1 latency and decreasing N1 amplitude. Processing of auditory–visual speech, as measured by the N1 component, was distinct from processing of acoustic speech with the nonspeech visual cues tested in this study. However, processing was also influenced by a nonspeech dynamic facial stimulus (chewing), in that faster N1 latencies also occurred with this visual cue in a quiet listening condition. In noisy acoustic environments, N1 latencies were faster with visible speech gestures than with chewing gestures.

### Quality of the acoustic environment

Reduced audibility influenced the morphology of the N1 response to auditory–visual signals in a manner similar to that demonstrated by Martin et al. (1999; Martin et al., 1997) for auditory-only speech. Thus, auditory–visual processing was influenced by the quality of the acoustic

environment—that is, slower and with reduced amplitude in noise. The results from the present experiments revealed an advantage for auditory–visual speech in noise as compared to other experimental stimuli, in both behavioral measures of task performance and the associated neural correlates. Comparisons were made at each SNR level, however, and when speech was presented in a noisy environment, accurate auditory–visual speech perception may have required less effort than is required with auditory-only speech (Fraser, Gagné, Alepins, & Dubois, 2010).

### Dynamic facial visual cues

In the present experiments, latency facilitation occurred with both types of dynamic facial cues, but amplitude decrease occurred only with auditory–visual speech. Van Wassenhove et al. (2005) theorized that the latency reduction with auditory–visual speech is related to predictive coding of sensory information, as well as to the saliency and redundancy of visual information. They developed the hypothesis that visible articulation that naturally precedes acoustic speech activates an internal prediction that is compared to later incoming acoustic information. Faster latencies would be the result of an earlier match between the prediction and new information. The data reported in the present study support this theory, as latencies were faster with visible speech gestures. However, the results of the present study also suggest an extension of this theory. As was discussed previously in relation to Experiment 3, faster N1 latencies occurred with chewing gestures than with static facial images. We hypothesize that a precue or a learned association between facial motion in general and acoustic speech may be involved in priming the auditory system to respond, thus leading to faster latencies. As van Wassenhove et al. would argue, the increased saliency or redundancy of the visual speech gesture could account for the additional latency decrease demonstrated with noise for the present Experiment 3. However, only visual information with linguistic content—speech gestures—would lead to a match between visual and auditory information and decreased amplitudes from less effortful processing. A chewing gesture could be considered a mismatch between visual and auditory information, resulting in greater N1 amplitudes for chewing gestures than for speech gestures.

Ponton, Bernstein, and Auer (2009) discussed the possibility of both integrative and modulatory effects of visual signals in auditory–visual processing. They suggested that different effects arise from parallel processing of features related to the visual speech (integrative) and of features not related to the visual speech (modulatory). The present results are consistent with this distinction. The decreased amplitudes observed with auditory–visual speech could be interpreted as an integrative effect of processing both acoustic speech and articulatory gesture. The faster latencies observed with dynamic facial motion could be interpreted as a modulatory effect of visual information not related to the speech gesture.

### Considerations for the ERP results

*Multiple components* The N1 is probably not a unitary response, but may instead consist of several subcomponents (Näätänen & Picton, 1987), which may be differentially affected by visible speech gestures, the temporal relation between visible speech gestures and acoustic speech, and multisensory interaction/integration effects. For example, Puce, Epling, Thompson, and Carrick (2007) have discussed two components occurring within the N1 response time window, one of which they identified as being sensitive to auditory information (the N140), and the other to visual motion (the N170). The timing of the N1 peak in the averaged waveforms in Experiment 3 (average latencies across electrodes at 170 ms for a static face, 140 ms for nonspeech facial motion, and 127 ms for auditory–visual speech), together with its susceptibility to decreased audibility (i.e., its absence in the loudest noise and in the visual-only condition) suggests that it consisted primarily of a response to the auditory stimulus rather than to the motion itself. However, it is possible that the decrease in N1 amplitude in the auditory–visual speech condition was due to an increase in the contribution from the slightly later neurons responding to visual information, since spreading the activity over a wider time range would lead to a decrease in amplitude in the averaged waveforms.

Auditory–visual stimuli have been found to influence other ERP components both earlier and later than the N1. Auditory brainstem responses to auditory–visual speech occurring within approximately 30 ms have been shown to have slower latencies and smaller amplitudes than those for auditory-only speech (Musacchia, Sams, Nicol, & Kraus, 2006). Changes over visual areas during a 40- to 90-ms time window have been demonstrated for bimodal relative to unimodal nonspeech object stimuli (Giard & Peronnet, 1999). The P50 component has been reported to be altered when both auditory and visual nonspeech stimuli are attended (Talsma, Doty, & Woldorff, 2007). Earlier-occurring components may also influence the morphology of later components, with auditory–visual effects perhaps culminating during the N1 time window. One effect that earlier components can have is to cause a desynchronization of neural activity, resulting in degraded morphology of later-occurring peaks, which could have contributed to the apparent absence of a P2 peak following the N1 in our waveforms at all but the most frontal sites.

*Neural oscillatory phase* It is possible that the observed differences in N1 latency and/or amplitude could be due to

shifts in the phase of the neuronal oscillations that make up the EEG. Recent evidence has demonstrated that the phase of alpha activity at the onset of a visual stimulus affects aspects of the averaged ERP response to those stimuli (Mathewson, Gratton, Fabiani, Beck, & Ro, 2009). It has also been hypothesized that a predictable temporal lag between auditory and visual stimuli modifies the phase of neural oscillations (Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008). In the experiments reported here, the time lags between visual motion onset and the acoustic burst were equated for the chewing and speaking faces. If detecting facial motion leads to a shift in the EEG phase, that could contribute to the earlier N1 latency observed in both the speaking and chewing face conditions. Since no conditions with nonfacial visual motion were included, we cannot conclude that only facial motion leads to such an effect, but our hypothesis is that individuals with normal hearing and normal vision are sensitive to the temporal relations between facial motion and acoustic speech, due to a lifetime of experience associating movements of the face with speech.

*Increase of inhibitory activity* Evidence from human fMRI studies (Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003) and monkey local-field potential studies (Ghazanfar, Maier, Hoffman, & Logothetis, 2005) has suggested that bimodal stimuli produce both increases in neural activity in some brain regions and decreases in other regions, as compared to unimodal stimuli. It is possible that an increase in activity in some regions may reflect increased inhibitory activity, and increased inhibition of N1 generators is one possible mechanism for the decrease in N1 amplitudes seen in Experiments 2 and 3 for auditory–visual speech only.

*Relation to buttonpress response* The behavioral measures tapped into processing at a different time than the N1 ERP component. It took an additional 500–800 ms after the N1 peak for the perceiver to make a decision and press a button. The difference in the timing between the two kinds of measures provides part of the explanation for why response times were faster for /ga/ but N1 latencies were faster for /ba/. The faster neural processing of /ba/ may be related to (1) earlier and/or more salient visual cues for /ba/ than for /ga/, and/or (2) the greater audibility of /ba/ because it is less masked by broadband white noise than /ga/ is. In contrast, faster response times for /ga/ may be related to a bias to wait for a visual cue for /ba/, since it normally has clear visual cues, and not to wait with /ga/, because it normally has weaker visual cues. This interpretation was supported by signal detection analyses that revealed that the buttonpress responses were affected by overall response biases as well as by sensitivity to the properties of the stimuli.

The N1 latency measure proved to be more sensitive to the effects of visual motion on speech processing than did the behavioral responses. The fact that N1 latencies were faster for both chewing and speech facial gestures than for a static facial image suggests that facial motion in general may serve to alert the speech-processing system that speech sounds are imminent. This was not revealed by the behavioral response times, which were faster only when speech was accompanied by a speaking face, again suggesting that the two kinds of measures tapped into different subsets of the processes involved in recognizing the auditory–visual stimuli. Physiological measures allow for an evaluation of processing without the ceiling or floor effects that can occur with behavioral measures. For instance, behavioral measures did not reveal any difference between auditory–visual speech and visual-only speech, presumably because the visual cues were sufficient to make the decision required by the task. In contrast, an N1 response was recorded for auditory–visual speech, but not for visual-only speech. The strength of behavioral measures is in determining how processing is related to the final percept and response, for example, in how well a person may be able to communicate. The evaluation of both early processing in physiological measures and later processing with a behavioral action or response would be helpful for determining at which point an obstacle or deficit in perception occurs, which could be crucial to identifying an appropriate treatment methodology.

## Implications for auditory–visual speech perception development

Sensory and language experience early in life impacts neurological development and processing (Bavelier et al., 2001; Capek et al., 2009; Neville et al., 1998; Ponton & Eggermont, 2001; Sharma, Dorman, & Spahr, 2002): If auditory stimulation is not received within a sensitive period, sensory processing follows a different developmental path, as measured by the P1 auditory evoked potential (see Sharma, Nash, & Dorman, 2009, for a review). Children with hearing loss have different experience not only with auditory sensory stimulation, but also with associating auditory and visual sensory stimulation. It is unknown how auditory—and thus auditory–visual integration—deprivation would influence the development of multisensory neural processing.

Even with early identification and treatment, children with hearing loss will lack experience associating auditory and visual stimuli very early in life. Infants with normal hearing demonstrate sensitivity to auditory–visual integration processes within the first few months of life (Burnham & Dodd, 2004; Lewkowicz, 2000; Patterson & Werker, 2003). Children with hearing loss who had received a cochlear implant after 30 months of age did not consistently report fused auditory–visual percepts in a McGurk stimuli

task (Schorr, Fox, van Wassenhove, & Knudsen, 2005). This suggests that auditory–visual integration deprivation during a critical period affects the ability to associate these multimodal stimuli, as measured by a behavioral response. In contrast to their normal-hearing peers, children with cochlear implants relied more upon the visual information when perceiving conflicting auditory–visual stimuli (Schorr et al., 2005). It is probably not the sensitivity to visual speech cues that is affected by a unimodal auditory sensory impairment, but rather the ability to associate and integrate information from both auditory and visual sensory modalities, which may develop differently or not at all. Individuals who lack this early multisensory experience might not demonstrate the decrease in N1 latencies shown by the normally hearing participants in our studies when viewing a facial chewing motion.

For children with hearing loss who utilize aural communication, the development of auditory skills with a sensory device has obvious importance. However, the value of associating auditory and visual information, particularly in poor-quality acoustic environments, should not be overlooked. Adults with hearing loss demonstrate changes in auditory processing in better acoustic environments than do adults with normal hearing (Oates, Kurtzberg, & Stapells, 2002; Whiting, Martin, & Stapells, 1998). Correspondingly, visual information may affect auditory processing in better acoustic conditions for individuals with hearing loss than for individuals with normal hearing.

### Limitations

The scope of the present experiments did not allow for an investigation of all possible characteristics of the visible speech gesture. Dynamic nonfacial stimuli were not tested. The stimuli were selected according to the hypothesis that facial images would influence processing to a greater degree than would nonfacial images. The results indicated that the relevant facial feature is motion. Syllable differences in the timing measures also suggest that future research should employ spectrally shaped maskers to investigate the processing of syllables with equal degrees of masking/audibility. Because the white noise was exactly the same for both syllables, it is clear that the differences are related to the syllable and not to the processing of different noise sources. Future studies may examine differential effects of spectrally shaped, high-pass, low-pass, or other types of noise. The present findings relate only to competition from a broadband white noise. Continued investigation will be required to elucidate the roles of audibility and visibility in understanding syllable differences. The present experiments have demonstrated the effect of type of visual cue within a particular speech syllable, with decreased N1 amplitudes being associated with meaningful speech gestures. Further research should investigate the extension of these results to a larger, more diverse set of syllables and also to more complex speech.

### Conclusions

The present experiments are unique in examining, in combination, the effects of reduced audibility and the influence of the features of visual cues on the N1 ERP corresponding to auditory processing in noise. The results indicated that the N1 response measured in these experiments was an auditory response, because it was not generated in response to visual-only stimuli and was influenced by reduced audibility in a manner similar to auditory-only speech (Martin et al., 1999; Martin et al., 1997). The processing of acoustic speech with reduced audibility was associated with slower latencies and decreased amplitudes, regardless of the type of visual cue presented. However, with visible speech gestures, latencies did not differ between the 0- and –9-dB SNR environments in Experiment 2. Thus, evidence from the present experiments suggests that concurrently seeing visible speech gestures while listening to acoustic speech in noise may make auditory processing less sensitive to an increase in noise intensity.

The results from the present study provide partial support for the hypothesis that auditory–visual facilitation is unique to the properties of the speech gesture. The neural correlates of auditory–visual speech perception differed from the measures in response to the other types of experimental stimuli tested. No difference was seen between a static geometric shape and a static facial image, indicating that seeing a face in itself does not influence auditory processing to a greater degree than does seeing a geometric shape. Two types of facial motion were associated with faster N1 latencies. As we noted above, a dynamic nonfacial stimulus was not included among the presentation conditions. However, the results revealed that one aspect of the influence of visible speech gestures on auditory speech (faster N1 latency) may not be specific to speech gestures, but instead may generalize to all facial motion. A distinct effect of visible speech gestures on auditory processing was seen in decreased N1 amplitudes.

Our empirical data support the hypothesis that auditory–visual perception of speech in noise is a distinct process related to the special characteristics of visible speech gestures; however, speech gestures are very complex, and the present experimental stimuli could not be used to evaluate all possible features of the visual speech gesture. Specifically, the influence of visual motion must continue to be investigated. In addition, hypotheses regarding a priming or alerting mechanism due to a reliable temporal lag between visual and auditory stimuli should be tested further. The impact of auditory

deprivation on the future ability to integrate auditory and visual cues needs to be considered in order to maximize the auditory–visual benefit for clinical populations—for instance, individuals with cochlear implants.

The interaction of auditory and visual cues influences speech perception processes. Auditory–visual facilitation was seen in both behavioral measures and the neural correlates of speech perception, particularly in noise. In adults with normal hearing and vision, the neural correlates of sensitivity to integrating facial motion, and especially facial speech motion, with acoustic speech may be measured in both quiet and noise. People are sensitive to associations between auditory and visual stimuli beginning at a very young age (see, e.g., Burnham & Dodd, 2004; Lewkowicz, 2000; Patterson & Werker, 2003). Auditory (re)habilitation treatment methodologies for all age groups must take into account the speech perception benefit that may be received from learning to associate and integrate auditory and visual cues.

**Author note** J.L.G. is currently affiliated with the University of Wisconsin, Stevens Point. This research was conducted as part of the dissertation requirements for the first author and was supported in part by funds from the University of Illinois Research Board. Portions of this research have been presented at meetings of the American Speech–Language–Hearing Association and the American Auditory Society.

## References

Ahlfors, S. P., Simpson, G. V., Dale, A. M., Belliveau, J. W., Liu, A. K., Korvenoja, A., & Ilmoniemi, R. J. (1999). Spatiotemporal activity of a cortical network for processing visual motion revealed by MEG and fMRI. *Journal of Neurophysiology, 82,* 2545–2555.

Baier, B., Kleinschmidt, A., & Müller, N. G. (2006). Cross-modal processing in early visual and auditory cortices depends on expected statistical relationship of multisensory information. *Journal of Neuroscience, 26,* 12260–12265.

Bavelier, D., Brozinsky, C., Tomann, A., Mitchell, T., Neville, H., & Liu, G. (2001). Impact of early deafness and early exposure to sign language on the cerebral organization for motion processing. *Journal of Neuroscience, 21,* 8931–8942.

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience, 8,* 551–565.

Bergeson, T. R., Pisoni, D. B., & Davis, R. A. O. (2003). A longitudinal study of audiovisual speech perception by children with hearing loss who have cochlear implants [Monograph]. *Volta Review, 103*(4), 347–370.

Bernstein, L. E., Auer, E. T., Jr., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication, 44,* 5–18.

Bernstein, L. E., Demorest, M. E., Coulter, D. C., & O'Connell, M. P. (1991). Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects. *Journal of the Acoustical Society of America, 90,* 2971–2984.

Bernstein, L. E., Demorest, M. E., & Eberhardt, S. P. (1994). A computational approach to analyzing sentential speech perception: Phoneme-to-phoneme stimulus–response alignment. *Journal of the Acoustical Society of America, 95,* 3617–3622.

Bernstein, L. E., & Eberhardt, S. P. (1986). Johns Hopkins Lipreading Corpus I–II: Disc I [Laser Video Disc). Baltimore, MD: Johns Hopkins University.

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience, 20,* 2225–2234.

Burnham, D., & Dodd, B. (2004). Auditory–visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology, 45,* 204–220.

Capek, C. M., Grossi, G., Newman, A. J., McBurney, S. L., Corina, D., Roeder, B., & Neville, H. J. (2009). Brain systems mediating semantic and syntactic processing in deaf native signers: Biological invariance and modality specificity. *Proceedings of the National Academy of Sciences, 106,* 8784–8789. doi:10.1073/pnas.0809609106

Chatrian, G. E., Lettich, E., & Nelson, P. L. (1985). Ten percent electrode system for topographic studies of spontaneous and evoked EEG activity. *The American Journal of EEG Technology, 25,* 83–92.

Chatrian, G. E., Lettich, E., & Nelson, P. L. (1988). Modified nomenclature for the "10%" electrode system. *Journal of Clinical Neurophysiology, 5,* 183–186.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cunningham, J., Nicol, T., King, C., Zecker, S. G., & Kraus, N. (2002). Effects of noise and cue enhancement on neural responses to speech in auditory midbrain, thalamus and cortex. *Hearing Research, 169,* 97–111.

Davis, H., & Silverman, S. R. (1978). *Hearing and deafness* (4th ed.). New York: Holt, Rinehart & Winston.

Demorest, M. E., & Bernstein, L. E. (1991). Computational explorations of speechreading. *Journal of the Academy of Rehabilitative Audiology, 24,* 97–111.

Demorest, M. E., & Bernstein, L. E. (1992). Sources of variability in speechreading sentences: A generalizability analysis. *Journal of Speech and Hearing Research, 35,* 876–891.

Eberhardt, S. P., Bernstein, L. E., Demorest, M. E., & Goldstein, M. H., Jr. (1990). Speechreading sentences with single-channel vibrotactile presentation of voice fundamental frequency. *Journal of the Acoustical Society of America, 88,* 1274–1285.

Fraser, S., Gagné, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research, 53,* 18–33.

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience, 25,* 5004–5012.

Giard, M. H., & Peronnet, F. (1999). Auditory–visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience, 11,* 473–490.

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory–visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America, 104,* 2438–2450.

Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology, 55,* 468–484.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research, 46,* 390–404.

Kaplan-Neeman, R., Kishon-Rabin, L., Henkin, Y., & Muchnik, C. (2006). Identification of syllables in noise: Electrophysiological

and behavioral correlates. *Journal of the Acoustical Society of America, 120,* 926–933.

Kim, J., & Davis, C. (2004). Investigating the audio-visual speech detection advantage. *Speech Communication, 22,* 19–30.

Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics, 65,* 536–552.

Lewkowicz, D. J. (2000). The development of intersensory temporal perception: An epigenetic systems/limitations view. *Psychological Bulletin, 126,* 281–308.

Martin, B. A., Kurtzberg, D., & Stapells, D. R. (1999). The effects of decreased audibility produced by high-pass noise masking on N1 and the mismatch negativity to speech sounds /ba/ and /da/. *Journal of Speech, Language, and Hearing Research, 42,* 271–286.

Martin, B. A., Sigal, A., Kurtzberg, D., & Stapells, D. R. (1997). The effects of decreased audibility produced by high-pass noise masking on cortical event-related potentials to speech sounds /ba/ and /da/. *Journal of the Acoustical Society of America, 101,* 1585–1599.

Mathewson, K. E., Gratton, G., Fabiani, M., Beck, D. M., & Ro, T. (2009). To see or not to see: Prestimulus alpha phase predicts visual awareness. *Journal of Neuroscience, 29,* 2725–2732.

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

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27,* 338–352.

Musacchia, G., Sams, M., Nicol, T., & Kraus, N. (2006). Seeing speech affects acoustic information processing in the human brainstem. *Experimental Brain Research, 168,* 1–10.

Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology, 24,* 375–425.

Neff, D. L., & Dethlefs, T. M. (1995). Individual differences in simultaneous masking with random-frequency, multicomponent maskers. *Journal of the Acoustical Society of America, 98,* 125–134.

Neville, H. J., Bavelier, D., Corina, D., Rauschecker, J., Karni, A., Lalwani, A., & Turner, R. (1998). Cerebral organization for language in deaf and hearing subjects: Biological constraints and effects of experience. *Proceedings of the National Academy of Sciences, 95,* 922–929.

Oates, P. A., Kurtzberg, D., & Stapells, D. R. (2002). Effects of sensorineural hearing loss on cortical event-related potential and behavioral measures of speech-sound processing. *Ear and Hearing, 23,* 399–415.

Otta, E., Lira, B. B. P., Delevati, N. M., Cesar, O. P., & Pires, C. S. G. (1994). The effect of smiling and of head tilting on person perception. *Journal of Psychology, 128,* 323–331.

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

Paulmann, S., Jessen, S., & Kotz, S. A. (2009). Investigating the multimodal nature of human communication: Insights from ERPs. *Journal of Psychophysiology, 23,* 63–76.

Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal of Speech, Language, and Hearing Research, 52,* 1073–1081.

Ponton, C. W., Bernstein, L. E., & Auer, E. T., Jr. (2009). Mismatch negativity with visual-only and audiovisual speech. *Brain Topography, 21,* 207–215.

Ponton, C. W., & Eggermont, J. J. (2001). Of kittens and kids: Altered cortical maturation following profound deafness and cochlear implant use. *Audiology & Neuro-Otology, 6,* 363–380.

Puce, A., Epling, J. A., Thompson, J. C., & Carrick, O. K. (2007). Neural responses elicited to face motion and vocalization pairings. *Neuropsychologia, 45,* 93–106. doi:10.1016/j.neuropsychologia.2006.04.017

Schorr, E. A., Fox, N. A., van Wassenhove, V., & Knudsen, E. I. (2005). Auditory–visual fusion in speech perception in children with cochlear implants. *Proceedings of the National Academy of Sciences, 102,* 18748–18750. doi:10.1073/pnas.0508862102

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences, 12,* 106–113.

Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio–visual interactions in speech identification. *Cognition, 93,* B69–B78.

Sharma, A., Dorman, M. F., & Spahr, A. J. (2002). A sensitive period for the development of the central auditory system in children with cochlear implants: Implications for age of implantation. *Ear and Hearing, 23,* 532–539.

Sharma, A., Nash, A. A., & Dorman, M. (2009). Cortical development, plasticity and re-organization in children with cochlear implants. *Journal of Communication Disorders, 42,* 272–279.

Sinnett, S., Spence, C., & Soto-Faraco, S. (2007). Visual dominance and attention: The Colavita effect revisited. *Perception & Psychophysics, 69,* 673–686. doi:10.3758/BF03193770

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

Talsma, D., Doty, T. J., & Woldorff, M. G. (2007). Selective attention and audiovisual integration: Is attending to both modalities a prerequisite for early integration? *Cerebral Cortex, 17,* 679–690.

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. doi:10.1073/pnas.0408949102

Weihing, J., Daniels, S., & Musiek, F. E. (2009). The effect of visual and audiovisual competition on the auditory N1-P2 evoked potential. *Journal of the American Academy of Audiology, 20,* 569–581.

Whiting, K. A., Martin, B. A., & Stapells, D. R. (1998). The effects of broadband noise masking on cortical event-related potentials to speech sounds /ba/ and /da/. *Ear and Hearing, 19,* 218–231.

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J., & McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex, 13,* 1034–1043.