Perception & Psychophysics
2004, 66 (4), 574-583

Spatial frequency requirements for audiovisual speech perception

K. G. MUNHALL
Queen's University, Kingston, Ontario, Canada

C. KROOS and G. JOZAN
ATR Human Information Science Laboratories, Kyoto, Japan

and

E. VATIKIOTIS-BATESON
ATR Human Information Science Laboratories, Kyoto, Japan
and University of British Columbia, Vancouver, British Columbia, Canada

Spatial frequency band-pass and low-pass filtered images of a talker were used in an audiovisual speech-in-noise task. Three experiments tested subjects' use of information contained in the different filter bands with center frequencies ranging from 2.7 to 44.1 cycles/face (c/face). Experiment 1 demonstrated that information from a broad range of spatial frequencies enhanced auditory intelligibility. The frequency bands differed in the degree of enhancement, with a peak being observed in a mid-range band (11-c/face center frequency). Experiment 2 showed that this pattern was not influenced by viewing distance and, thus, that the results are best interpreted in object spatial frequency, rather than in retinal coordinates. Experiment 3 showed that low-pass filtered images could produce a performance equivalent to that produced by unfiltered images. These experiments are consistent with the hypothesis that high spatial resolution information is not necessary for audiovisual speech perception and that a limited range of spatial frequency spectrum is sufficient.

When auditory speech perception is impaired by noise or by a degraded acoustic signal, being able to see a talker's face saying the words significantly increases intelligibility. In one of the first experimental demonstrations of this effect, Sumby and Pollack (1954) showed that perception of spoken words was enhanced across a range of acoustic signal-to-noise ratios when subjects looked at a talker seated a few feet away. This visual contribution to speech perception has been demonstrated under a variety of conditions, including point-light displays (e.g., Rosenblum, Johnson, & Saldaña, 1996), animated faces (e.g., Massaro, 1998), and degraded visual images (Vitkovich & Barber, 1996). In this study, we extended this work by manipulating the spatial frequency content of the facial videos in order to investigate the nature of visual image processing during audiovisual speech perception.

Understanding how visual information is integrated with auditory speech information requires a detailed specification of the nature of the visual speech information, as well as the nature of visual information processing. In recent years, some progress has been made in both areas. Studies of facial kinematics during speech production have indicated that dynamic visual information must be low in temporal frequency (see Munhall & Vatikiotis-Bateson, 1998, for a review). For example, Ohala (1975) showed that the modal temporal frequency of jaw motion in continuous oral reading is below 5 Hz. Although speech may contain some high-frequency movement components, the time course of opening and closing of the vocal tract for syllables is relatively slow.

The primary locus of visual speech information is around the mouth and jaw, owing to their principal role in speech sound generation. However, the motion of articulation spreads across the entire face (Vatikiotis-Bateson, Munhall, Hirayama, Kasahara, & Yehia, 1996). These motions have been shown to correlate with the changing acoustic spectrum and RMS amplitude of the speech (Yehia, Rubin, & Vatikiotis-Bateson, 1998). From a perceptual standpoint, this distributed facial information has been shown to contribute significantly to intelligibility: the more of the face that is visible, the greater the intelligibility (e.g., Benoît, Guiard-Marigny, Le Goff, & Adjoudani, 1996).

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada, the NIH, the National Institute of Deafness and Other Communications Disorders (Grant DC-05774), and the Communication Dynamics Project, ATR Human Information Science Laboratories, Kyoto, Japan. The authors thank Clare MacDuffee, Amanda Rothwell, and Helen Chang for assistance in testing subjects. Martin Paré and Doug Shiller made helpful suggestions about an earlier draft. Correspondence concerning this article should be addressed to K. G. Munhall, Department of Psychology, Queen's University, Kingston, ON, K7L 3NC Canada (e-mail: munhallk@psyc.queensu.ca).

Note—This article was accepted by the previous editorial team, headed by Neil Macmillan.

People are adept at using this rich visual information, and when image quality is degraded, visual speech has been shown to be quite robust. In one of the first studies of this kind, Brooke and Templeton (1990) manipulated spatial resolution by varying the quantization level of a display showing the mouth region articulating English vowels. Varying the quantization level between 8 × 8 and 128 × 128 pixels, Brooke and Templeton found that performance on a silent lipreading task decreased when the display was reduced below 32 × 32 spatial resolution—for example, 16 × 16 pixels. Similar methods have been used by C. Campbell and Massaro (1997) and MacDonald, Andersen, and Bachmann (2000) to study the McGurk effect and by Vitkovich and Barber (1996) to study speechreading of digits. Both Campbell and Massaro and MacDonald et al. found that the McGurk effect persisted to some degree even at low spatial resolution. Vitkovich and Barber found no effect of their pixel density manipulation but did find changes in performance when the grayscale resolution (i.e., the number of gray levels) of the images was drastically reduced. Most recently, Thomas and Jordan (2002) used Gaussian blurring to study the importance of fine facial detail in visual speech and face processing. In a series of studies, they found that visual speech perception was not impaired until the blurring cutoff was 8 cycles per face or less.

All of these studies suggest that a visual contribution to speech perception does not require the extraction of high spatial resolution information from the facial image. The performance of subjects with point-light displays of visual speech and the influence of gaze during audiovisual speech are consistent with this interpretation. When the facial surface kinematics are reduced to the motions of a collection of light points, subjects still show perceptual benefit in an acoustically noisy environment (Rosenblum et al., 1996) and can perceive the McGurk effect (Rosenblum & Saldaña, 1996). Recently, Paré, Richler, ten Hove, and Munhall (2003) examined the distribution of gaze during the McGurk effect. Their results showed that the eye/mouth regions dominated the gaze patterns (see also Lansing & McConkie, 1999; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998); however, the strength of the McGurk effect was not influenced by where in the face region the subjects fixated. More important, Paré et al. demonstrated that the McGurk effect did not disappear when subjects directed their gaze well beyond the facial region. Thus, high-acuity foveal vision does not seem to be a requirement for audiovisual integration.

Although these studies suggest that the visual information used for speech perception can be relatively crude, there have been no direct tests of the range of spatial frequencies that are critical for audiovisual speech perception. It is widely accepted that the perception of visual images is carried out in the nervous system by a set of spatial-frequency–tuned channels. For the purpose of studying this processing, images can be decomposed into sets of nonoverlapping spatial frequency bands, and studies of the recognition of static objects in which such stimuli have been used indicated that these different spatial frequency bands do not contribute equally to the object identification process (e.g., Parish & Sperling, 1991). Perception of such objects as letters and faces appears to involve distinct processing of portions of the spatial frequency spectrum. This may be due to task- or domain-specific processing (Wenger & Townsend, 2000). For example, Vuilleumier, Armony, Driver, and Dolan (2003) reported that the amygdala carries out selective processing of the low spatial frequency portions of faces showing emotional expression, whereas the fusiform cortex, which is thought to be more involved in identification, is influenced more by the high-frequency portion of the facial image. Alternatively, selective processing could be determined by the properties of a stimulus. Solomon and Pelli (1994) and Majaj, Pelli, Kurshan, and Palomares (2002) have argued that the identification of text is mediated by a small portion of broadband spectrum—a single channel one to two octaves wide, determined by information in the fonts.

In this research, we manipulated the spatial resolution of facial images in order to understand the separate contributions of different spatial frequency components to visual speech processing. Previous work on static face processing (Gold, Bennett, & Sekuler, 1999; Näsänen, 1999) has suggested that optimal recognition of faces involves a relatively narrow band of spatial frequencies. Gold et al., for example, tested the discrimination of facial identity across a range of bandwidths and spatial frequency ranges. They observed peak efficiency of face processing with a two-octave band-pass filtered stimulus with a center frequency of 6.2 cycles/face (c/face; see Näsänen, 1999, for similar results).

Our extension of this approach to the perception of spoken language is motivated by two factors. First, there is behavioral evidence that static and dynamic facial images convey independent information (e.g., Lander, Christie, & Bruce, 1999) and neurological evidence that distinct neural substrates process static and dynamic visual speech (Calvert & R. Campbell, 2003; R. Campbell, Zihl, Massaro, Munhall, & Cohen, 1997; Munhall, Servos, Santi, & Goodale, 2002). Second, the spatial frequency range that is important for a given stimulus type may be tied to the particular task being carried out with the stimuli (Schyns & Oliva, 1999; Wenger & Townsend, 2000). In natural face-to-face conversations, a wide range of visual tasks are carried out in parallel. Listeners perceive information about the talker's identity, emotional and physical state, focus of attention, and degree of social and conversational engagement, as well as linguistic information about the utterance. Each of these aspects of communication may involve different stimulus properties, with information conveyed by different spatial frequency ranges (e.g., Vuilleumier et al., 2003).

The experiments reported here extended the study of the role of visual stimulus properties to dynamic facial images that conveyed linguistic information. Our experimental task, speech perception in noise, was quite se-

lective, and subjects had no explicit demands to monitor any information beyond the spoken words. Thus, the aim of the series of experiments was to characterize the visual information properties involved in a single subtask in communication, the visual enhancement of speech intelligibility. In Experiment 1, we used band-pass filtered facial images in a speech-in-noise task and compared performance of different bands against the full-video and auditory-only conditions. In Experiment 2, the different band-pass filtered images were tested at different viewing distances. In Experiment 3, low-pass filtered versions of the same audiovisual sentences as those used in Experiments 1 and 2 were tested. In combination, these three experiments allowed us to examine what spatial resolution is sufficient for audiovisual speech. The use of the band-pass filtered images permitted tests of the information contained within each spatial frequency band and provided evidence about the retinal specificity of the visual information used in visual speech. The low-pass filtered images provided a direct test of the sufficiency of low-frequency information in visual speech perception.

GENERAL METHOD

Stimuli

The Central Institute for the Deaf (CID) "everyday sentences" (Davis & Silverman, 1970) were spoken by a native American English speaker and were recorded on high-quality videotape (Betacam SP). For Experiments 1 and 2, the sentences were digitized as image sequences, converted to grayscale, and band-pass filtered using a two-dimensional discrete wavelet transformation (for a description of the application of wavelets to image processing, see Prasad & Iyengar, 1997; for details of the specific adaptation of the process used here, see Kroos, Kuratate, & Vatikiotis-Bateson, 2002). Five one-octave, band-pass filtered sets of sentence stimuli were created (see Table 1) with filter center frequencies ranging from 2.7 to 44.1 c/face.[1] Figure 1 shows single images from the different filter conditions. Note that the set of filters tested did not cover the complete range of spatial frequency components in the original images. The low-frequency cutoff was 1.8 c/face, and high-frequency cutoff was 59 c/face (Table 1).

The original audio track and a commercial multispeaker babble track (Auditec, St. Louis, MO) were recorded to Betacam SP videotape in synchrony with the original grayscale image sequence and with each of the five band-pass image sets. From the tape, a CAV videodisc was pressed for use in the perception experiments.

Equipment

The videodisc was played on a Pioneer (Model LD-V8000) videodisc player. Custom software was used to control the videodisc trials.

Table 1
Spatial Frequency Bands Used in Experiment 1

Bandwidth (c/face) Width	Center Frequency (c/face) Width	Center Frequency (cycles/deg of visual angle) Width
1.8–3.7	2.7	0.36
3.7–7.3	5.5	0.73
7.3–15	11.0	1.46
15–29	22.0	2.92
29–59	44.1	5.85

Scoring

Loose key word scoring (Bench & Bamford, 1979) was carried out on the subjects' responses, using the standard CID key words (Davis & Silverman, 1970). This scoring method ignores errors in inflection, because these errors compound for subsequent words (e.g., plural nouns and verb agreement). Percentage of key words correctly identified was used as the dependent variable in all the statistical analyses.

EXPERIMENT 1

Static face images can be recognized with only a small range of spatial frequencies. Although this fact has been demonstrated numerous times, a number of different frequencies have been proposed as the critical bands. Gold et al. (1999) reviewed a broad collection of studies reporting critical identification bands ranging from 1 to 25 c/face. This wide discrepancy is presumably due to task differences and discriminability of the faces in the test set. Recent studies that have taken into account such design limitations have reported that recognition of facial images and static facial expressions is best for narrow band (around two octaves), mid-to-low spatial frequency images (5–20 c/face; e.g., Näsänen, 1999). In the present experiment, we examined whether the perception of dynamic facial images during audiovisual speech shows the same sensitivity to band-pass filtered images.

Method

Subjects. Twenty-one native speakers of English were tested. All had normal or corrected-to-normal vision, and none reported hearing problems or a history of speech or language difficulties.

Stimuli. Seventy sentences from the full set of 100 CID sentences were selected for use in the perceptual testing. The CID sentence set is organized in groups of 10 sentences that are balanced for the number of key words, phonetic content, and average duration of the utterance. Seven of these sentence groups were used to test the seven audiovisual conditions (full image, auditory only, and five band-pass filter).

Equipment. The subjects watched the displays on a 20-in. video monitor (Sony PVM 1910). The acoustic signals were amplified and mixed (Tucker-Davis, System II) and were played through speakers (MG Electronics Cabaret) placed directly below the monitor. The testing took place in a double-walled sound isolation booth (IAC Model 1204).

Design. Each subject participated in a single session consisting of 70 sentences presented in noise. There were seven conditions (unfiltered video, no-video, and five band-pass filter conditions). The 70 sentences were divided into groups of 10 sentences with an equal number of key words. Each group of 10 sentences was assigned to one of the seven stimulus conditions for a subject. Assignment of groups of sentences to a condition was counterbalanced across subjects. The presentation of stimuli was randomized across condition within subjects.

Procedure. A small table and chair were positioned 171 cm from the monitor, and the subjects viewed the stimuli with the head position fixed using a chin- and headrest. After the presentation of a trial, the subjects verbally repeated as much of the sentence as they could. When the experimenter recorded the response, the next trial was initiated. The auditory signal-to-noise level was held constant for all the subjects. Pilot testing was used to find a level that produced auditory-only response accuracy below 50%. This was necessary to permit the significant increase in intelligibility that occurs when visual stimuli are present without ceiling effects.

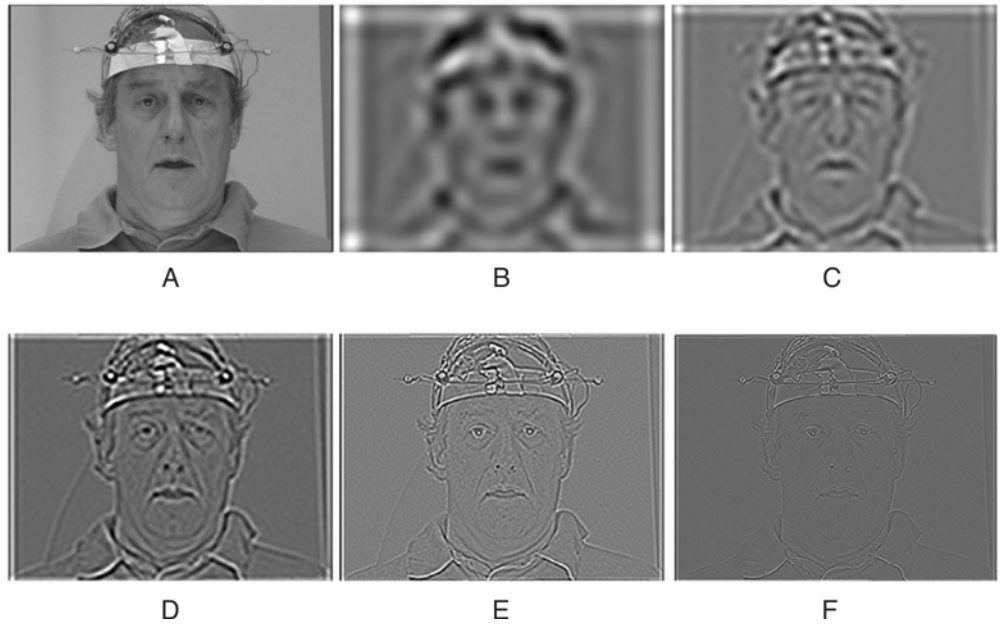

Figure 1. Images of the talker showing the different visual conditions used in Experiments 1 and 2: (A) full video, (B) 2.7-c/face center frequency, (C) 5.5-c/face center frequency, (D) 11-c/face center frequency, (E) 22-c/face center frequency, and (F) 44.1-c/face center frequency.

Results and Discussion

The percentage of key words correctly identified was used as the dependent variable. As can be seen in Figure 2, the average performance differed across the audiovisual conditions. The greatest number of key words were identified with the unfiltered visual stimuli, whereas the poorest performance was observed for auditory only. The spatial frequency filtered bands showed an inverted U-shaped performance curve with a peak in accuracy in the images with 11-c/face center frequency. A one-way analysis of variance (ANOVA) showed a main effect for audiovisual condition [$F(6,120) = 19.57$, $p < .001$]. Tukey's HSD post hoc tests revealed that the full-face video condition had a significantly ($p < .05$) higher number of key words identified than did all the conditions except the 11-c/face band ($p = .08$). The latter condition was also found to be different from the unfiltered face with additional statistical power.[2] Furthermore, all of the audiovisual conditions, with the exception of two spatial frequency filters (2.7 and 22 c/face), showed significantly higher intelligibility than did the auditory-only condition ($p < .05$).

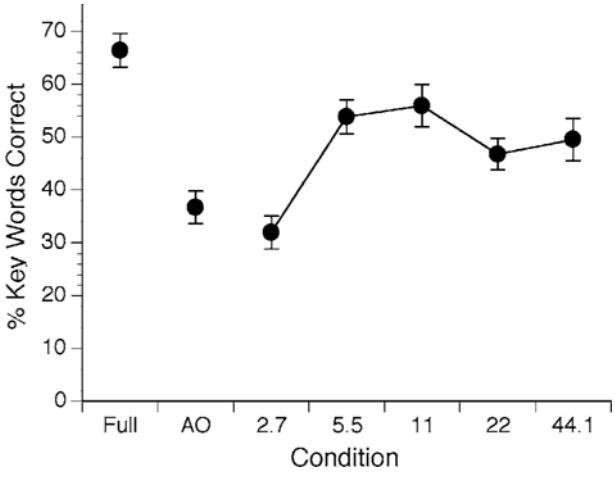

Figure 2. Percentage of key words correctly identified in Experiment 1 as a function of the spatial frequency bands (2.7-, 5.5-, 11-, 22-, and 44.1-c/face center frequencies). The unfiltered full-face and auditory-only (AO) conditions are also shown. The error bars show the standard errors of the means.

When the performances in the filter conditions were analyzed separately, a significant quadratic trend was observed ($p < .01$). The observed peak (11-c/face center frequency) corresponds well with Näsänen's (1999) 8- to 13-c/face maximum sensitivity and Costen, Parker, and Craw's (1996) 8- to 16-c/face critical spatial frequency range for static faces.

These results show that visual enhancement of speech perception in noise occurs across a range of spatial frequencies. All but two of the spatial frequency bands that we tested exceeded performance in the auditory-only condition. On the other hand, the full-face condition exceeds the accuracy level of all of the band-pass conditions, suggesting that the linguistic information contained within any single band is incomplete. Either the summing of some of the filter bands or the broadening of the bandwidth may raise performance to the level of the full-face condition. The filtering in the present experiment was carried out with a one-octave bandwidth.[3] Previous stud-

ies have shown that optimal performance for static face recognition can be found with a bandwidth close to two octaves (Gold et al., 1999; Näsänen, 1999). If audiovisual speech perception has the same optimal bandwidth, its use might have reduced differences between the full-face image and the band-pass conditions. We will return to this issue in Experiment 3.

One curious difference between Gold et al.'s (1999) studies of static face identity and the present dynamic speech perception experiment is that the one-octave stimuli used here produced performance that approached that of the unfiltered stimuli, whereas Gold et al.'s subjects found their one-octave stimuli impossible to discriminate. Although there are differences in task and stimulus dimensions (e.g., identity vs. speech information) between the two studies, the discrepancy in performance for the one-octave stimuli is more likely due to the presence of motion in the present experiment. Motion influences object perception in a number of ways. Motion can provide additional cues to shape, either through motion-defined form or simply by providing multiple viewpoints for the face. Independently, the facial kinematics can specify the spatiotemporal signature for an individual that can be used to perceive identity or gender, as well as be perceived itself as a form of biological motion (e.g., Hill & Johnston, 2001; Stone, 1998). The independence of facial motion information from form cues has been demonstrated with point-light audiovisual speech (Rosenblum et al., 1996) and preserved visual speech perception in a case of visual form agnosia (Munhall et al., 2002). Knappmeyer, Thornton, and Bülthoff (2003), on the other hand, have demonstrated that facial motion and facial form information are integrated during identity judgments. The study of static face processing is thus a highly impoverished condition removed from the natural context of dynamic object perception.

Despite this difference in the effectiveness of the one-octave bandwidth, the spatial frequency effects for static studies of identity and emotion discrimination show a remarkable similarity to the present dynamic study of facial behavior. For both dynamic and static conditions, very low frequencies seem less useful than high frequencies (see also Gold et al., 1999), and peak performance is observed in the low–mid spatial frequency range. This similarity of performance across different stimuli and tasks suggests either that visual stimulus properties of faces determine performance for all of these tasks in a bottom-up fashion or that the pattern of performance reflects a property of the visual system that is implicated in all facial processing (cf. Majaj et al., 2002).

EXPERIMENT 2

One approach to examining the relative contributions of the visual system and the object properties to perceptual performance is to vary the size of or the viewing distance to the object. In this experiment, we contrasted spatial frequency defined with reference to the object (rate of change of contrast per face) with spatial frequency defined in terms of retinal coordinates (rate of change of contrast per degree of visual angle). We did so by manipulating viewing distance to the display over a 3:1 range. Because image size on the retina is inversely proportional to viewing distance, retinal spatial frequency will vary proportionally with distance to the display (see Table 2). If the spatial frequency sensitivity in Experiment 1 is determined by retinal frequency, this viewing distance manipulation should shift the peak of the intelligibility function.

For a range of different stimuli, it has been reported that visual perception is scale invariant. When viewed from different distances, performance on the detection of sine wave gratings (e.g., Banks, Geisler, & Bennett, 1987), static face recognition (Näsänen, 1999), priming of face recognition (Brooks, Rosielle, & Cooper, 2002), letter identification (Parish & Sperling, 1991), and object recognition (Tjan, Braje, Legge, & Kersten, 1995) is quite stable. This remarkable scale stability for static object recognition is also found in dynamic audiovisual speech perception. Jordan and Sergeant (2000) manipulated distance from 1 to 30 m in an audiovisual task and found that vision improved performance with congruent auditory speech at all distances. In this experiment, we examined whether this scale invariance would extend to band-pass filtered images and, thus, whether subjects consistently use object-centered information to perceive spoken words.

Method

Subjects. Ninety individuals served as subjects. All were native speakers of English, had normal or corrected-to-normal vision, and reported no hearing problems and no history of speech or language difficulties.

Stimuli. Ninety CID sentences were used as stimuli. In order to test five filter conditions at three viewing distances, these sentences were divided into 15 groups of six sentences. The number of key words was approximately equal for each group.

Equipment. The subjects watched the displays on a 20-in. video monitor (Quasar QC-20H20R). The acoustic signals were amplified and mixed (Mackie Micro Series 1202-VLZ mixer) and were played through headphones (Sennheiser HD265). Because of the separation required between the monitor and the subject, the testing took place in a quiet laboratory room, rather than in a sound booth.

Design. Each subject participated in a single session consisting of 90 sentences presented in noise. There were five band-pass filtered conditions and three viewing distances. The 90 sentences

Table 2
Spatial Frequency Bands Used in Experiment 2

Center Frequency (c/face) Width	114-cm Viewing Distance (cycles/deg visual angle)	228-cm Viewing Distance (cycles/deg visual angle)	342-cm Viewing Distance (cycles/deg visual angle)
2.7	0.32	0.65	0.98
5.5	0.66	1.33	1.98
11.0	1.32	2.65	3.98
22.0	2.65	5.31	7.96
44.1	5.32	10.64	15.96

were divided into 15 groups of 6 sentences, with an approximately equal number of key words. Each group of 6 sentences was assigned to one of the five band-pass filter conditions at a viewing distance for a subject. The subjects were tested at each viewing distance separately. Order of viewing distance was counterbalanced across subjects. Assignment of a group of sentences to a condition was counterbalanced across subjects so that every sentence occurred in each condition combination (band-pass filter × viewing distance × order of viewing distance). The presentation of stimuli was randomized across band-pass filter condition within subjects within a viewing distance.

Procedure. The subjects watched the monitor from three viewing distances (114, 228, and 342 cm) with their head in a headrest. Table 2 shows the center frequencies in object and retinal coordinates. In order to equate listening conditions, the auditory stimuli were presented through headphones. The auditory signal-to-noise level was held constant for all the subjects. Pilot testing was used to find a level that produced auditory-only response accuracy below 50%. As in Experiment 1, the presentation of stimuli was subject paced. The subjects verbally repeated as much of each sentence as they could. When the experimenter recorded the response, the next trial was initiated.

Results and Discussion

As in Experiment 1, an inverted U-shaped function is present in the data, with a peak in intelligibility again being observed for the 11-c/face filter condition. In general, the three viewing distances did not influence the pattern of results, with all viewing distances showing the similar percentages of key words perceived for the different spatial frequency bands (see Figure 3).

A 3 × 5 ANOVA was carried out to test the effect of viewing distance (114, 228, and 342 cm) and band-pass filter (2.7-, 5.4-, 11-, 22-, and 44.1-c/face center frequency). As can be seen in Figure 3, no effect of distance was observed [$F(2,178) = 0.66, p > .5$]; however, the band-pass filters significantly altered the intelligibility of the speech [$F(4,356) = 72.7, p < .001$]. A significant quadratic trend was observed ($p < .001$) in these data. A small filter × distance interaction was present [$F(12,712) = 2.01, p < .05$]. The performance for the high-frequency filter (44.1 c/face) was significantly worse at 342 cm than at the average of the other two distances.

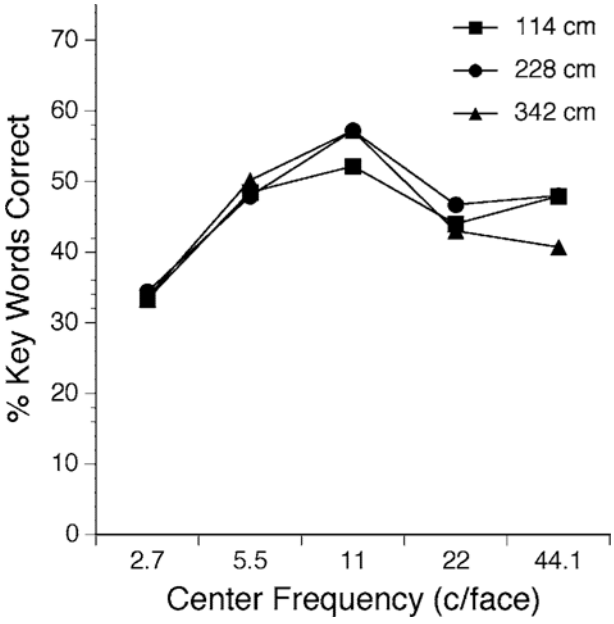

Figure 3. Percentage of key words correctly identified in Experiment 2 as a function of the five spatial frequency bands (2.7-, 5.5-, 11-, 22-, and 44.1-c/face center frequencies) and the three viewing distances (114, 228, and 342 cm).

The present findings are consistent with studies of static faces that have indicated that object spatial frequency, rather than retinal spatial frequency, best accounts for the perceptual results (faces, Hayes, Morrone, & Burr, 1986, and Näsänen, 1999; letters, Parish & Sperling, 1991; cf. Majaj et al., 2002). This presumably accounts for the ability of subjects to perceive audiovisual speech (Jordan & Sergeant, 2000) and letters (Legge, Pelli, Rubin, & Schleske, 1985) over a large range of viewing distances. Extracting speech information by using a face-based coordinate system has ecological advantages, since the spatial range of conversation varies considerably.

As in Experiment 1, the intelligibility function was not completely symmetrical around the 11-c/face band. For two of the viewing distances (114 and 228 cm), the high-frequency band showed better performance than did the corresponding low-frequency condition. For the furthest viewing distance, high-frequency performance dropped off significantly. This drop-off was presumably due to optical attenuation of the high spatial frequencies with increased distance (see Näsänen, 1999, for similar changes in sensitivity to high-frequency bands with distance for static facial images).

EXPERIMENT 3

The final experiment was motivated by two related observations: (1) In the first experiment, the band-pass filtered stimuli did not produce performance equal to the unfiltered facial images, and (2) the observed intelligibility function for the band-pass filtered images in Experiments 1 and 2 was not symmetrical around the peak at 11 c/face. One explanation for these observations is that visual speech information is distributed across the individual spatial frequency bands. This could occur because the optimal bandwidth for visual speech is actually wider than the one-octave bands that we tested (cf. Gold et al., 1999; Näsänen, 1999) or because nonredundant phonetic information is distributed across the spatial frequency spectrum. In the latter case, performance for the higher frequency bands may have been better than that for the lower frequency bands because the filter bandwidth is proportional to frequency. High frequency bands would contain more frequencies than would lower bands and, thus, perhaps more information. On the other hand, it may be that the high spatial frequency band captures important phonetic information that is not reflected in bands below its lower cutoff (29 c/face). C. Campbell and Massaro's (1997) and MacDonald et al.'s (2000) data

showed a monotonic increase in performance with each increase in level of quantization. Although the slope of this function decreased markedly for higher levels of quantization, these results support the idea that the higher frequency bands carry some unique information that aids visual speech.

In the present experiment, we addressed these issues by testing the same recordings and task as those used in the first two experiments with low-pass, rather than band-pass, spatial frequency filtered images. By doing this, we directly tested whether lower frequency information is sufficient for visual speech processing (e.g., MacDonald et al., 2000) and indirectly tested whether the optimal bandwidth for visual speech is greater than the one-octave bandwidth used in the first two experiments. Previous work has supported the idea that subjects can reach maximum performance with low-pass filtered stimuli; however, it is unclear to what extent C. Campbell and Massaro's (1997) and MacDonald et al.'s findings showing high-frequency contributions are due to the use of quantization rather than other forms of image filtering. Quantization introduces spurious high-frequency information, and the accuracy function shown in those articles may have been due in part to release from the effects of this high-frequency artifact with increasing quantization, rather than to the information below the cutoff frequency. This interpretation is supported by Thomas and Jordan's (2002) findings.

Method

Subjects. Two groups of 24 subjects (48 in total) served as subjects. All were native speakers of English, had normal or corrected-to-normal vision, and reported no hearing problems and no history of speech or language difficulties. Each of the two groups of subjects was tested at a different auditory signal-to-noise level.

Stimuli. Eighty CID sentences were used as stimuli. The same recordings as those used in Experiments 1 and 2 were processed in a different manner for this study. The video was digitized as a sequence of images, converted to grayscale, and low-pass filtered using rotationally symmetric Gaussian filters (for details of the process used here, see Jozan, 2001). Six low-pass filtered sets of sentence stimuli were created (see Table 1), with cutoff frequencies ranging from 1.8 to 59 c/face.

The original audio track and a commercial multispeaker babble track (Auditec, St. Louis, MO) were recorded to Betacam SP videotape in synchrony with the original grayscale image sequence and with each of the six low-pass image sets. From the tape, a CAV videodisc was pressed for use in the perception experiments.

Equipment. The subjects watched the displays on a 20-in. video monitor (Sharp Aquos). The acoustic signals were amplified and mixed (Tucker-Davis, System II) and were played through speakers (MG Electronics Cabaret) placed directly below the monitor. The testing took place in a double-walled sound isolation booth (IAC Model 1204).

Design. Each subject participated in a single session consisting of 80 sentences presented in noise. These 80 stimuli were divided into groups of 10 sentences, with each group of 10 sentences assigned to one of the eight stimulus conditions for a subject (unfiltered video, auditory only, and the six low-pass conditions). Assignment of groups of sentences to a condition was counterbalanced across subjects. The presentation of stimuli was randomized across conditions within subjects. A second set of subjects carried out the same experiment at a different signal-to-noise ratio.

Procedure. A small table and chair were positioned 171 cm from the monitor, and the subjects viewed the stimuli with the head position fixed using a chin- and headrest. After the presentation of a trial, the subjects verbally repeated as much of the sentence as they could. When the experimenter recorded the response, the next trial was initiated.

Two different auditory signal-to-noise levels were chosen for the two groups of subjects. For each group, the signal-to-noise level was held constant for all the subjects. As in the previous experiments, pilot testing was used to find noise levels that produced two different auditory-only response accuracies below 50%.

Results and Discussion

The different signal-to-noise levels resulted in distinct levels of performance. The mean percentages of key words correct averaged across conditions for the two groups of subjects were 44.7% and 59.1%. However, both groups showed the same pattern across the different audiovisual conditions. Figure 4 shows mean percentages of key words correct for the eight audiovisual conditions averaged across the two signal-to-noise levels. As can be seen, all but the lowest spatial frequency filter condition showed visual enhancement over the auditory-only condition. Significantly, the performance in the low-pass conditions asymptotes with the condition composed of images filtered with a 7.3-c/face cutoff frequency. For this condition and conditions with higher cutoffs, performance equaled that found for the unfiltered images.

A 2×8 ANOVA showed these patterns to be statistically reliable. Main effects were found for signal-to-noise level [$F(1,46) = 11.25$, $p < .01$] and audiovisual condition [$F(7,322) = 35.7$, $p < .01$], with no significant interaction. Tukey's HSD post hoc tests showed that all the audiovisual conditions, with the exception of the lowest spatial frequency condition (1.8 c/face), were reliably better than the auditory-only condition ($p < .05$) and that there was no significant difference between the unfiltered facial stimuli and the stimuli with cutoffs of 7.3, 15, 29, and 59 c/face ($p > .1$).

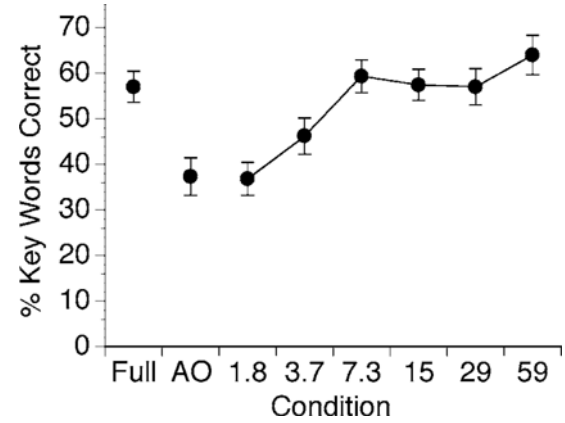

Figure 4. Percentage of key words correctly identified in Experiment 3 as a function of the low-pass filter cutoffs (1.8, 3.7, 7.3, 15, 29, and 59 c/face). The unfiltered full-face and auditory-only (AO) conditions are also shown. The error bars show the standard errors of the means.

These data help to clarify the interpretation of the first two experiments. The asymptotic performance at a relatively low spatial frequency suggests that the high spatial frequency speech information either was redundant with that contained in other bands or, at the very least, is not useful in this context. Lower spatial frequency information is sufficient to equate performance with the unfiltered condition. This raises the strong possibility that a broader bandwidth (e.g., Gold et al., 1999) would have raised the accuracy level in Experiments 1 and 2 for the optimal band to equal performance in the unfiltered images.

GENERAL DISCUSSION

In all the experiments, the intelligibility of speech in noise varied as a function of the spatial frequency content of the accompanying video images. Experiment 1 showed that all but the lowest spatial frequency band that we tested enhanced auditory speech perception; however, none of the individual spatial frequency bands reached the accuracy level of the unfiltered images. The band-pass conditions showed a quadratic intelligibility pattern, with peak intelligibility occurring in the mid-range filter band with a center frequency of 11 c/face. Experiment 2 showed that this pattern did not vary as a function of viewing distance and, thus, that object-based spatial frequency best characterized the data. Experiment 3 indicated that a summation of the lower frequency bands provided all of the information required for audiovisual speech.

The visual enhancement of intelligibility even from the 5.5-c/face center frequency band is consistent with studies using quantized facial images for audiovisual speech perception (C. Campbell & Massaro, 1997; MacDonald et al., 2000). Low-frequency images in those studies and in the present experiments influenced auditory perception. This suggests that listeners do not need to foveate on the mouth to acquire visual phonetic information. Studies of gaze during audiovisual speech have indicated that listeners scan the eyes and mouth in a stereotyped manner (Paré et al., 2003; Vatikiotis-Bateson et al., 1998). Our findings predict that, irrespective of the location of gaze, audiovisual speech perception will not vary, and indeed, Paré et al. have demonstrated that this is the case. Whether subjects fixated on the eyes or the mouth, they showed the same level of McGurk effect.

The band-pass stimuli show a maximum enhancement of intelligibility for the band with a center frequency of 11 c/face. This corresponds with values found for static face identification and facial expression discrimination (e.g., Näsänen, 1999), suggesting a commonality between these disparate perceptual processes. This contrasts with previous research supporting a distinction between the perceptual processes and neural architectures responsible for face identification and visual speech processing (e.g., Bruce & Young, 1986). The present findings are more consistent with data indicating that facial speech processing and face identification are not independent (Walker, Bruce, & O'Malley, 1995). In face-to-face communication, there may be advantages to using the same set of spatial frequencies for multiple face-related tasks. It allows resources to be focused on a limited scope of information, and by rigidly restricting the processing bandwidth, an efficient automated skill can evolve. Pelli and colleagues have demonstrated such "rigid" processing in the selective spatial frequency processing of letters (Majaj et al., 2002; Solomon & Pelli, 1994; cf. Schyns & Oliva, 1999) and the use of letter versus word processing of written text (Pelli, Farell, & Moore, 2003).

Whether the spatial frequency patterns are the result of the information distribution within the face itself (object statistics) or a property of the visual system cannot be directly tested here.[4] One technique that has been used for face (Gold et al., 1999) and letter (Parish & Sperling, 1991) perception is the calculation of the efficiency of the subjects' perceptual processing. This involves comparing the performance of subjects with that of an *ideal observer* (e.g., Geisler, 1989) to see whether the human subjects are using all of the available information. Given the dynamic nature of the speech stimuli and the open-set identification task potentially drawing on the full lexicon, this is not a tractable approach for audiovisual speech. However, ideal observer analysis for static face identification indicates that human observers do not optimally use all of the information available in those facial images, and thus the peak in sensitivity associated with a mid-range spatial frequency band is not replicated in the ideal observer's behavior (e.g., Gold et al., 1999). Since the observed spatial frequency sensitivity curves appear to be similar between static faces and dynamic speech information, we think it likely that our results also do not reflect a data limitation. Indeed, statistical analysis of the facial kinematics during speech indicates a remarkably rich information structure (Yehia et al., 1998) that is strongly correlated with the time-varying acoustic spectrum of speech.

What do these findings suggest about the process of audiovisual speech perception? In practical terms, the data indicate that viewing distance and gaze are not tightly constrained by the requirements of visual speech perception. During natural conversations, gaze has many social and information-gathering functions. Direction of gaze signals turn taking and social attention. Objects and gestures attract the gaze away from the face, and a lateral reflective gaze accompanies some cognitive processing. Our data and the recent findings of Paré et al. (2003) suggest that visual speech processing can tolerate considerable deviations in fixations beyond the mouth region and, thus, that speech visual processing might not be significantly impaired by the parallel uses of gaze during conversation.

Lansing and McConkie (1999) have recently proposed that subjects foveate the region of the face that carries the best information for a task. This *gaze direction assumption* predicts that subjects will fixate on the mouth

for tasks requiring segmental perception and will fixate on other regions of the face during prosodic and emotion judgments. Although their data are consistent with this assumption, it is likely that their particular task, silent speechreading within a laboratory setting, influenced their findings. Different distributions of gaze have been reported by Vatikiotis-Bateson et al. (1998) and Paré et al. (2003) for different tasks. The determinants of these patterns of gaze during speech perception are likely to be a combination of oculomotor factors and cognitive processing strategies. However, it is an entirely separate issue whether foveation on specific facial areas is *necessary* for audiovisual speech perception or even helpful when it does occur (see the discussion above).

The conditions that promote integration of auditory and visual information in speech are not completely understood. The integration of simpler nonverbal signals appears to involve synchronous stimulation across the modalities (e.g., Calvert, Hansen, Iversen, & Brammer, 2001), but behavioral evidence suggests that this is not the case for speech (e.g., Grant & Greenberg, 2001; Munhall, Gribble, Sacco, & Ward, 1996). One possible solution that has been proposed is that the dynamics of articulation provide a temporal signature in both visual and auditory speech that is important for cross-modal integration. For example, slow changes in the acoustics that reflect the syllabic alternation of opening and closing the vocal tract may directly correspond to the visual kinematics (Greenberg & Arai, 2001; Munhall et al., 1996; Remez, Fellowes, Pisoni, Goh, & Rubin, 1998; Yehia et al., 1998). By this view, the tracking of visual motion is the basis of audiovisual integration.

Facial motion is a combination of rigid motion of the head and nonrigid deformation of the face. Whereas head motion is strongly associated with prosody (Nicholson, Baum, Cuddy, & Munhall, 2001; Yehia, Kuratate, & Vatikiotis-Bateson, 2002), the soft tissue deformations of the mouth and face provide segmental phonetic information. Most research on motion perception has focused on rigid motion (see Lu & Sperling, 2001), and work on basic processes in the perception of shape deformation is only beginning (e.g., Loffler & Wilson, 2001). Since spatial-frequency–tuned channels within the visual system play an important role in motion perception, determining the contribution of these two classes of motion to speech perception will require studies that systematically control spatial and temporal frequency of the facial images.

REFERENCES

Banks, M., Geisler, W., & Bennett, P. (1987). The physical limits of grating visibility. *Vision Research*, **27**, 1915-1924.

Bench, J., & Bamford, J. M. (Eds.) (1979). *Speech-hearing tests and the spoken language of hearing-impaired children.* London: Academic Press.

Benoît, C., Guiard-Marigny, T., Le Goff, B., & Adjoudani, A. (1996). Which components of the face do humans and machines best speechread? In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines: Models, systems, and applications* (pp. 315-328). Berlin: Springer-Verlag.

Brooke, N. M., & Templeton, P. D. (1990). Visual speech intelligibility of digitally processed facial images. *Proceedings of the Institute of Acoustics*, **12**, 483-490.

Brooks, B. E., Rosielle, L. J., & Cooper, E. E. (2002). The priming of face recognition after metric transformations. *Perception*, **31**, 297-313.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, **77**, 305-327.

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

Calvert, G. A., Hansen, P. C., Iversen, S. D., & Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *NeuroImage*, **14**, 427-438.

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

Campbell, R., Zihl, J., Massaro, D., Munhall, K., & Cohen, M. M. (1997). Speechreading in the akinetopsic patient, L.M. *Brain*, **120**, 1793-1803.

Costen, N. P., Parker, D. M., & Craw, I. (1996). Effects of high-pass and low-pass spatial filtering on face identification. *Perception & Psychophysics*, **58**, 602-612.

Davis, H., & Silverman, S. R. (Eds.) (1970). *Hearing and deafness* (3rd ed.). New York: Holt, Rinehart & Winston.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review*, **96**, 267-314.

Gold, J., Bennett, P., & Sekuler, A. (1999). Identification of band-pass filtered letters and faces by human and ideal observers. *Vision Research*, **39**, 3537-3560.

Grant, K., & Greenberg, S. (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. In *Proceedings of AVSP-01* (pp. 132-137). Scheelsminde, Denmark.

Greenberg, S., & Arai, T. (2001). The relation between speech intelligibility and the complex modulation spectrum. In *Proceedings of the 7th European Conference on Speech Communication and Technology* (Eurospeech-2001, pp. 473-476). Aalborg, Denmark.

Hayes, T., Morrone, M. C., & Burr, D. C. (1986). Recognition of positive and negative bandpass-filtered images. *Perception*, **15**, 595-602.

Hill, H., & Johnston, A. (2001). Categorizing sex and identity from the biological motion of faces. *Current Biology*, **11**, 880-885.

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

Jozan, G. (2001). Analysis of talking faces: Tools for filtering and feature tracking (ATR Tech. Rep. No. TR-HIS-0004, pp. 1-18). Kyoto: ATR Laboratories.

Knappmeyer, B., Thornton, I., & Bülthoff, H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research*, **43**, 1921-1936.

Kroos, C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Video-based face motion measurement. *Journal of Phonetics*, **30**, 569-590.

Lander, K., Christie, F., & Bruce, V. (1999). The role of movement in the recognition of famous faces. *Memory & Cognition*, **27**, 974-985.

Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, & Hearing Research*, **42**, 526-538.

Legge, G. E., Pelli, D. G., Rubin, G. S., & Schleske, M. M. (1985). Psychophysics of reading: I. Normalization vision. *Vision Research*, **25**, 239-252.

Loffler, G., & Wilson, H. R. (2001). Detecting shape deformation of moving patterns. *Vision Research*, **41**, 991-1006.

Lu, Z.-L., & Sperling, G. (2001). Three-systems theory of human visual motion perception: Review and update. *Journal of the Optical Society of America A*, **18**, 2331-2370.

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

Majaj, N. J., Pelli, D. G., Kurshan, P., & Palomares, M. (2002). The role of spatial frequency channels in letter identification. *Vision Research*, **42**, 1165-1184.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

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

Munhall, K. G., Servos, P., Santi, A., & Goodale, M. (2002). Dynamic visual speech perception in a patient with visual form agnosia. *NeuroReport*, **13**, 1793-1796.

Munhall, K. G., & Vatikiotis-Bateson, E. (1998). The moving face during speech communication. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye: Pt. 2. The psychology of speechreading and audiovisual speech* (pp. 123-139). London: Taylor & Francis, Psychology Press.

Näsänen, R. (1999). Spatial frequency bandwidth used in the recognition of facial images. *Vision Research*, **39**, 3824-3833.

Nicholson, K. G., Baum, S., Cuddy, L. L., & Munhall, K. G. (2001). A case of multimodal aprosodia: Impaired auditory and visual speech prosody perception in a patient with right hemisphere damage. In *Proceedings of AVSP-01* (pp. 62-65). Scheelsminde, Denmark.

Ohala, J. J. (1975). The temporal regulation of speech. In G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 431-453). London: Academic Press.

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

Parish, D. H., & Sperling, G. (1991). Object spatial frequencies, retinal spatial frequencies, noise, and the efficiency of letter discrimination. *Vision Research*, **31**, 1399-1415.

Pelli, D. G., Farell, B., & Moore, D. C. (2003). The remarkable inefficiency of word recognition. *Nature*, **423**, 752-756.

Prasad, L., & Iyengar, S. S. (1997). *Wavelet analysis with applications to image processing*. Boca Raton, FL: CRC Press.

Remez, R. E., Fellowes, J. M., Pisoni, D. B., Goh, W. D., & Rubin, P. E. (1998). Multimodal perceptual organization of speech: Evidence from tone analogs of spoken utterances. *Speech Communication*, **26**, 65-73.

Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech & Hearing Research*, **39**, 1159-1170.

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

Schyns, P., & Oliva, A. (1999). Dr. Angry and Mr. Smile: When categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition*, **69**, 243-265.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, **24**, 1193-1216.

Solomon, J. A., & Pelli, D. G. (1994). The visual filter mediating letter identification. *Nature*, **369**, 395-397.

Stone, J. V. (1998). Object recognition using spatiotemporal signatures. *Vision Research*, **38**, 957-951.

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

Thomas, S. M., & Jordan, T. R. (2002). Determining the influence of Gaussian blurring on inversion effects with talking faces. *Perception & Psychophysics*, **64**, 932-944.

Tjan, B. S., Braje, W. L., Legge, G. E., & Kersten, D. (1995). Human efficiency for recognizing 3-D objects in luminance noise. *Vision Research*, **35**, 3053-3069.

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.

Vatikiotis-Bateson, E., Munhall, K. G., Hirayama, M., Kasahara, Y., & Yehia, H. (1996). Physiology-based synthesis of audiovisual speech. In *Proceedings of the 4th Speech Production Seminar: Models and Data* (pp. 241-244). Autrans, France.

Vitkovich, M., & Barber, P. (1996). Visible speech as a function of image quality: Effects of display parameters on lipreading ability. *Applied Cognitive Psychology*, **10**, 121-140.

Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2003). Distinct spatial frequency sensitivities for processing faces and emotional expressions. *Nature Neuroscience*, **6**, 624-631.

Walker, S., Bruce, V., & O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics*, **57**, 1124-1133.

Wenger, M. J., & Townsend, J. T. (2000). Spatial frequencies in short-term memory for faces: A test of three frequency-dependent hypotheses. *Memory & Cognition*, **28**, 125-142.

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, **30**, 555-568.

Yehia, H. C., Rubin, P. E., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, **26**, 23-44.

NOTES

1. The spatial frequency values used throughout the article are based on image width measures.

2. The lack of statistical difference ($p = .08$) between the unfiltered face and the 11-c/face band could be due to a lack of power or simply to the fact that there is no real difference between these conditions. Given the great intersubject variability found in speech-in-noise experiments, we strongly suspected the former explanation. We tested an additional group of 21 subjects and found exactly the same pattern of results as that found in Experiment 1. The combined data for the two groups of subjects showed that the 11-c/face band was reliably lower in intelligibility than the unfiltered face by Tukey's HSD ($p < .01$). For the remainder of this article, we will treat this unfiltered face as showing better performance than do all of the filtered conditions.

3. The stimuli used in this experiment are a by-product of a video-based face-tracking project (Kroos et al., 2002). The one-octave bandwidth was determined by the requirements of that work, rather than being the optimal bandwidth for the perception of facial stimuli. Since the tracking system becomes less accurate at the highest spatial frequencies, an important validation of the adequacy of that system was to show that linguistically relevant speech motion is recoverable at medium and lower frequencies.

4. Of course, the object statistics and properties of the visual system might be equivalent (see Simoncelli & Olshausen, 2001).

(Manuscript received October 21, 2001;
revision accepted for publication September 25, 2003.)