JSLHR

Research Article

Spatial Frequency Requirements and Gaze Strategy in Visual-Only and Audiovisual Speech Perception

Amanda H. Wilson,[a,b] Agnès Alsius,[a] Martin Paré,[b] and Kevin G. Munhall[a,b]

Purpose: The aim of this article is to examine the effects of visual image degradation on performance and gaze behavior in audiovisual and visual-only speech perception tasks.
Method: We presented vowel–consonant–vowel utterances visually filtered at a range of frequencies in visual-only, audiovisual congruent, and audiovisual incongruent conditions (Experiment 1; *N* = 66). In Experiment 2 (*N* = 20), participants performed a visual-only speech perception task and in Experiment 3 (*N* = 20) an audiovisual task while having their gaze behavior monitored using eye-tracking equipment.
Results: In the visual-only condition, increasing image resolution led to monotonic increases in performance, and proficient speechreaders were more affected by the removal of high spatial information than were poor speechreaders. The McGurk effect also increased with increasing visual resolution, although it was less affected by the removal of high-frequency information. Observers tended to fixate on the mouth more in visual-only perception, but gaze toward the mouth did not correlate with accuracy of silent speechreading or the magnitude of the McGurk effect.
Conclusions: The results suggest that individual differences in silent speechreading and the McGurk effect are not related. This conclusion is supported by differential influences of high-resolution visual information on the 2 tasks and differences in the pattern of gaze.

Understanding speech is a vital part of our daily life for social, emotional, and informational purposes. Although audition plays an important role in perceiving speech information, vision is also an integral part of this process. People with profound hearing impairment can use speechreading as the primary source for understanding speech, and some talented speechreaders are even capable of very high accuracy in the complete absence of any auditory signal (Bernstein, Demorest, & Tucker, 2000). People with normal hearing also experience a benefit from vision when the auditory signal is degraded. This can be seen in experiments demonstrating that the intelligibility of speech presented in noisy environments (speech-in-noise tasks) can be enhanced by presenting visual cues of the corresponding articulation (Cotton, 1935; Erber, 1969; Middelweerd & Plomp, 1987; Neely, 1956; O'Neill, 1954; Sumby & Pollack, 1954). The influence of vision on speech perception can also be observed when the acoustics are not degraded, such as in the phenomenon known as the *McGurk effect* (McGurk & MacDonald, 1976). In this effect, observers presented with conflicting auditory and visual information (e.g., an auditory /aba/ and a simultaneous visual /aga/), often perceive /aga/ or a sound that is entirely different from either stimuli (e.g., /ada/ or /atha/).

These examples highlight the importance of the visual cues to speech processing, but exactly which characteristics of the visual image influence speech perception remain to be determined. Research conducted on the resolution requirements for the visual cues in speech perception have suggested that, in general, fine facial detail is not critical for audiovisual speech enhancement to be observed (Jordan & Sergeant, 1998, 2000; MacDonald, Andersen, & Bachmann, 2000; Munhall, Kroos, Jozan, & Vatikiotis-Bateson, 2004; Neely, 1956; Rosenblum, Johnson, & Saldana, 1996).

Jordan and Sergeant (2000), for instance, manipulated the distance between the perceiver and the talker and showed that audiovisual speech recognition was still resilient when the talker was viewed from 20 m (Jordan & Sergeant, 2000; Neely, 1956; Small & Infante, 1988; however, see Erber, 1971). Another finding that suggests a crude visual

[a]Psychology Department, Queen's University, Kingston, Ontario, Canada
[b]Centre for Neuroscience Studies, Queen's University, Kingston, Ontario, Canada
Correspondence to Agnès Alsius: aalsius@gmail.com

Editor: Jody Kreiman
Associate Editor: Ewa Jacewicz
Received March 4, 2015
Revision received September 16, 2015
Accepted October 7, 2015
DOI: 10.1044/2016_JSLHR-S-15-0092

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

signal may be sufficient for audiovisual speech integration to occur is that of Paré, Richler, ten Hove, and Munhall (2003). In one of their experiments, participants fixated on different points of the talker's face (mouth, eyes, hairline) while being presented with McGurk stimuli. There was no significant difference in the McGurk effect whether fixations were on the mouth or eyes, and there was only a slight decline when participants fixated on the hairline (from 77% at the mouth to 65% at the hairline). In a second experiment, the viewing angle was manipulated more substantially by having participants fixate at 0°, 20°, 40°, and 60° from the talker's mouth. Even at viewing angles of 20° and 40°, participants reported significant illusory percepts, although the effect was reduced from that which occurred when participants fixated on the talker's mouth. Consistent with these results, a recent study has shown that the intelligibility of speech presented in noise is not affected when participants stare 10° from the mouth, and it is only slightly diminished at 15° (Yi, Wong, & Eizenman, 2013). Altogether, these findings suggest that high-acuity foveal vision may not be necessary for much of the visual gain in speech tasks.

Similar trends are found when the spatial resolution of the image is directly degraded using spatial quantization and spatial frequency filtering. The *spatial quantization technique,* which reduces the number of pixels in the image, has been used in studies testing the McGurk effect (MacDonald et al., 2000) as well as studies testing silent consonant perception (Campbell & Massaro, 1997). These have found that speechreading performance and the illusory effect decrease monotonically as the coarseness of the spatial quantization increases; however, both audiovisual and speechreading performance appear to be quite robust to stimulus degradation resulting from such artificial visual manipulation.

The *spatial frequency* of an image is the rate of change of contrast per spatial unit, and in general, higher frequencies represent higher detailed information. By filtering out certain frequencies, it is possible to determine what spatial frequencies are necessary for visual speech perception. Munhall et al. (2004) degraded the image of a talker by applying low- and band-pass filters and showed that perception of speech in noise can be enhanced to a degree equivalent to that produced by unfiltered images, even by an image that contains only very low spatial frequencies (i.e., 7.3 cycles/face [c/f]). This suggests that high spatial frequency information is not needed for speech perception.

One question that has not been addressed, however, is whether spatial frequency filtering affects visual speech perception in the same way that it affects audiovisual speech perception. There are reasons to believe that visual and audiovisual speech perception might use distinct visual processes. First, the successful integration of audiovisual information (as evidenced by the McGurk effect) is not related to speechreading ability (Cienkowski & Carney, 2002; however, see Strand, Cooperman, Rowe, & Simenstad, 2014). However, the relationship between silent speechreading performance and audiovisual speech perception in noise may show a different pattern than exhibited by the McGurk effect; with sentence material, significant positive correlations with speechreading have been reported (MacLeod & Summerfield, 1987). Second, research has shown that speechreading produces different gaze patterns than audiovisual speech (Lansing & McConkie, 2003), again providing evidence for the existence of apparently distinct visual mechanisms. The present study examines the point at which high spatial frequency information becomes redundant (i.e., does not add any significant benefit in performance) for an audiovisual McGurk task and a silent speechreading task.

Another question that has not been addressed is whether degrading the visual image affects eye movements during speech perception. Only a handful of studies have monitored how perceivers gather visual speech information by tracking eye movements (e.g., Buchan & Munhall, 2012; Buchan, Paré, & Munhall, 2007, 2008; Everdell, Marsh, Yurick, Munhall, & Paré, 2007; Lansing & McConkie, 1999, 2003; Paré et al., 2003; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998; Yi et al., 2013). One study, examining fixations during a speech-in-noise task using extended monologues (Vatikiotis-Bateson et al., 1998), found that people tend to alternate fixations between the eyes and the mouth of the talker. The proportion of mouth fixations increased from approximately 35% with no noise present to approximately 55% at the highest noise level with a corresponding decrease in eye fixations (see also Buchan, Paré, & Munhall, 2005, and Yi et al., 2013, for consistent results showing an increased number and longer fixations on the nose and mouth in the presence of noise). Lansing and McConkie (2003) carried out another study monitoring gaze, this time with a sentence identification task in two modalities: audiovisual and visual-only. They found that participants tended to look toward the mouth more during the visual-only condition, and they suggested that this was due to the difficulty of the speech perception task. They propose that people fixate on the mouth when they require more visual speech information. However, it was surprising that they did not find a correlation between speechreading accuracy and the proportion of gaze time directed at the mouth. In a study monitoring eye gaze during a McGurk task, Paré et al. (2003) again found that gaze was often directed to the mouth and the eyes and that gaze fixations could not predict the likelihood of perceiving the McGurk effect, in vowel–consonant–vowel (VCV) utterances (e.g., /ada/).

In summary, studies monitoring gaze behavior suggest that fixations tend to cluster primarily on the mouth and eyes of the talker and that gaze varies depending on the type of task (Lansing & McConkie, 1999), the degree of auditory information available (Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998), and the point in time in the trial at which gaze is measured (see Lansing & McConkie, 2003; Paré et al., 2003). However, it is not clear if gaze behavior during audiovisual speech perception varies depending on the resolution of the image. If gaze patterns differ as a function of the image resolution, then drops in speech intelligibility in visually degraded conditions could be attributed to a loss of visual detail, a change in gaze behavior, or both.

In this research, we conducted three studies to examine the effects of low-pass spatial frequency filtering on performance and gaze behavior using the McGurk effect and a visual-only speech perception task. In the first experiment, a low-pass filtering technique was used to filter images of VCV utterances, which were presented in an audiovisual condition and a visual-only condition. In the second set of experiments, participants performed the visual-only task (Experiment 2) and the audiovisual task (Experiment 3) while having their gaze behavior monitored using eye-tracking equipment.

The studies contribute to our understanding of audiovisual communication in conditions that have significant clinical importance. It is surprising how little we know about individuals with reduced vision who nonetheless depend on the visual modality to augment speech perception. When individuals have central scotomas, peripheral vision is sufficient for audiovisual speech perception (Wilson, Wilson, ten Hove, Paré, & Munhall, 2008). Legault, Gagné, Rhoualem, and Anderson-Gosselin (2010), using a simulated decrease in visual acuity, showed that quite reduced acuity still provided significant visual enhancement in both young and older individuals. However, it is important to understand how audiovisual integration occurs in various visual conditions and how information-gathering routines could contribute to multisensory processing. Dickinson and Taylor (2011) explored the effect of mild degrees of visual impairment (i.e., reduced contrast sensitivity) on audiovisual speech perception using sentences masked with noise. They found that, although performance was overall resilient to all levels of visual degradation, audiovisual speech intelligibility fell significantly with even minimal loss of contrast sensitivity.

By studying both visual- and auditory-only processing as well as how the two modalities are combined at various visual resolutions, we can see how individuals vary their gaze and their performance in response to these challenges.

Experiment 1

Method

Participants

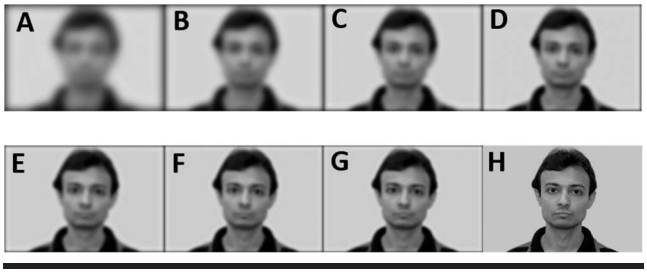
Sixty-six adults were tested (55 women, 11 men, *M* age = 21.45 years, *SD* = 5.14). All participants were fluent speakers of English with no known hearing, speech, or language disorders and normal or corrected-to-normal vision.

Stimuli

The stimuli consisted of eight VCV utterances (/aba/, /ada/, /aga/, /ava/, /aθa/, /ala/, /aʒa/, /aɹa/) spoken by a native English speaker and recorded full-face on digital videotape. The consonants were selected to represent a good sample of the visually distinct consonants (*visemes*) of English (Jackson, 1988).

The video recording was transferred to a computer as image sequences, converted to gray scale, and low-pass filtered using a second-order Butterworth filter, designed to achieve an attenuation of 3 dB (half power) at the desired cutoff frequency. The following frequency cutoffs (FCs) were used to create seven sets of VCVs: 4.0, 6.0, 8.0, 10.0, 12.1, 14.1, and 16.1 horizontal c/f along with the unfiltered condition, which averaged 228 c/f (see Figure 1).[1]

Figure 1. Reproduction of the stimuli used in the study. The images of the talker were filtered with a frequency cutoff of 4.0 (A), 6.0 (B), 8.0 (C), 10.0 (D), 12.1 (E), 14.1 (F), and 16.1 (G) cycles/face, together with an unfiltered image (H).



The original gray scale image sequences and each of the filtered image sequences were recorded to DVD along with the original audio track. All the auditory clips were band-pass filtered using Praat digital signal processing software (Boersma & Weenink, 2004) with cutoff frequencies of 75 and 10000 Hz. The filters removed high- and low-frequency noise from the recordings. The stimuli were presented in two modalities: visual-only and audiovisual. In the visual-only modality, the utterances were presented with no sound. In the audiovisual modality, the same utterances were presented in congruent and incongruent forms. In the congruent form, the auditory stimulus matched the visual consonants. In the incongruent form, an auditory /aba/ was dubbed onto each of the visual utterances, maintaining the timing with the visual stimulus of the original sound track. This editing was done using a custom program in MATLAB (The Mathworks, Inc.).

Equipment

Participants were tested in a single-walled sound isolation booth (Eckel Model C-17). They were seated at a table with their head positioned in a chin rest with their eyes approximately 57 cm from a 20-in. video monitor (JVC Model TM-H1950G). The stimuli were played with a Pioneer DVD Player, Model V7400, with a resolution of 720 × 480 pixels. The auditory signal was amplified (InterM R300 reference amplifier) and played through speakers (Paradigm Reference Studio 20) located on each side of the screen. The DVD trials were controlled by custom software, which also recorded the participants' responses. Responses were made on a keyboard with nine possible key responses: one for each consonant, plus a key labeled *other* that was used if they heard something different from the responses provided; chance performance thus being 0.11.

[1]Because the video pixels had a 10/11 vertical to horizontal ratio, the vertical FC are slightly higher in c/f.

Design

Each participant completed two tasks: a visual-only and an audiovisual task. The tasks were tested in separate blocks, and the order of tasks was counterbalanced across participants. There were a total of 64 distinct video clips, created from the eight tokens, and filtered with eight different FCs (unfiltered video and seven low-pass conditions). For the visual-only task, each of these distinct video clips was played three times for a total of 192 trials. The order of the trials was randomized for each participant within each block.

In the audiovisual task, the congruent stimuli consisted of 64 video clips: eight matched consonants in each of the eight filtered conditions. The incongruent stimuli consisted of 56 video clips: seven consonants (visual /aba/ was excluded) played with a simultaneous auditory /aba/ in each of the eight filtered conditions. These incongruent trials were expected to produce some form of the McGurk effect in which the visual stimulus changed the auditory perception. That is, the McGurk effect was defined here as categorical change in auditory perception induced by incongruent visual speech, resulting in a single percept of hearing something other than what the voice is saying (see McGurk & MacDonald, 1976; Munhall, Gribble, Sacco, & Ward, 1996; Paré et al., 2003; Tiippana, 2014). Again, the video clips were played three times in each condition, for a total of 360 trials: (64 + 56) × 3. The trials were broken up into two blocks in order to provide participants with a break. The first block contained 240 trials, and the second contained 120.

Procedure

Participants were instructed to watch the screen during the video clip and to press the key corresponding to the consonant they had heard (in the audiovisual condition) or seen (in the visual-only condition). Participants were told that their response would trigger the next stimulus presentation. The auditory stimuli were played at approximately 70 dB sound pressure level (A-weighting).

Data Analysis

Performance in the audiovisual and visual-only tasks was reported as a proportion of correct responses (auditory and visual, respectively) as a function of FC. The effects of FC on stimulus type (consonant) were assessed by conducting repeated-measures analyses of variance (ANOVAs). The Greenhouse–Geisser method of adjusted degrees of freedom was to correct for any biases due to violations of the assumption of sphericity.

Proportion values across participants were also fit with a Weibull function of the form $W(FC) = \gamma - (\gamma - \delta) \times \exp[-(FC/\alpha)^{\beta}]$ with threshold (α) and slope (β) as free parameters and the initial (γ) and end (δ) limits set to chance (1/9) and to the mean proportion of correct responses in the unfiltered condition; the reciprocal of this function was used to fit the audiovisual McGurk task data. The spatial frequency at which the proportion of correct responses reached an asymptote was defined as the value at which the 95th percentile of the end limit of the function was reached.

Results

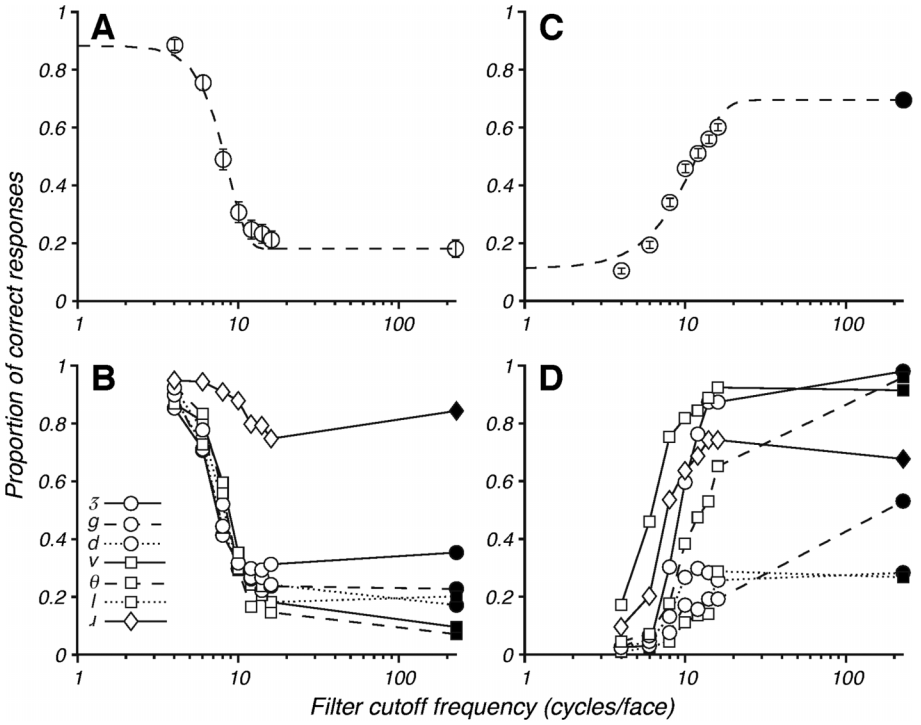
Audiovisually Congruent Condition

Performance was high across all levels of the audiovisually congruent condition regardless of FC and consonant —that is, the mean proportion of correct responses never dropped below 0.93 for any consonant at any FC. A within-subject 8 × 8 (FC × Consonant) ANOVA revealed no significant interaction and no significant main effects of consonant; however, a significant main effect of FC was found, $F(5.52, 385.59) = 2.41$, $p = .03$. None of the post hoc contrasts reached significance when corrected for multiple comparisons (Holm–Bonferroni method). Because the mean of correct responses was at ceiling in all FC (4.0 = .98, 6.0 = .99, 8.0 = .98, 10.0 = .98, 12.1 = .99, 14.1 = .99, 16.1 = .98, unfiltered = .99) and the effect size for this contrast is small ($\eta^2 = .02$), we believe it is explained by the low variability in some FCs.

Audiovisually Incongruent Condition

In general, performance in the incongruent condition varied as a function of the FC with less accurate perception of the auditory consonant /aba/ (stronger McGurk effect) at the higher FC and unfiltered condition (see Figure 2A). This pattern is evident for each of the individual consonants with a single exception, /aɹa/ (see Figure 2B and confusion matrices in the online supplemental materials, Supplemental Figure 1). The deviance of /aɹa/ may be explained by the fact that the temporal characteristics of the visual and auditory stimuli were too different to permit audiovisual integration or the lip rounding for the /ɹ/ was visually similar to /b/. For different reasons, both of these factors would lead participants to simply hear the auditory stimuli /aba/ (i.e., no McGurk effect). Because this stimulus was apparently processed differently than the others, we excluded it from further analyses. The 8 × 6 (FC × Consonant, excluding /aba/ and /aɹa/) ANOVA revealed a significant main effect of FC, $F(3.30, 211.28) = 284.24$, $p < .001$, and a significant interaction FC × Consonant, $F(20.9, 1{,}337.395) = 5.89$, $p < .001$. As can be seen in Figure 2B, the proportion of correct reports of the auditory consonant are much more spread out in the highest FCs than they are in the low spatial FCs. In particular, /ada/, /ava/, /aθa/—the visual utterances that influence the auditory signal less at lower FCs—are the ones that show the highest increase of the illusory percepts at high FCs. For other utterances (i.e., /aga/, /ala/, /aʒa/), the visual information found in the highest frequency range of the image does not seem to influence the magnitude of the illusion. The main effect of consonant was not significant. When /aɹa/ was included in the analyses, the main effect of consonant was significant, $F(3.33, 213.34) = 133.56$, $p < .001$. The best-fit function ($R^2 = .50$) of the proportion of correct reports of the auditory consonant as a function of FC reached an asymptote at 11.6 c/f (see Figure 2A), suggesting that

Figure 2. Mean (± *SE*) proportion of auditory correct responses as a function of frequency cutoff for incongruent audiovisual stimuli averaged across consonants (A) and for each individual consonant (B) in Experiment 1. Mean (± *SE*) proportion of visual correct responses as a function of frequency cutoff in the visual-only task averaged across consonants (C) and for each individual consonant (D) in Experiment 1. Dashed line in Panels A and C indicates the best-fit function (see text for details). Solid symbol corresponds to the unfiltered condition (228 cycles/face).



high spatial frequency information does not significantly alter audiovisual speech processing.

Audiovisually Congruent and Incongruent Conditions Contrasted

Eight paired-sample *t* tests were conducted to compare performance (averaged across consonants) in the congruent trials for each FC to the performance in the incongruent condition on the same FC. Familywise error rate across these tests was controlled by using Holm–Bonferroni correction. The *t* tests showed significant differences between the average performance in the congruent condition and the performance in the incongruent condition at each of the FCs (all *p*s < .001). This indicates that even at the lowest FC (4.0 c/f), the visual stimulus is having an influence on speech perception. This influence was most likely interference from the incongruent condition, although an increase in intelligibility for the congruent condition cannot be eliminated without an auditory-only condition, which our design did not include.

Visual-Only Condition

In the visual-only condition, there was a strong upward trend for consonant identification accuracy across the increasing FCs. This pattern can be seen in Figure 2, which depicts speechreading performance as a function of FC averaged across consonants (Figure 2C) and for each consonant individually (Figure 2D). The breakdown of data into consonants reveals that the individual patterns of different consonants varied considerably. In fact, the main effect of consonant was significant, $F(5.85, 380.28) = 167.36$, $p < .001$, as was the main effect of FC, $F(5.27, 342.24) = 370.35$, $p < .001$. The interaction Consonant × FC was also significant, $F(36.51, 2{,}373.20) = 22.97$, $p < .001$, indicating that the pattern for each consonant identification varied as a function of FC.

The best-fit function ($R^2 = .69$) of the proportion of visually correct responses as a function of FC reached an asymptote at 18.3 c/f (see Figure 2C). This suggests that, on average, speechreading is more greatly influenced from higher visual resolution than audiovisual speech processing; from 12 c/f and beyond, the proportion correct in speechreading changes by 0.236, ANOVA: $F(3, 260) = 27.03$, $p < .001$, whereas it changes only by 0.066 for audiovisual, ANOVA: $F(3, 260) = 0.86$, $p = .46$.

Speechreading Ability

We tested how individual differences in speechreading were reflected across spatial frequency by correlating the mean proportion of correct responses in the unfiltered visual-only condition (the most natural condition indexing speechreading ability) with the mean proportion of correct responses

in each of the filtered conditions. Spearman correlations were found to be significant for all FCs except for 4 c/f (4 c/f: $r = .05$, $p = .66$; 6 c/f: $r = .29$, $p = .018$; 8 c/f: $r = .32$, $p = .009$; 10 c/f: $r = .40$, $p < .001$; 12.1 c/f: $r = .46$, $p < .001$; 14.1 c/f: $r = .52$, $p < .001$; 16.1 c/f: $r = .56$, $p < .001$). These results suggest that good speechreaders extract more information from visual speech even at low visual resolution (i.e., 6 c/f).

Speechreading ability ranged from 0.38 to 0.92 and was found not to follow a unimodal distribution (dip test[2]; D = 0.0682, $p = .017$) with a first mode at a correct response proportion of 0.58 and a second at 0.71 (see Figure 3). To explore the frequency range that contributes the most to speechreading ability, we contrasted the performance of the best and worst participants. The worst speechreaders ($n = 18$) were taken as those with a proportion of correct responses < 0.6 in the unfiltered condition, which encompassed the first mode and its lower tail. The best speechreaders ($n = 12$) were taken as those with a proportion of correct responses > 0.8—that is, the upper tail of the second mode of the distribution. The difference in performance between the two groups was highly significant (0.86 ± 0.01 vs. 0.53 ± 0.02; t test, $t = 16.52$, $p < .0001$). A 2 × 8 split-plot ANOVA with speechreading ability as a between-participants factor (best and worst speechreaders) and FC as a within-participant factor revealed a main effect of FC, $F(3.99, 111.71) = 159.97$, $p < .001$, a main effect of speechreading ability, $F(1, 28) = 22.64$, $p < .001$, and significant interaction, $F(3.99, 111.717) = 10.75$, $p < .001$. Multiple independent samples t tests of performance of the two groups, corrected for multiple comparisons with the Holm–Bonferroni method, revealed that the best speechreaders did significantly better than the worst speechreaders at FCs >10 c/f (12.1 c/f: $t = 3.14$, $p = .004$; 14.1 c/f: $t = 3.64$, $p = .0011$; 16.1 c/f: $t = 3.47$, $p = .0017$). These results suggest that poor speechreaders gain significantly less from increasing spatial frequency information than do skilled speechreaders.[3]

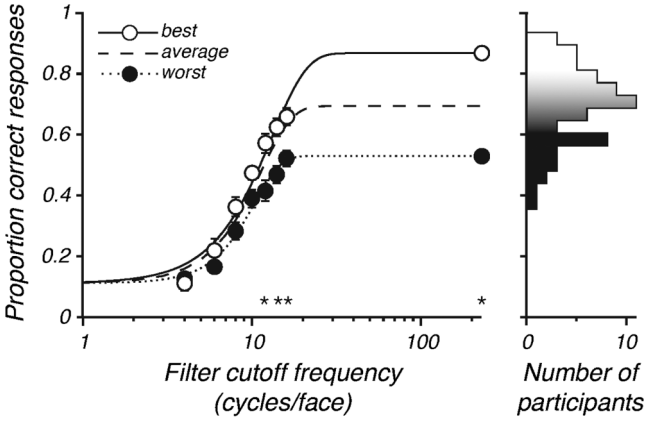
Speechreading Ability and the McGurk Effect

Individual differences in performance were compared across modalities by computing correlations of the mean proportion of correct responses in the unfiltered visual-only condition to the mean proportion of identifications of /aba/ in the unfiltered incongruent audiovisual condition (McGurk effect). A Pearson product–moment correlation revealed a nonsignificant correlation coefficient of .104 ($p = .40$). The correlations at the other FC were also nonsignificant.

[2]A *dip test* is a nonparametric test of bimodality in single-variate data (Hartigan & Hartigan, 1985).

[3]Further analyses revealed that the size and pattern of the difference between the highest and lowest skilled groups varied as a function of consonant. For instance, for the consonant /ɑθɑ/, the high and low skill groups cluster very closely across all the FC. For the consonant /ɑɹɑ/, the two groups are very distinct both at low and high FC. For the consonant /ɑgɑ/, the two groups cluster closely for the lower FC, but deviate from each other substantially in the highest FC.

Figure 3. Speechreading performance of participants as a function of frequency cutoff and of their speechreading ability assessed from performance in the unfiltered condition (228 cycles/face; histogram) in Experiment 1. Mean (± *SE*) proportion of visual correct responses (averaged across consonants) of the participants with the best speechreading ability (open) is contrasted with that of the participants with the worst speechreading ability (dark) along with the best-fit function of the performance across all participants (see Figure 2C). The best-fit Weibull function with limits set to chance and the mean of unfiltered data saturated at 22.2 cycles/face for the best speechreaders ($R^2 = .76$) and at 15.4 cycles/face for the worst ($R^2 = .60$). *Statistically significant ($p < .01$) difference between the two groups.



Discussion

The proportion of correct reports of the auditory consonant in the McGurk condition showed a downward trend across increasing visual resolution, but the psychometric function reached an asymptote at 11.7 c/f, showing the robustness of audiovisual speech processing to spatial frequency filtering. The general finding that lower FCs led to a decrease in altered percepts is consistent with previous research, showing a reduction of the McGurk effect with viewing distances of 20 m and above (Jordan & Sergeant, 2000) with eccentric viewing angles > 10°–20° (Paré et al., 2003) and with the image quantized below 29 pixels/face (MacDonald et al., 2000; Campbell & Massaro, 1997, found a decrease in speechreading performance for quantization below 16 pixels/face). The finding that high-resolution information is redundant for audiovisual speech perception is also consistent with Munhall et al. (2004), who found that performance in a speech-in-noise task reaches asymptote with low-pass filtered images with a FC of 7.3 c/f and above (although see Dickinson & Taylor, 2011). However, it is worth noting that in the current study the asymptote occurred at a higher FC (11.7 c/f and above). Two critical differences between Munhall et al. (2004) and the current study may explain the different outcomes. First, the stimuli were of different nature—that is, whereas Munhall et al. involved the perception of audiovisual sentences presented in noise, the present study required the performance of a closed-set identification task (i.e., syllable categorization). It is therefore possible that the distinctive features that allow participants to identify (or distinguish between) syllables

can only be extracted at higher spatial frequencies, or that context allows performance to reach asymptote for sentences at lower spatial frequencies. Second, the two experiments have different talkers as stimuli, and the clarity of speech of the two talkers' articulation may differ. It is well known that the visual and auditory characteristics of the talker can have a substantial impact on how the signal is processed by the observer (Bernstein et al., 2000; Cienkowski & Carney, 2002; Demorest & Bernstein, 1992; Jiang & Bernstein, 2011; Paré et al., 2003). It is thus possible that the talker in Munhall et al. articulated more clearly than the one here, thus allowing the extraction of visual cues at a lower visual resolution.

A comparison of the audiovisual and visual-only data suggests that the two modalities led to different perceptual patterns at high spatial frequencies—that is, whereas information above 11.7 c/f appeared to be redundant in the audiovisual modality, the asymptote in the visual-only modality was reached at 18.3 c/f. Thus, higher spatial frequency information seems to be less important for audiovisual speech processing than for speechreading, a finding consistent with previous literature indicating that higher detail information is required for visual-only more than for audiovisual speech stimuli (Lansing & McConkie, 2003; Munhall & Vatikiotis-Bateson, 2004). A potential explanation for this effect is that, when present, auditory information provides complementary cues to that provided by the higher visual frequency information. Abel, Barbosa, Black, Mayer, and Vatikiotis-Bateson (2011) found that subtle visual information for some manners of articulation or for voicing of phonemes belonging to the same viseme group (/p/, /b/, /m/) can indeed be observed when analyzing the facial kinematics of the talker's orofacial motion in an unfiltered setting. It is thus possible that such subtle visual cues—present in the high-frequency range of the image—are available but not used by most perceivers because they provide information that is redundant to that gathered through the auditory signal.

However, before interpretations of these differences can be considered, it is important to note that comparisons between the two modalities are tempered by the fact that the two modalities are scored in slightly different ways. The performance in the speechreading task is represented by the choice of a certain response (the consonant for that trial) whereas the McGurk effect is represented by the absence of a certain response choice (not /b/). This means that there are far more choices that indicate perception of the McGurk effect than there are that indicate speechreading accuracy. To explore if the manner in which the conditions are scored might be influencing the observed data patterns, two consonants were examined individually: /th/ (from /aθa/) and /v/ (from /ava/). These consonants were chosen because participants responded consistently to the visual component of the stimuli in the unfiltered incongruent condition. In this analysis, the visual consonants /th/ and /v/ in the incongruent condition were scored in such a way that a response was considered correct if the participant chose /th/ or /v/ (respectively) as their response. When scoring /th/ with this new method, we observed a pattern that is much closer to that of the visual-only modality and different greatly from the pattern observed with the "non-/b/" scoring. It is interesting to note that the patterns for the consonant /v/ are remarkably different from those of /th/. Unlike /th/, the proportion of /v/ responses in the audiovisual condition was generally lower than that of the visual /v/ performance. The patterns for /v/ bring to question if we can simply dismiss the differences observed between the audiovisual and visual-only modalities as an artifact of the different scoring methods. It is, however, impossible to reach any conclusions without analyzing the other consonants as well. It is unfortunate that it is not possible to use a similar representation for the majority of the consonants because most consonants elicit a variety of responses from participants. Further research will be needed to confirm the source of the differences between the two modalities.

In the visual-only condition, performance for all consonants improved with increasing FCs (i.e., visual resolution), but the change in improvement varied considerably between consonants. Several reasons may explain the large variability between consonants in the visual-only condition. First, the differences between consonants might in part reflect variations in the absolute ease of their recognition. In other words, consonants that are easier to distinguish in unfiltered speechreading might tend to be more easily perceived at the lower filter levels, reaching ceiling at a lower FC (e.g., /v/). Second, the differences may have occurred because of the combination of consonants in the set. Certain pairs of consonants within the set used in this study are more easily confused. If two consonants were easily confused (e.g., /d/ and /g/), it would result in a lower overall accuracy for both consonants as compared with the other consonants in the set. This analysis, however, is beyond the scope of the present publication, and it will be left to future investigation. A third reason for the differences in the patterns of different consonants is that the consonants vary in terms of what features distinguish them (Campbell & Massaro, 1997), and some of these features might be more resilient to FCs than others. Because these visual features are dependent on the characteristics of the talker, however, this hypothesis cannot be tested without a larger sample of talkers.

This study has also addressed the topic of speechreading ability. There is tremendous variety in individual speechreading ability, and the basis for this variance is a source of debate (Bernstein et al., 2000; MacLeod & Summerfield, 1990; Rönnberg, 1995; Rönnberg, Arlinger, Lyxell, & Kinnefors, 1989; Summerfield, 1992). Our results show that variance in speechreading ability can be accounted for, in part, by the extent to which people benefit from high-frequency spatial information. It is not clear, however, what enables the better speechreaders to benefit from the higher spatial frequency. One possibility is that better speechreaders have an underlying visual processing ability that enables them to extract more information from the visual image (Auer & Bernstein, 1997; Feld & Sommers, 2009; Gagné, Charbonneau, & Leroux, 2011). Another possibility is that the ability to use spatial frequency is the result of a different speechreading strategy. For instance, it may be that participants who are

better speechreaders tend to focus their attention on an area of the face that benefits more from the high definition provided by high spatial frequencies, a question that we address in Experiments 2 and 3.

Another finding related to the speechreading proficiency sheds new light on the controversy regarding the relationship between speechreading performance and audiovisual speech perception. We found no correlation between speechreading ability and the McGurk effect, a finding that is consistent with that of Cienkowski and Carney (2002) and Munhall and Vatikiotis-Bateson (2004), who found no correlation between the McGurk illusion and sentence speechreading. Our results are also consistent with that of Strand et al. (2014) who used VCV utterances and found no correlations between the ability to speechread consonants and the magnitude of McGurk effect. It is important to note, however, that Strand et al. did find moderate correlations when quantifying speechreading ability using a broader measurement (i.e., the ability to identify the point of articulation of the utterance).

Experiment 2

The goal of Experiment 2 was to determine how information is gathered from a talker's face during silent speech perception in conditions of poor visual resolution. Furthermore, we explored the relationship between gaze and performance by examining the individual speechreading ability of each participant and their gaze fixations toward the mouth. To do so, we monitored gaze behavior as participants performed the visual-only task from Experiment 1.

Method

Participants

Twenty native English speakers participated in Experiment 2 (16 women, four men; *M* age = 22.3 years; *SD* = 7.5). The participants had no known hearing, speech, or language disorders and had normal or corrected-to-normal vision. The first 10 participants had previously participated in Experiment 1.[4]

Eye-Tracking Equipment

Gaze position was recorded with the EyeLink-II system (SR Research Ltd.). Gaze position was measured with a sampling rate of 500 Hz and with an average error of less than 0.5° of visual angle. A calibration and validation procedure was used, using nine points that were located in the center of the screen and spaced equidistantly in three rows around the periphery (73 pixels from the side of the screen, 60 pixels from the bottom of the screen, and 52 from the top of the screen). If there was a discrepancy between the calibration and validation of more than 1.5° for any point or an average discrepancy of more than 1.0°, the calibration procedure was repeated. A drift correction was performed between trials with a maximum acceptable error of 1.5°.

Design

Each of the 64 original stimuli (eight CVCs filtered at eight levels) was played three times for a total of 192 trials. The 192 trials were divided into two blocks in order to provide participants with a break if needed. Within each block, the trials were played in a random order.

Procedure

Participants were fitted with the Eyelink-II headband, the cameras were adjusted and focused, and participants performed a nine-point calibration and validation procedure. The 10 participants who had participated in Experiment 1 were already familiar with the procedure. They were told they would be performing the same task as the visual-only section of the first experiment and were refreshed on the possible consonants they would be seeing. The 10 participants who had not been in a previous study were shown eight practice clips in which they saw the unfiltered congruent clips of each consonant. They were then told that they would be asked to perform a silent speechreading task in which they would make a key press to identify the consonant they had seen. Participants were told that their response would trigger the next stimulus presentation.

Gaze Analysis

To analyze gaze position with respect to points on the talker's face in the video, three regions of interest (ROIs) were defined, on the basis of frame-by-frame coding of all stimuli (see Figure 4; also see Buchan et al., 2007, for similar ROIs). The first ROI, corresponding to the right eye plus the left eye, consisted of ellipses centered on each pupil with a horizontal radius of 70 pixels and a vertical radius of 50 pixels (3.91° × 3.14° of visual angle). The second ROI, corresponding to the nose region, consisted of an equilateral triangle. The two bottom corners of the triangle were located 15 pixels horizontally out from each nostril (total 2.8° of visual angle), and the upper corner was determined on the basis of the equation of an equilateral triangle (2.5°). The third ROI, corresponding to the mouth, consisted of an ellipse, which was calculated to pass through the following four points: the left and right corners of the mouth with five pixels added horizontally and the upper and lower edges of the mouth with five pixels added vertically. The eyes and nose ROIs remained constant in size across the frames;

[4]To check for differences in the effect of spatial frequency filtering on participant groups, the data from the two groups of participants (i.e., naïve, previously trained) was analyzed using 2 × 8 (Group × FC) mixed ANOVA for each region of interest (ROI) to test for any Group × Filter interactions. The main effects of group were not significant, eye: $F(1, 18) = 0.276$, $p = .11$; nose: $F(1, 18) = 0.11$, $p = .74$; mouth: $F(1, 18) = 0.56$, $p = .46$, and neither were the Group × FC interactions in the eye, $F(1.55, 27.9) = 1.07$, $p = .34$, and nose ROIs, $F(2.62, 47.21) = .91$, $p = .43$. The Group × FC interaction in the mouth ROI did not reach significance, $F(2.36, 42.46) = 3.01$, $p = .052$. The two groups were also compared in terms of their speechreading performance. Neither the main effects of group or the Group × FC interactions were significant, $F(1, 18) = 0.36$, $p = .56$; $F(3.7, 66.3) = 1.62$, $p = .18$.

Figure 4. Reproduction of the stimuli used in the study. Depiction of the regions of interest used for the gaze analyses in Experiments 2 and 3.

however, the mouth ROI varied in size as the talker opened and closed his mouth throughout speech.

In addition, to ensure that any additional fixations were primarily falling within the face region, an ellipse was defined that encircled the face. The ellipse was centered on the two nostrils and extended 224 pixels horizontally and 320 pixels vertically (12.46° × 19.90°).

Results

Speechreading Performance

Performance across FCs was similar to that in Experiment 1 (see Figure 5A). A within-subject 8 × 8 (FC × Consonant) ANOVA revealed a significant effect of FC, $F(7, 67.89) = 51.12$, $p < .001$, and consonant, $F(7, 74) = 96.29$, $p < .001$, and a significant interaction, $F(49, 246.05) = 8.2$, $p < .001$. The best-fit function ($R^2 = .72$) of the proportion of visually correct responses as a function of FC reached an asymptote at 17.4 c/f, a result again similar to that of Experiment 1.

Gaze Behavior and Image Resolution

As shown in Figure 6A, when there was higher spatial frequency content in the images (i.e., better image resolution), there was an increased tendency to focus on the mouth and a decreased tendency to focus on the nose. Very little time was spent with gaze on the eye region. Indeed, a within-subject, one-way ANOVA revealed a significant effect of FC for the mouth, $F(2.11, 40.08) = 18.08$, $p < .001$, and the nose ROIs, $F(2.74, 52.19) = 11.85$, $p < .001$. The effect of FC on the eye region did not reach significance, $F(1.58, 30.00) = 3.27$, $p = .06$.

Gaze Behavior: Individual Consonants

To explore if performance and gaze behavior differed across consonants, the data from the unfiltered condition was analyzed by consonant for speechreading performance and for the percentage of time spent with gaze directed to the mouth. Gaze behaviors toward the mouth (the dominant ROI) were quite consistent across consonants, although participants looked more frequently to /ada/ than to the other consonants (84% vs. approximately 76%). A within-subject, one-way ANOVA of consonant on the percentage of time gazing at the mouth did not reach significance ROI, $F(7, 133) = 1.85$, $p = .08$.

To verify if similar trends were observed when high spatial frequency information was removed, the same analysis was also carried out for two filtered conditions, one at an intermediate FC (10 c/f) and the other at a low FC (4 c/f). In both conditions, there was a significant effect of consonant on performance, $F(3.17, 60.17) = 18.17$, $p < .001$ and $F(7, 133) = 15.98$, $p < .001$ for 10 c/f and 4 c/f, respectively, but no significant effect of consonant on the percentage of time with gaze at the mouth ROI, 10 c/f: $F(4.18, 79.49) = 1.38$, $p = .25$; 4 c/f: $F(7, 133) = 0.731$, $p = .64$. Even in

Figure 5. (A) Mean (± *SE*) proportion of visual correct responses (averaged across consonants) as a function of frequency cutoff in the visual-only task in Experiment 2. (B) Mean (± *SE*) proportion of auditory correct responses (averaged across consonants) as a function of frequency cutoff for incongruent audiovisual stimuli in Experiment 3. Dashed line represents the best-fit function (see text for details). Solid symbol corresponds to the unfiltered condition (228 cycles/face).

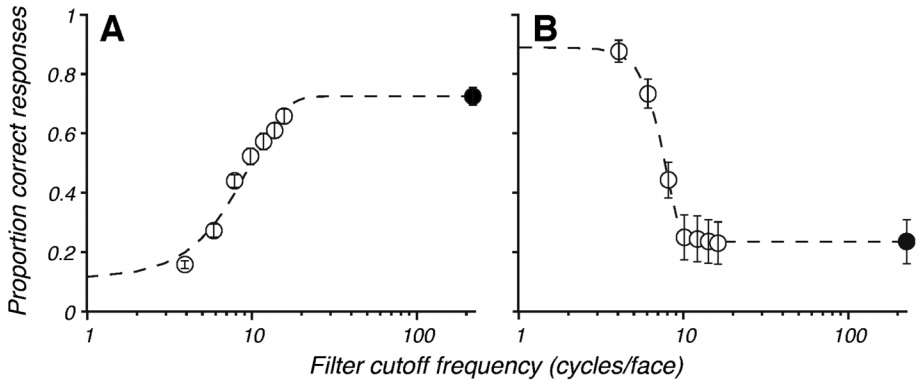
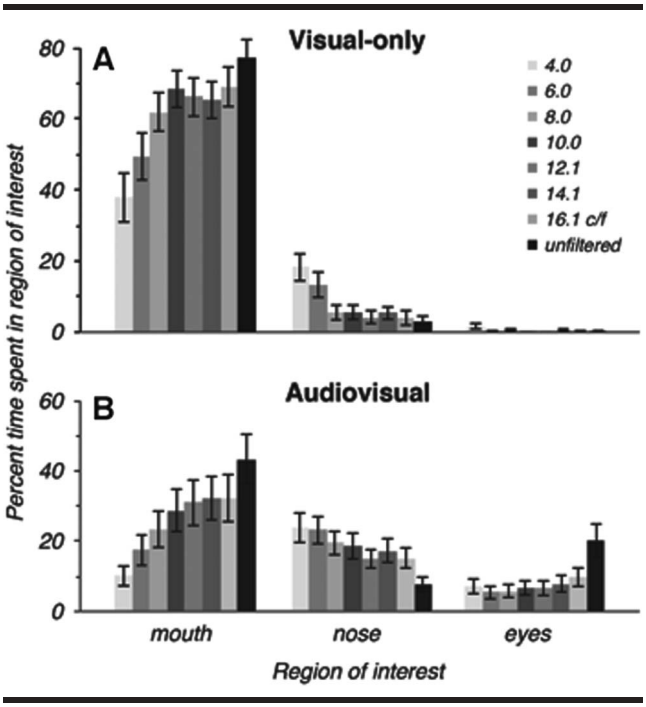


Figure 6. (A) Percentage of time (± *SE*) spent with gaze directed to each region of interest as a function of frequency cutoff in the visual-only task of Experiment 2. (B) Average percentage time (± *SE*) spent with gaze directed to each region of interest as a function of frequency cutoff in the audiovisual task (congruent and incongruent collapsed) of Experiment 3.



these filtered conditions, gaze behavior did not seem to be influenced by the consonants.

Eye Gaze and Accuracy

To analyze the relationship between speechreading performance and gaze strategy, the proportion of correct responses produced across participants and time spent gazing at the mouth region for each subject was assessed for each filter condition. A Pearson product–moment correlation revealed nonsignificant correlation coefficients between the mean percentage of time spent with gaze in the mouth ROI and the mean proportion of correct responses (see Table 1, Experiment 2). These results suggest that differences between proficient speechreaders and poor speechreaders are not a function of their gaze behaviors toward the mouth.

Discussion

Performance on the visual-only task across FCs was similar to that in Experiment 1, demonstrating again that performance increases monotonically with spatial frequency. In addition, fixations increasingly rested on the mouth as higher spatial frequency information was provided. When visual resolution decreased, perceivers tended to spend less time focused on the mouth and more time with gaze focused centrally on the nose. It is interesting to note that, although the stimulus utterances differed in their visual distinctiveness,

Table 1. Pearson correlation coefficients between the mean percentage of time spent with gaze in the mouth region of interest to the mean proportion of correct responses in the visual-only condition (Experiment 2) and to the proportion of identifications of /aba/ in the incongruent audiovisual condition (Experiment 3), for each frequency cutoff.

Cycles/face	Experiment 2	Experiment 3
4.0	$r = .21$, $p = .37$	$r = .17$, $p = 48$
6.0	$r = .09$, $p = .71$	$r = .11$, $p = .65$
8.0	$r = –.24$, $p = .30$	$r = .12$, $p = .60$
10.0	$r = –.13$, $p = .59$	$r = .23$, $p = .34$
12.1	$r = –.19$, $p = .42$	$r = .35$, $p = .13$
14.1	$r = –.06$, $p = .79$	$r = .35$, $p = .13$
16.1	$r = –0.5$, $p = .85$	$r = .24$, $p = .30$
Unfiltered	$r = .01$, $p = .95$	$r = .30$, $p = .20$

gaze behaviors toward the mouth (the dominant ROI) were quite consistent across consonants. Note, however, that the absence of a difference in gaze fixations for the different consonants might be explained by the fact that consonants were presented in a random order in the present study, therefore potentially preventing participants from adopting a strategy that would allow them to optimize the extraction of visual information in each speech item. Moreover, fixation lengths were averaged across trials. It is possible, therefore, that differences between consonants could be found when examining the temporal gaze patterns within each stimulus (Lansing & McConkie, 2003; Paré et al., 2003). Future analyses will shed some light on this issue. Last, the time spent gazing at the mouth region appeared to be unrelated to speechreading accuracy.

Although previous studies have not looked directly at the effects of visual manipulations on gaze behavior during speechreading, useful comparisons can be drawn from earlier research examining speech perception and gaze behavior. One topic that has been of considerable interest in eye-tracking studies of speech perception is the extent to which people gaze at the mouth. Gaze toward the mouth is higher for tasks that require discrimination of the phonetics of speech rather than the stress patterns (Lansing & McConkie, 1999) and is higher during sections of the trial when the talker is speaking, compared with sections when the talker is silent (Lansing & McConkie, 2003). In addition, the duration of fixations to the mouth increases when the audio signal is absent (Lansing & McConkie, 2003) and may increase with higher auditory noise levels (Buchan et al., 2008). On the basis of these findings, Lansing and McConkie (2003) suggest that people fixate increasingly on the mouth when the difficulty of the speech perception task is increased. The findings from the current study, however, suggest that increasing the difficulty of a speech perception task will not necessarily lead to increased gaze toward the mouth. When the task difficulty is manipulated through a degradation of visual resolution, gaze to the mouth actually decreases rather than increasing.

One potential explanation for the different gaze behavior observed as a function of spatial filtering is that the visual degradation technique used in the present stimuli

produced a modulation of the visual saliency maps of the image. Visual saliency is known to produce systematic changes in participants' gaze patterns (e.g., Enns & MacDonald, 2013; Itti & Koch, 2000; Latif, Gehmacher, Castelhano, & Munhall, 2014; Parkhurst, Law, & Neibur, 2002), and thus it is possible that the differences in the low-level properties of the image caused by the spatial filtering led participants' gaze to be drawn toward the mouth when the resolution was good and toward different areas when it was poor.

Another likely explanation for the decrease in gaze toward the mouth when the resolution is low is that perceivers do not gain as much information by fixating on that region. When the image is clear, the details of tongue and lip movement can be more clearly perceived by placing the mouth on the fovea. When the resolution is degraded, however, participants can gain the same amount of information from more peripheral areas of their retina. Furthermore, by keeping the eyes stationary at a central region, they can keep both the eyes and the mouth as close as possible to central vision and minimize the cost of saccadic planning and integration across saccades, allowing them to extract more temporal information from the visual image (Buchan et al., 2007).

It would seem, therefore, that people gaze toward the lips when there is more information to be gained in doing so. It is surprising, however, that our results provide no evidence that looking at the mouth actually benefits performance. There was no correlation found between a person's speechreading accuracy and the extent to which a person fixated on the mouth. These results are consistent with Lansing and McConkie (2003), who found no improvement in speechreading as people directed their gaze at the mouth, and with Paré et al. (2003), who found no difference in perception of the McGurk effect when people were instructed to fixate on the eyes compared with the mouth.

The results therefore challenge the hypothesis that better speechreaders use a more effective gaze strategy, fixating on regions of the face that offer greater benefit from the high definition provided by high spatial frequencies. Thus, there must be another explanation for the findings in Experiment 1 showing that good speechreaders benefit more from high-frequency spatial information than do bad speechreaders. One possibility is that better speechreaders have a low-level visual processing ability that enables them to extract more information from the visual image, independent of where they are looking on the face, such as a superior processing speed capacity or a better capacity to encode speech elements from details (Auer & Bernstein, 1997; Feld & Sommers, 2009; Gagné et al., 2011).

Gaze toward the eyes in this experiment was extremely low. This might seem quite surprising given that some speech perception studies have shown gaze to the eyes to represent as much as 45% to 70% of fixations (Vatikiotis-Bateson et al., 1998). One reason for the different outcomes in Vatikiotis-Bateson et al. (1998) and the present study is that Vatikiotis-Bateson et al. used a larger ROI for the eyes than the ROI defined here. However, other factors might explain the reduced fixations to the eyes in the present study. First, gaze behavior was only examined during speech periods. Lansing and McConkie (2003) found that gaze toward the eyes was higher for nonspeech periods in the trial. Another factor is the limited duration of the stimuli we used. In another study using VCV stimuli that were similar to the stimuli used in this study (Paré et al., 2003), the proportion of fixations toward the eyes was also quite low (11%). It seems likely that during the extended monologues in the study by Vatikiotis-Bateson et al., participants looked toward the eye regions for social reasons, whereas during the more limited time frame of the short VCV stimuli they do not have as much incentive to gaze toward the eyes (see also Saalasti et al., 2012).

In summary, the results of Experiment 2 suggest that people show a tendency to focus on the mouth during a silent speechreading task, especially when the image contains high spatial frequency information. This tendency does not appear to be related to a gain in performance, nor does it seem to be related to the consonant. In Experiment 3, we explore if the same trends hold true for audiovisual perception of McGurk stimuli.

Experiment 3

Experiment 3 explores if spatial frequency filtering affects eye movements for an audiovisual speech perception task, examining gaze behavior during the audiovisual McGurk task used in Experiment 1. In this experiment, we also assessed the relationship between performance and gaze behavior by examining if there was a correlation between the occurrence of the McGurk effect and time spent with eye gaze at the mouth ROI.

Method

Participants

Twenty new participants were tested (13 women, seven men; *M* age = 19.8; *SD* = 1.4). All participants were fluent speakers of English with no known hearing, speech, or language disorders and normal or corrected-to-normal vision.

Eye-Tracking Equipment

The eye-tracking equipment and procedures were the same as in Experiment 2.

Design

The same stimuli from the audiovisual task in Experiment 1 were presented. The 120 distinct stimuli (eight audiovisually congruent and seven audiovisually incongruent in eight FCs) were presented three times in separate blocks.

Procedure

Participants were shown three practice clips of congruent stimuli (/aba/, /aga/, /ava/). They were instructed to watch the screen during the trial and then make a key press corresponding to the consonant they heard. A short rest period was provided between each block.

Results

Performance

The results were similar to those in Experiment 1 (see Figure 5B). Identification of the auditory target was almost perfect for congruent stimuli across all filter conditions. Neither the main effects of FC, consonant, nor the interaction were significant. For the incongruent stimuli, increasing FC led to decreased identification of the auditory target (/aba/, increased McGurk effect). Again, this pattern was observed for all consonants except for /aɹa/, which was excluded from the analyses. The main effect of FC, $F(1.7, 32.45) = 62.88$, $p < .001$, was significant with McGurk reports decreasing as filtering severity increased, and so was the interaction of FC × Consonant, $F(9.8, 186.38) = 3.15$, $p = .006$, with some consonants being more affected by spatial filtering than others. When /aɹa/ was included in the analyses, the main effect of consonant was also significant, $F(2.28, 43.34) = 41.88$, $p < .001$, as participants consistently reported the auditorily specified consonant across filters. In a similar manner to Experiment 1, the best-fit function ($R^2 = .43$) of the proportion of correct reports to the auditory consonant as a function of FC reached an asymptote at 9.8 c/f.

Gaze Behavior and Visual Resolution

Gaze behavior was studied to determine if increasing the spatial resolution of the image would cause participants to increasingly focus on the mouth, nose, and eyes during an audiovisual speech perception task. We initially compared whether different gaze behavior was observed in the congruent and incongruent conditions as a function of FC in each ROI. Within-subject 8 × 2 (FC × Congruency) ANOVA calculations for each ROI showed main effects of congruency in the mouth ROI only, $F(1, 19) = 7.34$, $p = .014$, with participants in the incongruent condition looking for slightly longer periods of time to the mouth region. Because no interaction was found between Congruency × FC ($p = .31$), the data for the congruent and incongruent trials were averaged to study how eye behavior changed as a function of FC.

As in Experiment 2, participants spent the majority of their time focused on the mouth and nose regions. When the visual resolution improved, there was an increased tendency to focus on the mouth and eyes and a decreased tendency to focus on the nose. These results are summarized in Figure 6B, which displays the percentage of time in each ROI as a function of filter level.

A within-subject, one-way ANOVA was conducted for each ROI to analyze the effect of filtering on the percentage of time spent in each ROI. A significant effect of filter was found for every ROI, eyes: $F(1.71, 32.40) = 10.66$, $p = .001$; nose: $F(2.79, 52.96) = 8.02$, $p < .001$; mouth: $F(1.96, 37.39) = 22.69$, $p < .001$, with participants looking more at the eye and mouth regions (and less at the nose) as the resolution increased. With poor spatial resolution, participants tended to focus centrally on the nose or regions of the face outside the ROIs. Across the filter conditions, participants spent on average between 29% and 59% of their time with gaze outside the ROIs. The majority of these fixations fell on the face, with time outside the face region corresponding to only 4% to 8% of gaze.

Gaze Behavior: Individual Consonants

To explore if McGurk effect and gaze behavior differed across consonants, data from the unfiltered condition were analyzed by consonant for auditorily correct responses and for the percentage of time spent with gaze directed to the mouth. The proportion of correct responses did not vary across the consonants as shown by a within-subject one-way ANOVA, $F(3.07, 58.43) = 2.04$, $p = .17$, nor did gaze fixations at the mouth (the dominant ROI), $F(5, 95) = 0.72$, $p = .62$.

The same analysis was also carried out for filtered conditions with FCs of 4 c/f and 10 c/f to evaluate if similar findings held true across the filter conditions. There was no significant effect of consonant on the McGurk effect, either with a FC of 10 c/f, $F(3.27, 64.71) = 1.64$, $p = .18$, or with a FC of 4 c/f, $F(3.72, 74.33) = 2.05$, $p = .10$. In both conditions, there was no significant effect of consonant on the percentage of time with gaze at the mouth ROI, $F(5, 100) = 1.00$, $p = .42$ for 10 c/f and $F(5, 100) = 0.10$, $p = .42$ for 4 c/f. Again, these results suggest that gaze behavior does not vary for different consonants. When /aɹa/ was included in the analyses, the ANOVA that tested the proportion of auditorily correct responses across the consonants revealed significant differences in the unfiltered condition, $F(3.54, 67.36) = 27.96$, $p < .001$, and in FC of 10 c/f, $F(3.68, 73.64) = 30.1$, $p < .001$. All the other effects remained nonsignificant.

Eye Gaze and the McGurk Effect

To test the relationship between McGurk effect and time fixated on the mouth, a Pearson product–moment correlation was conducted, comparing the percentage of time spent with gaze in the mouth ROI with the proportion of identifications of /aba/ in the unfiltered incongruent audiovisual condition. Nonsignificant correlation coefficients were found for all seven FCs and the unfiltered conditions (see Table 1, Experiment 3)—that is, the percentage of time spent with gaze at the mouth was very similar regardless of whether participants perceived the McGurk effect. This suggests that individual differences in the McGurk effect are not due to differences in gaze strategies (Paré et al., 2003).

Comparison of Results of Experiments 2 and 3

To explore if the audiovisual task in Experiment 3 led to different gaze behavior than the visual-only task in Experiment 2, findings from the two studies were contrasted. Gaze behaviors across FCs within each ROI were similar in Experiments 2 and 3. In both experiments, there was a higher percentage of time spent looking at the mouth and a lower percentage of time spent looking at the nose and outside the ROIs as the video resolution increased. Despite these similarities, however, the average amount of time spent in each ROI differed substantially between experiments.

A split-plot ANOVA with ROI and filter as within-participant factors and experiment (audiovisual, visual-only) as a between-participants factor revealed a significant interaction of ROI × Experiment, $F(1.31, 49.95) = 20.80$, $p < .001$. The interaction FC × Experiment was not significant, $F(3.03, 115.32) = 1.15$, $p = .33$. A series of two-sample separate variance t tests (corrected with Holm–Bonferroni) comparing the percentage of time spent in each ROI in Experiments 2 and 3 revealed significant differences between the two experiments for every ROI (eyes: $t = -3.63$, $p = .002$; nose: $t = 2.88$, $p = .007$; mouth: $t = 4.69$, $p < .001$). These results suggest that when performing a visual-only task, participants spend more time focused on the mouth, whereas for an audiovisual task, participants spend more time with their gaze focused on other regions of the face.

Discussion

As in Experiment 1, the McGurk effect increased when higher spatial frequency information was available, but the size of the effect reached an asymptote at moderate levels of spatial FCs. The trends observed in eye movements toward the mouth in Experiment 3 were similar to those in Experiment 2—that is, when higher spatial frequency information was provided, gaze was increasingly directed toward the mouth regions and less time was spent with gaze fixated on the nose or outside the features of the face. Gaze toward the eyes also increased with the available high spatial frequency information in the image.

Although the trends across filter conditions were similar for Experiments 2 and 3, the proportion of time spent fixated on the mouth was considerably higher for the silent speechreading task in Experiment 2 than for the audiovisual McGurk task of Experiment 3. Gaze toward the eyes ROI, in contrast, was significantly higher in Experiment 3. This replicates Lansing and McConkie's (2003) finding of increased mouth fixations for a silent speech perception task compared with an audiovisual task, a finding that they interpreted as suggesting that people gaze toward the mouth when the (auditory) speech perception becomes more difficult and requires more visual information.

Another possible interpretation for the results is that perceivers may use different gaze strategies because the processes involved in extracting visual speech information in the presence of sound are different from the processes involved in silent speechreading (Cienkowski & Carney, 2002; MacLeod & Summerfield, 1990; Munhall & Vatikiotis-Bateson, 2004)—that is, it may be that the perception of visual-only speech is not simply more difficult than the perception of visual cues in audiovisual speech, but also differs in other ways. For example, the observed differences could be due to the fact that the extraction of speech information in complete silence (i.e., visual-only condition) is an unusual and unnatural task for most people.

The unnatural characteristics of visual-only speechreading may also mean that the conventional habits that we have developed for audiovisual speech perception need not apply. During face-to-face communication, gaze toward the face is not simply for comprehension of the spoken utterances, but also serves a number of social purposes. For example, through gaze, listeners display their interest in and their reactions to the conversation topic, show their relationship with the talker (degree of intimacy), show their emotional state, and give cues to the talker about when conversational turn-taking should take place (Mirenda, Donnellan, & Yoder, 1983). In most communication settings, the social cues that the perceiver relays through his or her gaze behavior may be more important than the extraction of visual speech information. On the other hand, during (silent) visual-only speech perception, perceivers might increasingly look toward the mouth because they are no longer constrained by social gaze requirements, and the social cues become secondary to the main task—that is, understanding speech.

As in Experiment 2, the results of Experiment 3 do not show any correlation between mouth fixations and visual speech perception as evidenced by the lack of association between gaze to the mouth and the McGurk effect (see also Paré et al., 2003). Also, there were no significant differences in gaze behavior across consonants, which suggests that the focus of perceivers is not necessarily driven by how visually informative the stimulus is. Again, perceivers look to the mouth when there is no benefit to doing so from a speech standpoint. It is likely that perceivers are gleaning information from the mouth through their peripheral vision and that increased gaze to this region is a result of eye movements correlated with attention.

General Conclusions

The data from these studies document the impact of visual resolution on audiovisual and visual speech perception. The results suggest that there is a difference between the use of high spatial frequency information in visual-only and audiovisual speech processing. It is significant that the results provide evidence of a relationship between speechreading ability and the benefit from high-resolution visual information. Such high spatial frequency information does not modify the perception of the McGurk effect, however. Last, speechreading ability is not correlated to the perception of the McGurk effect.

The data suggest that, for some individuals, the fine facial detail information found in the higher frequency range of the image is necessary in order to achieve optimal performance in a visual-only task. Although individuals on average modified their gaze for visual-only speech, with more fixations on the mouth region than for the McGurk effect, the location of gaze did not correlate with individual differences in either visual-only speech intelligibility or the susceptibility to the McGurk effect. This suggests that the interindividual differences in visual-only speech intelligibility do not stem from different gaze strategies adopted by participants.

It is puzzling that the better performers on visual-only perception did not show more mouth fixations because Experiment 1 showed that the best speechreaders could use

high spatial frequency information to their advantage. There are a number of possible explanations for this. The first possibility is simply sample size. Our first experiment had a large number of subjects, an important feature for individual difference studies. The eye-tracking study, on the other hand, had a sample size about a third of the size of Experiment 1 and thus may not have sampled enough good speechreaders.

Another possible account is that our eye-tracking measures are not sensitive to subtle differences in the timing of gaze. Almost all of the research on gaze in visual speech, including this study, has taken measures that do not take into account the dynamics of fixations during a trial. Most studies of speech perception and eye movements have focused on average position of gaze or on location of fixation at selected moments in time. Fixations occur in sequence at different locations across the face. Although the visual information from the face is distributed spatially and temporally, the exact sequence of fixations and their locations may make a difference. The scan paths (i.e., sequences of fixations) may account for some of the variance in visual speech identification. It might be that gaze must be focused at just the right moment on the mouth or, alternately, that many "moments" over the course of a gesture are sufficient to infer the dynamics of oral movement. This will require future research.

The variance observed between different aspects of visual speech perception and audiovisual speech perception is still a profound problem for researchers (e.g., Strand et al., 2014). These individual differences are crucial issues for clinical science (Åsberg Johnels, Gillberg, Falck-Ytter, & Miniscalco, 2014; Dickinson & Taylor, 2011), and continued large studies are an essential ingredient of a resolution to such problems.

In terms of spatial filtering, we found that as the image resolution decreased, observers tended to gaze less at the mouth and eyes and more at central regions in both perception tasks. Although in some clinical populations, such as patients with central scotomas, there are beneficial interventions to train individuals to place visual information in certain retinal positions (e.g., Nilsson, Frennesson, & Nilsson, 1998; Wilson et al., 2008), it appears that gaze training would have limited impact on speechreading or audiovisual speech perception in a broad population. Our study and findings by others (e.g., Paré et al., 2003) show that fixation location on the face does not predict performance.

Acknowledgments

Amanda H. Wilson, Agnès Alsius, and this research have been supported by the Canadian Institutes of Health Research (CIHR) and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Abel, J., Barbosa, A. V., Black, A., Mayer, C., & Vatikiotis-Bateson, E. (2011). The labial viseme reconsidered: Evidence from production and perception. *The Journal of the Acoustical Society of America*, *129,* 2456.

Åsberg Johnels, J., Gillberg, C., Falck-Ytter, T., & Miniscalco, C. (2014). Face viewing patterns in young children with autism spectrum disorders: Speaking up for a role of language comprehension. *Journal of Speech, Language, and Hearing Research, 57,* 2246–2252.

Auer, E. T., Jr., & Bernstein, L. E. (1997). Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *The Journal of the Acoustical Society of America, 102,* 3704–3710.

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics, 62,* 233–252.

Boersma, P., & Weenink, D. (2004). *Praat: Doing phonetics by computer* (Version 4.2) [Computer Software]. Amsterdam, the Netherlands: Institute of Phonetic Science.

Buchan, J. N., & Munhall, K. G. (2012). The effect of a concurrent cognitive load task and temporal offsets on the integration of auditory and visual speech information. *Seeing and Perceiving, 25,* 87–106.

Buchan, J. N., Paré, M., & Munhall, K. G. (2005). The influence of task on gaze during audiovisual speech perception. *The Journal of the Acoustical Society of America, 115,* 2607.

Buchan, J. N., Paré, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience, 2,* 1–13.

Buchan, J. N., Paré, M., & Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Research, 1242,* 162–171.

Campbell, C., & Massaro, D. (1997). Perception of visible speech: Influence of spatial quantization. *Perception, 26,* 129–146.

Cienkowski, K. M., & Carney, A. E. (2002). Auditory-visual speech perception and aging. *Ear and Hearing, 23,* 439–449.

Cotton, J. C. (1935, December 20). Normal "visual hearing." *Science, 82,* 592–593.

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.

Dickinson, C. M., & Taylor, J. (2011). The effect of simulated visual impairment on speech-reading ability. *Ophthalmic & Physiological Optics, 31,* 249–257.

Enns, J. T., & MacDonald, S. C. (2013). The role of clarity and blur in guiding visual attention in photographs. *Journal of Experimental Psychology: Human Perception and Performance, 39,* 568–578.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research, 12,* 423–425.

Erber, N. P. (1971). Effects of distance on the visual reception of speech. *Journal of Speech and Hearing Research, 14,* 848–857.

Everdell, I. T., Marsh, H., Yurick, M. D., Munhall, K. G., & Paré, M. (2007). Gaze behaviour in audiovisual speech perception: Asymmetrical distribution of face-directed fixations. *Perception, 36,* 1535–1545.

Feld, J., & Sommers, M. (2009). Lipreading, processing speed, and working memory in younger & older adults. *Journal of Speech, Language, and Hearing Research, 52,* 1555–1565.

Gagné, J. P., Charbonneau, M., & Leroux, T. (2011). Speed of processing phonological information presented visually and speechreading proficiency. *Journal of the Academy of Rehabilitative Audiology, XLIV,* 11–27.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *Annals of Statistics, 13,* 70–84.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489–1506.

Jackson, P. L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. In C. L. De Filippo & D. G. Sims (Eds.), *The Volta Review: New Reflections on Speechreading* (Vol. 90) (pp. 99–115). Washington, DC: Alexander Graham Bell Association for the Deaf and Hard of Hearing.

Jiang, J., & Bernstein, L. E. (2011). Psychophysics of the McGurk and other audiovisual speech integration effects. *Journal of Experimental Psychology: Human Perception and Performance, 37,* 1193–1209.

Jordan, T., & Sergeant, P. (1998). Effects of facial image size on visual and audiovisual speech recognition. In F. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 155–176). London, England: Psychology Press.

Jordan, T., & Sergeant, P. (2000). Effects of distance on visual and audiovisual speech recognition. *Language and Speech, 43,* 107–124.

Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, and Hearing Research, 42,* 526–539.

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.

Latif, N., Gehmacher, A., Castelhano, M. S., & Munhall, K. G. (2014). The art of gaze guidance. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 33–39.

Legault, I., Gagné, J.-P., Rhoualem, W., & Anderson-Gosselin, P. (2010). The effects of blurred vision on auditory-visual speech perception in younger and older adults. *International Journal of Audiology, 49,* 904–911.

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

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 21,* 131–141.

MacLeod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology, 24,* 29–43.

McGurk, H., & MacDonald, J. (1976, December 23). Hearing lips and seeing voices. *Nature, 264,* 126–130.

Middelweerd, M. J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *The Journal of the Acoustical Society of America, 82,* 2145–2147.

Mirenda, P., Donnellan, A. M., & Yoder, D. E. (1983). Gaze behavior: A new look at an old problem. *Journal of Autism and Developmental Disorders, 13,* 297–309.

Munhall, K., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect *Perception and Psychophysics, 58,* 351–362.

Munhall, K. G., Kroos, C., Jozan, G., & Vatikiotis-Bateson, E., (2004). Spatial frequency requirements for audiovisual speech perception. *Perception & Psychophysics, 66,* 574–583.

Munhall, K. G., & Vatikiotis-Bateson, E. (2004). Spatial and temporal constraints on audiovisual speech perception. In G. Calvert, J. Spence, & B. Stein (Eds.), *Handbook of multisensory processing* (pp. 177–188). Cambridge, MA: MIT Press.

Neely, K. K. (1956). Effect of visual factors on the intelligibility of speech. *The Journal of the Acoustical Society of America, 28,* 1275–1277.

Nilsson, U. L., Frennesson, C., & Nilsson, S. E. G. (1998). Location and stability of a newly established eccentric retinal locus suitable for reading, achieved through training of patients with a dense central scotoma. *Optometry and Vision Science, 75,* 873–878.

O'Neill, J. J. (1954). Contributions of the visual component of oral symbols to speech comprehension. *Journal of Speech and Hearing Disorders, 19,* 429–439.

Paré, M., Richler, R. C., ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Perception & Psychophysics, 65,* 533–567.

Parkhurst, D., Law, K., & Neibur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42,* 107–123.

Rönnberg, J. (1995). What makes a skilled speechreader? In G. Plant & K. Spens (Eds.), *Profound deafness and speech communication* (pp. 393–416). London, England: Whurr.

Rönnberg, J., Arlinger, S., Lyxell, B., & Kinnefors, C. (1989). Visual evoked potentials: Relation to adult speechreading and cognitive function. *Journal of Speech and Hearing Research, 32,* 725–735.

Rosenblum, L. D., Johnson, J. A., & Saldana, H. M. (1996). Visual kinematic information for embellishing speech in noise. *Journal of Speech and Hearing Research, 39,* 1159–1170.

Saalasti, S., Kätsyri, J., Tiippana, K., Laine-Hernandez, M., von Wendt, L., & Sams, M. (2012). Audiovisual speech perception and eye gaze behavior of adults with Asperger syndrome. *Journal of Autism and Developmental Disorders, 42,* 1606–1615.

Small, L. H., & Infante, A. A. (1988). Effects of training and visual distance on speechreading performance. *Perceptual and Motor Skills, 66,* 415–418.

Strand, J., Cooperman, A., Rowe, J., & Simenstad, A. (2014). Individual differences in susceptibility to the McGurk effect: Links with lipreading and detecting audiovisual incongruity. *Journal of Speech, Language, and Hearing Research, 57,* 2322–2331.

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

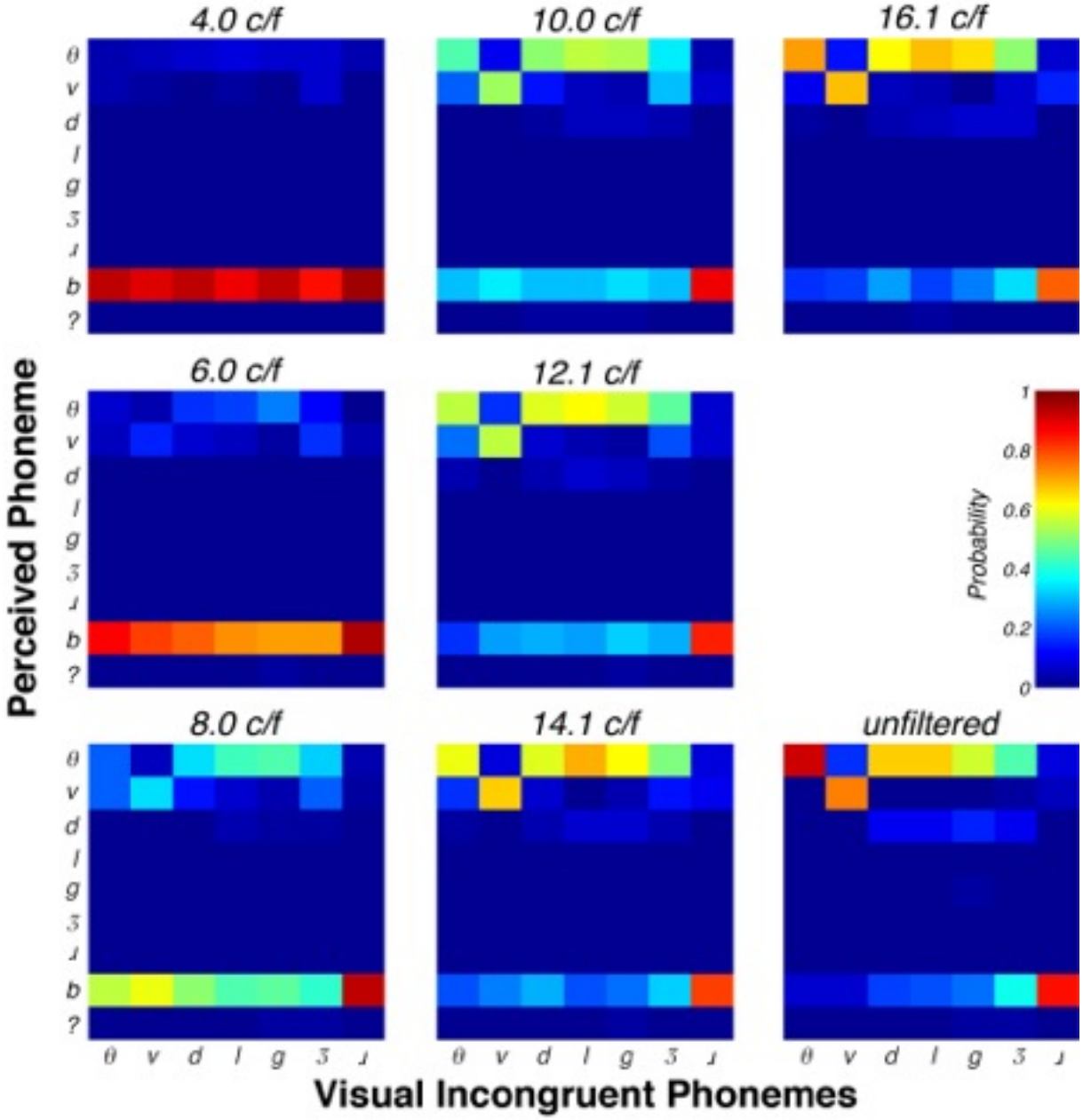
Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London B, 335,* 71–78.

Tiippana, K. (2014). What is the McGurk effect? *Frontiers in Psychology, 5,* 725.

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics, 60,* 926–940.

Wilson, A., Wilson, A., ten Hove, M., Paré, M., & Munhall, K. G. (2008). Loss of central vision and audiovisual speech perception. *Visual Impairment Research, 10,* 23–34.

Yi, A., Wong, W., & Eizenman, M. (2013). Gaze patterns and audiovisual speech enhancement. *Journal of Speech, Language, and Hearing Research, 56,* 471–480.


4.0 c/f
10.0 c/f
16.1 c/f
6.0 c/f
12.1 c/f
8.0 c/f
14.1 c/f
unfiltered
Perceived Phoneme
Visual Incongruent Phonemes
Probability
1
0.8
0.6
0.4
0.2
0
θ v d l g ʒ ɹ b ?