Available online at www.sciencedirect.com

SCIENCE DIRECT®

Vision Research 46 (2006) 3243–3258

Vision Research

www.elsevier.com/locate/visres

An ideal observer analysis of variability in visual-only speech

Brianna Conrey, Jason M. Gold *

Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405, USA

Received 8 September 2005; received in revised form 16 February 2006

Abstract

Normal-hearing observers typically have some ability to "lipread," or understand visual-only speech without an accompanying auditory signal. However, talkers vary in how easy they are to lipread. Such variability could arise from differences in the visual information available in talkers' speech, human perceptual strategies that are better suited to some talkers than others, or some combination of these factors. A comparison of human and ideal observer performance in a visual-only speech recognition task found that although talkers do vary in how much physical information they produce during speech, human perceptual strategies also play a role in talker variability.
© 2006 Elsevier Ltd. All rights reserved.

1. Introduction

People in a crowd often understand conversation better when they can see their conversation partners' faces. In fact, Sumby and Pollack (1954) reported over 50 years ago that being able to see a talker produce speech visually while listening to the talker's auditory speech presented in white noise significantly improved speech perception for normal-hearing observers. This improvement has been calculated to be as much as a 15 dB gain in signal-to-noise ratio for visual plus auditory speech compared with auditory speech alone (Summerfield, 1987). Furthermore, although the ability to understand speech through a visual signal alone ("lipread" or "speechread") varies considerably from person to person, normal-hearing observers are typically able to understand speech to some degree when it is presented in the visual-only modality, without an accompanying auditory signal (Bernstein, Demorest, & Tucker, 2000).

One issue that has surfaced in many studies of visual speech is the variability in the accuracy of visual-only speech perception (often referred to as "visual-only speech intelligibility") across different talkers. Results from the literature indicate that some talkers are consistently easier than others to speechread. While not every pair of talkers differs significantly in visual intelligibility, talker variability has been reported to result in visual intelligibility differences of anywhere from 4% to 23% between talkers (Bernstein, Jiang, & Alwan, 2001; Gagné, Querengesser, Folkeard, Munhall, & Masterson, 1995; Kricos & Lesner, 1982, 1985; Lachs & Hernandez, 1998; Montgomery & Jackson, 1983). Similarly, auditory intelligibility has been reported to vary by 5–20% between talkers, depending on listening conditions (Cox, Alexander, & Gilmore, 1987; Lachs & Hernandez, 1998).

Understanding the sources of variability in visual-only talker intelligibility is potentially important because, like auditory talker variability, visual talker variability may have cognitive consequences for speech perception. For example, changing talkers from trial to trial during a speech perception task has been found to affect cognitive processing whether the speech is auditory or visual. Auditory talker variability decreases speech intelligibility and increases word-naming latencies when the words in a list are spoken by different talkers; also, recognition memory is better for an old word presented in the original presentation voice than for an old word presented in a novel voice, suggesting that talker-specific information is encoded during auditory speech perception (Pisoni, 1997). Although fewer studies have examined the cognitive effects of talker variability in visual speech, Yakel, Rosenblum, and Fortier

[*] Corresponding author. Fax: +1 812 855 4691.
E-mail address: jgold@indiana.edu (J.M. Gold).

0042-6989/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2006.03.020

(2000) reported that changing talkers from trial to trial produced a decrease in speechreading performance compared with using the same talker for all trials. This result suggests that talker variability in visual speech also affects the cognitive processing of speech.

A potential pitfall of measuring variability in visual talker intelligibility is that speechreading ability varies a good deal across observers. If these individual differences in speechreading ability affect the magnitude and particularly the direction of visual talker intelligibility differences, then the variability in speechreading ability would have a larger effect than talker variability and perhaps render talker variability irrelevant. The evidence indicates that although individual differences in speechreading scores can be large, the direction of visual intelligibility differences for different talkers tends to be consistent across observers. Demorest and Bernstein (1992) conducted a systematic study of the effects of talker variability on the speechreading performance of 104 normal-hearing adults. They found a main effect of talker on speech intelligibility scores, but no Talker × Observer interaction in their results. These findings indicate that visual intelligibility differences between talkers existed but that the direction of the differences was consistent across observers.

Although the evidence suggests that visual intelligibility differences among talkers are consistent across observers, the specific factors underlying these differences are unknown. Two general types of factors may underlie talker visual intelligibility differences—those due to the talkers' production patterns and those due to the observer's perceptual strategies. To illustrate these factors, take a situation in which it is known that Talker A is less intelligible than Talker B. It is also known that Talker A moves only her lips during speech, whereas Talker B moves his lips in a similar way but also provides additional information about what he is saying through his jaw movements. One possible cause of the intelligibility difference is that the increased physical availability of information makes Talker B easier to understand than Talker A. Alternatively, it might be that the observer does not pay attention to either talker's lips, but instead watches only the talkers' jaw movements. In this case, because Talker B produces informative jaw movements but Talker A does not, Talker B would again be more intelligible than Talker A, but this time the intelligibility difference would be due to the observer's perceptual strategy rather than to differences in the amount of physical information produced by each talker. Finally, a combination of both factors is possible: the talkers might produce different amounts of information, and also the observer's perceptual strategies might be better suited to some talkers than others.

The extant models that posit the variability is due to physical information differences across talkers either describe physical characteristics of highly intelligible talkers or attempt to relate variability in visual intelligibility with observable physical differences in speech production. Lesner (1988) listed several factors that characterize a "good" talker (i.e., one who is highly visually intelligible) or tend to make any talker more visually intelligible. According to Lesner, visual intelligibility increases as more facial information is made available; seeing the lips alone is not as good as seeing the whole face. Normal lip movements produce more intelligible speech than exaggerated ones, and more intelligible talkers are claimed to have thinner lips. Facial hair produces lower visual intelligibility, and an "expressive" face increases visual intelligibility. Further characteristics of highly intelligible visual speech include facial and other message-related gestures, simpler linguistic messages, and familiarity with the talker. Some studies also report that the female talkers they used are more visually intelligible than the male talkers (e.g., Lachs, 1999), although this is not always the case and may be a question of individual talker differences rather than more general gender differences.

Although Lesner (1988) provided a description of what seems to make a good talker, this description does not predict specific relationships between physical characteristics of the talker and perceptual performance of speechreaders. Other studies have attempted to quantify these relationships more precisely. For instance, Montgomery and Jackson (1983) measured the height, width, and area of lip opening during vowel production and the acoustic and visual durations of vowels to test whether these factors would predict the vowel identification performance of observers. Measurements of lip opening and visual vowel duration were moderately good predictors of perceptual confusions and multidimensional scaling (MDS) space distances among vowels produced by some talkers. However, the best predictors and the strength of the predictions varied by talker and were not significant for all talkers. The measures were good predictors for the least intelligible talker and two others, but not for a talker who was close in intelligibility to the latter two.

In a more recent study of visual intelligibility, Bernstein et al. (2001) measured the physical distances between consonant articulations using markers placed on the lips, cheeks, and chin. The perceptual distances between consonant phonemes were also calculated with a MDS analysis. Bernstein et al., reported that more visually intelligible talkers displayed higher correlations between physical distances among consonant articulations and perceptual distances between phonemes on the MDS analysis. However, the perceivers were presented with visual stimuli that included the markers used in the physical analysis, so the perceivers may have been biased to use these locations in generating their responses.

As described above, previous research generally indicates that the physical characteristics of the talkers and perhaps the spatiotemporal cues they produce during speech may be important for visual-only speech identification. In addition to these investigations of talkers' production patterns, observations about listeners' perceptual strategies have been made in a study of perceivers' gaze behavior during visual-only speech. This evidence from the eye-movement literature has suggested that different eye-gazing behaviors may occur during the perception of a more versus less visually intelligible talker. Lansing and McConkie (2003) reported when visual-only speech was presented, observers tended to spend more time gazing at the mouth than when visual speech was presented with auditory speech in low-level noise. Similarly, within the visual-only speech conditions, observers tended to gaze more at the mouth of the talker who had lower intelligibility scores than at the mouth of the other talker. Lansing and McConkie suggested that the observers directed more attention toward the mouth when they had more difficulty gleaning the necessary information from visual speech. These changes in processing strategies may have been prompted by a lack of necessary physical information or by a discrepancy in the locations of physical information and of typical gazes.

Despite considerable evidence that talkers differ in visual intelligibility, many studies of auditory-visual or visual-only speech fail to account for this in their experimental designs or the interpretation of their results. Studies that use two or more talkers sometimes fail to take talker differences into account in the interpretation of their results. For instance, Vatikiotis-Bateson, Eigsti, Yano, and Munhall (1998) studied the eye movements of English-speaking subjects who watched and listened to a talker speaking English and Japanese-speaking subjects who watched and listened to a talker speaking Japanese. The Japanese speakers fixated on one of the talker's eyes less frequently than the English speakers during speech presented at high auditory signal-to-noise levels. The authors concluded that this difference was probably a language-specific effect but did not examine the possibility that differences in the visual appearance or intelligibility of the talkers might have influenced the results. Another study, by Yakel et al. (2000), looked at the effects of changing talkers from trial to trial on speechreading performance and found that performance fell by an average of 7.9% for mixed-talker sentence lists compared to lists in which only a single talker appeared. They attempted to control for non-specific stimulus differences between talkers by using different-colored filters on each trial for both single and multiple-talker lists. However, they acknowledged that they had not controlled for differing effects of the colors on perceptual salience of the talkers' faces, nor did they examine differences due to visual intelligibility of the talkers.

Part of the difficulty with such studies of variability in visual speech is that perceptual measures of visual intelligibility do not account for how much information is physically available in the stimulus. This means that variability in the physical information available in different talkers' speech has been confounded with how human observers are using that information. Separating the two factors of physical variability in the available information and perceptual strategies used to process that information can be accomplished through a technique called *ideal observer analysis*. Ideal observer analysis is a technique that can be used to quantify the amount of physical information available in a perceptual task. The ideal observer is defined in such a way as to produce the best possible performance on a particular task, and so is limited only by the amount of physical information available in the stimuli (Geisler, 2004). When the signals in a task are specified exactly and presented in uncorrelated Gaussian white noise, the ideal observer acts as a template matcher, choosing the response that maximizes the cross-correlation between the presented stimulus and the stored templates of the possible signals (Green & Swets, 1966; Tjan, Braje, Legge, & Kersten, 1995). Ideal observers cannot be formulated for certain types of tasks, such as those with subjective or open-ended responses (Geisler, 2004), but an ideal observer analysis is tractable and computationally feasible in many psychophysical tasks such as those involving the measurement of the detectability index or of contrast energy thresholds.

The ideal observer provides an objective, assumption-free measure of the level of performance possible in various conditions of a task, resulting in a yardstick for human performance and a method for separating the physical availability of information from perceptual strategies used to process that information. The performance of a human or other sub-ideal model observer relative to the ideal is described by a measure called *efficiency*, which is the ratio of the ideal to the sub-ideal contrast energy required to perform the given task at threshold. Because the ideal observer uses all the physical information available, efficiency provides an index (of up to 100%) of how much information a given observer is using when performing a task. An observer with an efficiency of 100% has the same contrast energy threshold as the ideal observer and thus is optimally efficient. An observer with an efficiency of less than 100% has a higher contrast energy threshold than the ideal observer and by definition is not using all the physical information available to perform the task. It is assumed that human performance will generally not be optimally efficient in a given task (Geisler, 1989). However, examining efficiency across conditions of a task provides an index of whether human observer performance is consistent with the physical availability of information. For example, if efficiency is relatively constant across the conditions of a task, then human contrast energy thresholds display a similar pattern to ideal observer contrast energy thresholds and are consistent with the physical availability of information. However, if efficiency varies across conditions, then factors other than the physical availability of information, specifically perceptual strategies, must be affecting human thresholds.

In this study, we use ideal observer analysis to test the hypothesis that differences in visual talker intelligibility are due to physical factors inherent to the talkers' speech, as suggested by the descriptions of Lesner (1988) and the results of Montgomery and Jackson (1983) and Bernstein et al. (2001). The alternative hypothesis is that differences in talker visual intelligibility are due to the observer's perceptual strategies, as suggested by Lansing and McConkie's (2003) eye movement results. Because it is not possible to formulate an ideal observer for tasks with open response sets, such as the open-ended identification of spoken words, we chose to measure contrast energy thresholds for the identification of a set number of videotaped spoken words presented without sound (“word identification thresholds”). These words were presented in uncorrelated Gaussian white pixel noise, as obtaining human thresholds in this type of noise greatly facilitates the ideal observer analysis. Because previous studies have found variability across talkers in the percentage of words identified correctly, we predicted that human word identification thresholds measured for a given percentage correct would also vary across talkers. If the ideal observer’s thresholds vary in a similar way, the results of the present study would be consistent with the hypothesis that the variability in human speech recognition across talkers is due to variability in the physical stimulus alone. However, if human thresholds vary but ideal thresholds either remain constant or vary in a different fashion across talkers, the results would indicate that human observers are more efficient at using the available information for some talkers than for others. This would be consistent with the hypothesis that the perceptual strategies used by the human observers are more optimally suited for understanding some talkers than others.

2. Method

2.1. Observers

Eight observers participated in this study (4 females, 4 males). Seven of the 8 observers were naïve to the purpose of the experiment; the other observer was an author (BC). All had normal or corrected-to-normal vision. Their ages ranged from 23 to 40, with a mean age of 28.5. Each observer took four approximately 1-h sessions to complete the experiment. With the exception of BC, they were compensated at a rate of $10 per hour for their time.

2.2. Apparatus

Stimuli were displayed on Sony Trinitron Multiscan G520 monitors. Two monitors in separate testing rooms were used during testing of human observers. Each monitor had a resolution of 1024×768 pixels, subtending $16.4 \times 12.4°$ of visual angle at the viewing distance of 130 cm. The frame rate was set to 85 Hz. The monitors were each controlled by an Apple G4 computer running Mac OS 9.2.2. The experiment was conducted in the MATLAB programming environment (version 5.2.1 for MacIntosh) using the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997).

Luminance calibrations were performed for each monitor with a Minolta Luminance Meter LS-100 photometer, and a 2025-element look-up table was built from the calibration data (Tyler, Chan, Liu, McBride, & Kontsevich, 1992). Luminance on one of the monitors ranged between 0.8 and 102.0 cd/m^2, with an average luminance of 33.2 cd/m^2; on the other monitor, luminance ranged between 0.9 and 145.1 cd/m^2, with an average luminance of 49.6 cd/m^2. On each trial of the experiment, appropriate luminance values from the calibrated look-up tables were selected by the computer software and stored in the 8-bit look-up tables for the monitor.

2.3. Stimuli

The stimuli comprised eight highly familiar, monosyllabic English words matched for frequency (Kucera and Francis, 1967). All words contained three different speech sounds (phonemes) and were of the form *consonant-vowel-consonant* (CVC). To the extent possible, the words were selected so as to maximize the number of different phonemes present in the stimulus set. The words were *far*, *gave*, *job*, *house*, *put*, *should*, *thought*, and *voice*.

Eight Indiana University undergraduates (4 females, 4 males) served as talkers for this experiment and were videotaped saying the stimulus words using a Canon ZR60 digital video camera. The talkers were requested to shave if applicable on the morning of the recording. When they arrived for the recording session, they were asked to pull their hair back from their face, remove their glasses if necessary, and put on a white T-shirt over their clothing. During a recording session, the talker sat in a chair and was asked to place his or her arms on the armrests and keep as still as possible while looking directly at the camera. The camera angle was adjusted so that the talker was recorded from the top of the shoulders to the top of the head. The list of stimulus words as read by one of the authors (BC) was played from a CD recording over computer speakers at a comfortable listening volume that was kept constant for all talkers. The words were spaced approximately 5 s apart on the recording, and the talker was instructed to repeat each word after it was played. Each talker was given at least three opportunities to say all the stimulus words so that there would be several tokens of each word to choose from in constructing the experimental stimuli. All talkers were played the same recording of the stimulus words so as to minimize phonetic and prosodic variability in the words they were to pronounce and also to eliminate excessive head and eye movements that could have arisen if talkers had been asked to read the words from a printed list.

The videos were recorded at a sampling rate of 30 frames per second. Simultaneously with each video recording, a high-quality audio recording of the talker’s speech was also recorded at a sampling rate of 44,100 Hz using a TASCAM DA-P1 digital audio tape (DAT) recorder.

Once the video and audio recordings had been made, they were uploaded into iMovie (version 3.0.3) and saved. The saved recordings were then imported into Final Cut Pro HD (version 4.5). The video and auditory recordings for each talker were aligned manually by using a sharp clicking sound produced at the beginning of each recording and also through agreement between the audio track from the video and the higher-quality DAT recording. The aligned recordings were then segmented into individual Quicktime movie files, one for each word token. Each talker’s best token for each word was chosen for use in the experiment as follows: Two raters (BC and a research assistant) independently ranked the three tokens of each word for how good they would be as experimental stimuli. “Good” tokens were those without eyeblinks and with minimal head motion. The raters rarely disagreed on which token was the best for use in the experiment, but in cases of disagreement BC’s best-rated token was used.

After the word tokens had been chosen, the duration of each word clip in frames was measured from the last still frame before visual articulation of the word began to the first still frame after visual articulation of the word had been completed and no sound was audible. The longest word clip for any talker was 30 frames (1 s) in duration. So that talker differences would not be confounded with word clip duration, all other word clips from all talkers were extended to be 30 frames in duration by including more recorded frames as needed on either side of the word’s active articulations.

For purposes of the present experiment, only the visual portion of each word clip was used; no auditory information was presented. Each visual-only 30-frame word clip was exported from Final Cut as a series of TIFF images. These images were scaled so that the talkers’ faces were all the same height in pixels. So as to avoid problems of upsampling, the scaling factor for each talker was determined by comparing his or her face height with the shortest face height. The width of the talkers’ faces was scaled using the same scaling factor as the height.

The TIFF images were converted to grayscale and were recombined into movies represented as three-dimensional matrices in MATLAB. The values in the matrices were converted to contrast values such that the contrast (c_{xy}) at pixel location (x,y) was given by

$$c_{xy} = \frac{l_{xy} - L}{L}$$

with L equal to the average luminance and l_{xy} equal to the pixel luminance. For stimulus presentation, each talker's face was contained within an elliptical region. The size of the elliptical region was constant for all talkers, at 97 pixels wide and 150 pixels high. The ellipse was embedded in a rectangle that was also 97 × 150 pixels; portions of the rectangle that fell outside the ellipse were set to a constant background of zero contrast (i.e., average luminance) for all talkers. The elliptical region was placed according to the average position (over all 30 frames) of the inner corner of the left eye on each word for each talker, and it contained the facial region from the forehead to the bottom of the chin and between the outer corners of the eyes. Example still frames from the stimulus movies are shown in Fig. 1.

For the final movies used in stimulus presentation, the first and last frames of each word were duplicated six times each so that the movie began and ended with seven identical still frames. Pilot data had suggested that these still frames gave the observer a better chance of focusing his or her attention on the stimulus before articulation began. The movies presented to the human observers thus contained 42 frames, each one presented for three screen refreshes at a refresh rate of 85 Hz, for a total movie duration of ~1.5 s.

Prior to each trial, a word movie was read into memory, the contrast energy (i.e., integrated squared pixel contrast) of the movie was set to the desired value by multiplying the movie matrix by an appropriate constant, and contrast values were converted to luminance values. A linear 8-bit look-up table for the display was constructed using these luminance values. Finally, the movie luminance values were mapped onto the look-up table values.

2.4. Noise fields

A unique sample of dynamic Gaussian white contrast noise was generated on each trial and added to the trial's word movie. The size of the Gaussian noise field was identical to the size of the word movie (97 pixels wide × 150 pixels high × 42 frames). The noise field values were taken from a Gaussian pseudo-random number generator with a mean of 0 and a variance of 0.01. As with the word movies, each pixel's value in the noise matrix was treated as a contrast value. The variance chosen for the noise ensured that at least 95% of the values in the distribution would fall within the linear contrast range of the noise display. Values that exceeded ±2 standard deviations from the mean were resampled to fall within the range of displayable contrast values. The spectral density of the noise (energy per unit bandwidth) at the viewing distance of 130 cm was $2.675 \times 10^{-6}\ \mathrm{deg}^2$ in all experimental conditions.

2.5. Viewing conditions

Each 97 × 150 pixel movie frame subtended a visual angle of 1.6 × 2.5° at the 130 cm viewing distance. Viewing was binocular through natural pupils, and a combination forehead and chinrest stabilized the observer's head. The monitor supplied the only source of illumination during the experiment.

2.6. Procedure

A one-of-eight word identification task was used to estimate word identification thresholds for each talker. The contrast energy of the word movies was manipulated across trials using an adaptive one-up, one-down staircase procedure to obtain each observer's 50% correct word-identification threshold for each talker (chance performance = 12.5%) on a fixed number of trials. In the first session, each observer completed blocks of 50 practice trials for each of the eight talkers. In the subsequent three test sessions, the observers completed blocks of 150 trials for each of the eight talkers; only these 150 trials were used to estimate word identification thresholds. Trials were blocked by talker in all sessions, and talker order was randomized and differed for the practice and test trials for each observer.

Throughout each block of trials, a white fixation rectangle at the center of the screen framed the location where the stimulus appeared. On each trial, one of the eight word movies was played, without sound, within the fixation rectangle. The word movies were presented in random order but each one appeared approximately the same number of times during a block of trials. After the movie was played, the monitor was set to average luminance, and the observers were presented with a selection screen containing written representations of the eight words. Four of the words (*far*, *gave*, *house*, and *job*) appeared at the top of the screen above the fixation rectangle, and the other four words (*put*, *should*, *thought*, and *voice*) appeared at the bottom of the screen below the fixation rectangle. Observers chose with the mouse the word they thought had been presented. After a choice was made, auditory feedback indicated whether the response was correct, and the display was set to average luminance prior to the beginning of the next trial. During the test sessions, observers received breaks every 50 trials.

The psychometric data were fit by Weibull functions, and the word identification threshold was identified as the contrast energy yielding 50% correct responses.

3. Results

3.1. Human observers

Contrast energy thresholds were obtained for all of the human observers on all eight talkers. Because human word identification thresholds could be measured in uncorrelated Gaussian white noise, the formulation of the ideal observer was tractable and greatly simplified, as described in the next section. Individual human observer thresholds are shown in Fig. 2. In this figure and all subsequent ones, F1 through F4 correspond to the four female talkers and M1 through M4 correspond to the four male talkers.

Consistent with findings in the visual-only speech literature (Bernstein et al., 2001; Gagné et al., 1995; Kricos & Lesner, 1982, 1985; Lachs & Hernandez, 1998; Montgomery & Jackson, 1983), there was considerable variability in

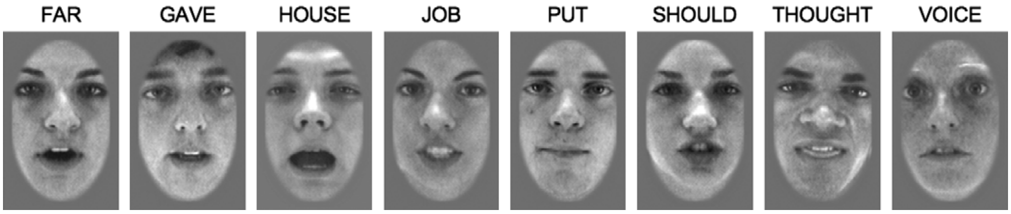

Fig. 1. Example static frames from talker movies. The talkers are, from left to right, F2, M2, M1, F4, M4, F3, M3, and F1. F, female talker, M, male talker.

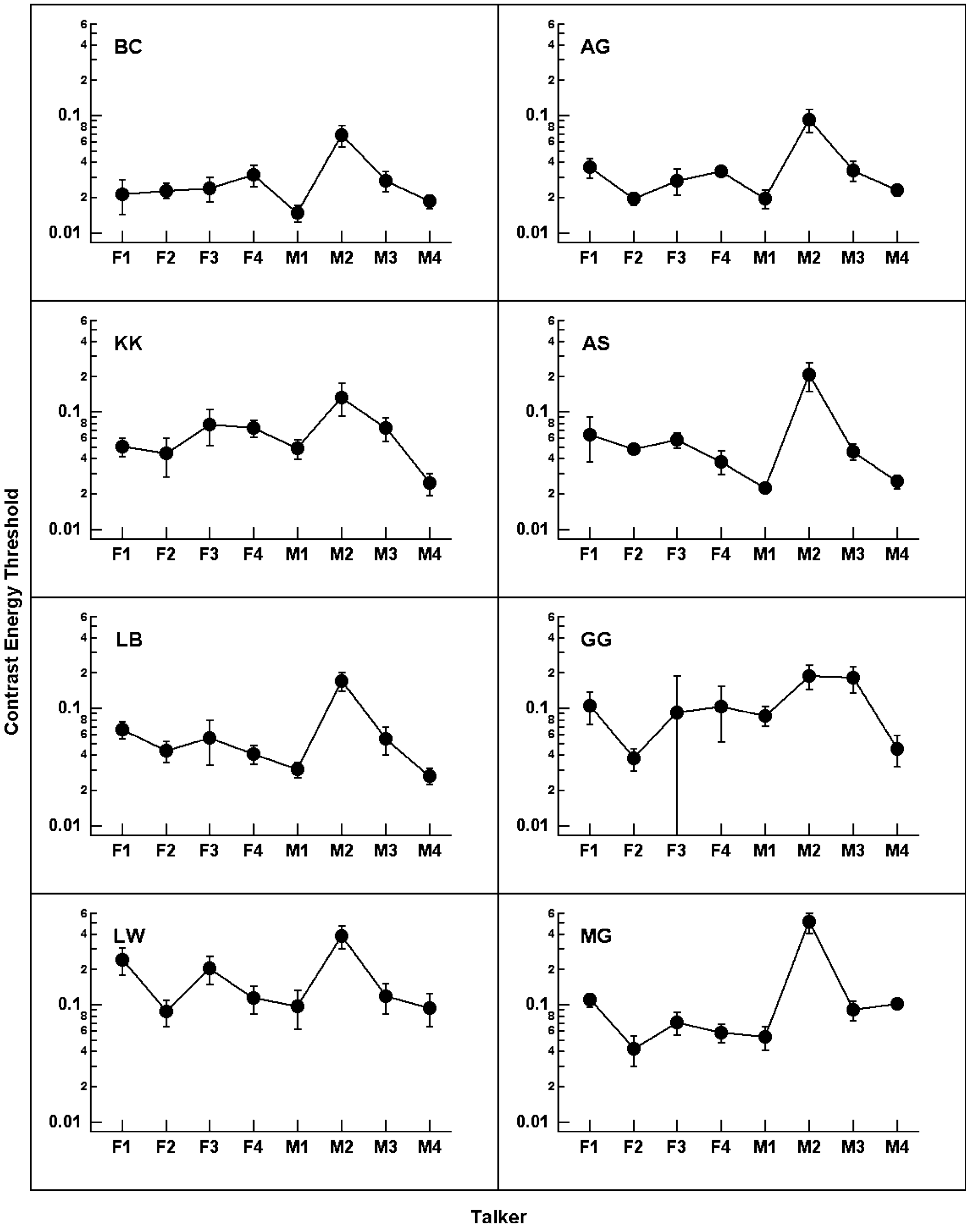

Fig. 2. Contrast energy thresholds for the eight human observers in the word-identification task. F, female talker, M, male talker. The panels on the left show thresholds for the female observers and the panels on the right show thresholds for the male observers. Error bars correspond to ±1 SD.

word identification thresholds across talkers for the human observers. The range of contrast energy thresholds across talkers spanned anywhere from 0.65 to 1.1 log units within observers, with an average range of 0.88 log units. Based on the patterns of contrast energy thresholds, if contrast energy were held constant, visual-only intelligibility, or the percent of words identified correctly, would be expected to vary considerably across talkers. Also consistent with the literature (Demorest & Bernstein, 1992), the pattern of cross-talker variability was similar for the different observers (although human observers varied in their overall levels of performance, here indexed by contrast energy thresholds). For example, talker M2 produced the highest word-identification threshold for all the human observers, and talker M1 had one of the lowest thresholds for all the human observers.

3.2. Ideal observer

The ideal observer for this task maximizes the cross-correlation between the word movie (signal + noise) stimulus and each of the eight possible noise-free word movie stimulus matrices, or templates (Green & Swets, 1966; Tjan et al., 1995). Ideal observer thresholds were obtained for each talker using Monte Carlo simulations in which the signal + noise stimulus was compared with all of the noise-free templates. On each trial, the template with the highest cross-correlation with the stimulus was chosen. The contrast energy of the signal was varied to obtain an estimate of the 50% correct word-identification threshold for each talker. Each contrast energy threshold was estimated by fitting a Weibull function to the data from 1000 simulated trials. The ideal observer's thresholds are shown in the closed symbols in Fig. 3. The corresponding human efficiencies are shown in the closed symbols of Fig. 4.

Like the human thresholds, the thresholds obtained for the ideal observer also varied somewhat across talkers, indicating that the amount of physical information available to perform the word identification task varied. Also like the human observers, the ideal observer's threshold for M2 was the highest; however, the pattern of variability was generally not very similar to that of the human observers. For instance, talker M1 had one of the higher thresholds for the ideal observer but one of the lower thresholds for most human observers. Also, the range of the thresholds across talkers was much smaller for the ideal observer (0.47 log units) than for any of the human observers.

The absolute levels of efficiency were low, with average efficiency ranging from 0.03 to 0.3% for individual observers (mean across observers = 0.09%). Some factors that may have contributed to the low efficiencies are discussed in the "Summary and Conclusions" section below. However, the variability in efficiencies across talkers, rather than the absolute magnitude of efficiencies, is the primary focus of this study.

If the differences in the physical availability of information that caused the variability across talkers in the ideal observer thresholds had also accounted for all of the variability across talkers in the human observer thresholds, the efficiencies, which are ratios of human to ideal observer thresholds, would have been relatively constant across talkers. Because efficiencies varied across talkers, differences in the physical availability of information do not fully account for the variability in human observer thresholds across talkers. This implies that human perceptual strategies for using that information must also play a role in the variability of human observer thresholds across talkers. In the following section we describe several possible models of human perceptual strategies for performing the word identification task that we explored through a series of computer simulations.

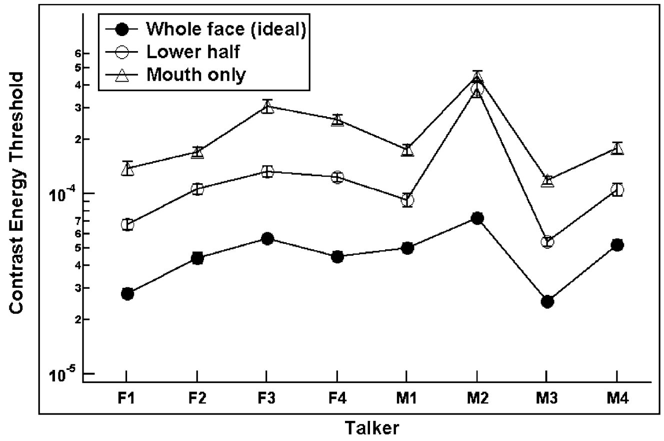

Fig. 3. Contrast energy thresholds for the ideal observer (filled circles), which used information from the whole face of each talker, and for the model observers whose information use was spatially restricted to the lower half of the face (open circles) or the mouth only (open triangles). Error bars correspond to ±1 SD.

A critical assumption of ideal observer analysis is that once information is lost at a particular stage, it cannot be recovered and so must play a role in the ultimate measurement of the threshold (Geisler, 1989). This is not an airtight assumption—some alternate configuration of circumstances can always be imagined that would also result in the same threshold function—but it is the most parsimonious method of accounting for the data. So, the model comparisons described below focus on comparing the variability in efficiencies across talkers rather than the magnitude of the efficiencies, in a manner analogous to Geisler's (1989) concept of *relative efficiency*.

4. Simulations

4.1. Spatially restricted models

Previous work on speech perception suggests a couple of perceptual strategies that humans might adopt when perceiving visual-only speech. Some studies have used only the lower half of the face as a visual stimulus (Bernstein et al., 2001; Campbell, 2000), which assumes implicitly that humans only need information from the lower half of the face for effective visual-only speech perception. In addition, it has been reported that humans tend to concentrate their eye movements around the mouth during visual-only speech perception (Lansing & McConkie, 2003). To explore these ideas, we developed two constrained, sub-ideal observer models based on the ideal observer's decision rule. Specifically, we conducted computer simulations to test whether restricting the use of spatial information to either to the lower half of the face or to the mouth only would provide a better description than the ideal observer model of the variability we observed in human thresholds across talkers.

4.1.1. Lower half of the face

One possible human perceptual strategy in the visual-only word identification task is to look only at the lower half of the face. The lower half of the face, from the bottom of the nose down, includes the mouth, jaw, and chin, which all move during speech production. For purposes of this simulation, the lower half of the face was defined as all pixels in the word movies falling below the nose. Because

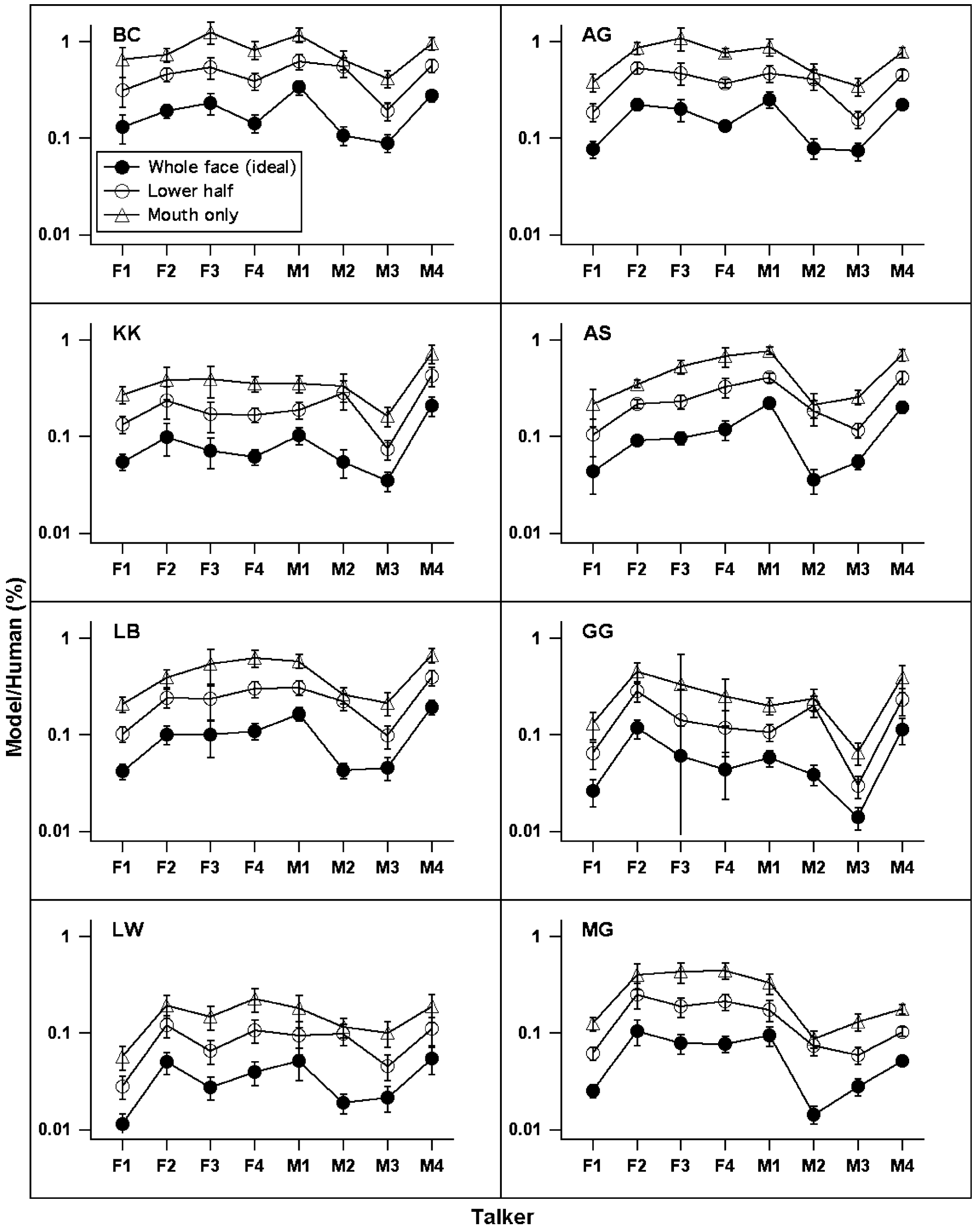

Fig. 4. Model-to-human threshold ratios for each of the human observers. Shown in the figure are ratios for the human observers vs. the ideal, whole-face observer (efficiencies; filled circles), the human observers vs. the lower-half model observer (open circles), and the human observers vs. the mouth-only model observer (open triangles). Error bars correspond to ±1 SD.

the talkers' faces had been scaled to the same height, the bottom of the nose was in nearly the same position for all talkers. Accordingly, the same dimensions were used for all talkers in defining the lower half of the face. The resulting stimuli were the same width and approximately half the height of the original movies, at 97 pixels wide × 76 pixels high × 42 frames. An example of a still frame from the lower half of the face only (taken from Talker F2) is shown in the left panel of Fig. 5.

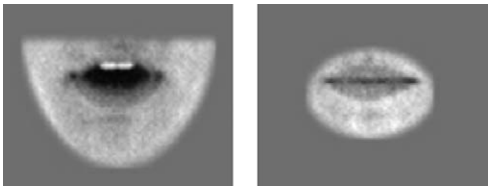

Fig. 5. Example static frames from the lower-half (left, Talker F2) and mouth-only (right, Talker F4) talker movies.

To simulate the effect of looking only at the lower half of the face, information use for a model observer was spatially restricted to the lower half of the face only. The data used to estimate contrast energy thresholds of the model observer for each talker were obtained using Monte Carlo simulations (1000 trials per talker), as described for the ideal observer. Specifically, on each trial, the model observer's noise-free templates of the lower half of the face only were compared with signal-plus-noise stimulus information from the lower half of the face only. To set the contrast energy of the movie on each trial, the entire movie matrix for the whole face was multiplied by an appropriate constant, and then the information from the lower half of the face only was presented as the stimulus in noise. The contrast energy was set for the entire movie rather than for the lower half of the face only so that the distribution of contrast energy across the stimulus could be compared with that for the whole face, ideal observer condition.

4.1.2. Mouth only

Another possible human perceptual strategy in the visual-only word identification task is to look only at the talker's mouth, as suggested by Lansing and McConkie (2003). To simulate this strategy, a model observer was implemented with information use spatially restricted to the mouth only. The mouth-only area of the word movies was defined as the smallest region that could accommodate the entire mouth from the top of the upper lip to the bottom of the lower lip and from corner to corner for all the talkers. The resulting stimulus measured 69 pixels wide × 54 pixels high × 42 frames. An example of a still frame from the mouth-only stimuli (taken from Talker F4) is shown in the right panel of Fig. 5. Monte Carlo simulations were conducted as described for the lower half of the face stimuli to obtain contrast energy threshold estimates for each talker.

4.1.3. Evaluation of ideal and spatially restricted models

Contrast energy thresholds for the spatially restricted model observers are shown in the open symbols in Fig. 3; ratios of the model-to-human performance, similar to efficiency for the ideal observer, are shown in open symbols in Fig. 4.[1] The open circles correspond to the lower half of the face observer model and the open triangles correspond to the mouth-only observer model.

There are several aspects of these data worth noting. First, the lower half and mouth-only models had approximately the same threshold for M2, the most difficult talker for the human and ideal observers, suggesting that little information was available in the lower half of his face outside the region of the mouth. Second, the contrast energy thresholds for the lower half of the face model covered a range of 1.1 log units, which is on the upper end of the variability in the human thresholds (human mean = 0.88 log units; range = 0.65 to 1.1 log units). The thresholds for the mouth-only model covered a range of 0.74 log units, which falls in the middle of the range of human thresholds. As expected, the thresholds for model observers that were spatially restricted in information use were higher than for the ideal observer, which had access to information from the whole face. Thresholds for the lower half model were an average of 0.4 log units higher than for the whole face model, and thresholds for the mouth-only model averaged 0.7 log units higher. Because the spatially restricted observer thresholds were higher than ideal observer thresholds, model-to-human threshold ratios were also higher. Average model-to-human threshold ratios for individual observers ranged from 0.08% to 0.5% for the lower half model (mean = 0.2%) and from 0.1% to 0.8% for the mouth-only model (mean = 0.4%).

Finally, the model-to-human threshold ratios for the mouth-only observer appeared to be more constant across talkers than for the ideal observer or the lower half observer, indicating that the mouth-only model provided the best fit to the variability in the human data. To quantitatively compare the goodness-of-fit of the ideal and spatially restricted models to the human data, the models were each scaled for the best fit to the human data, using the logic for relative efficiency described above (Geisler, 1989). Specifically, the base-10 logarithms of the contrast energy thresholds from each of the three models were scaled by a factor that minimized the summed, squared deviation of the model's pattern of thresholds from a given human observer's log-transformed contrast energy thresholds. The thresholds were log transformed to allow for properly scaled comparisons of the model thresholds with the human thresholds, which were much higher. The average deviation of the model's contrast energy thresholds from the human thresholds was computed using a statistic called the root mean squared deviation (RMSD; cf. Massaro, Cohen, Campbell, & Rodriguez, 2001; Massaro & Friedman, 1990). RMSD is calculated as follows:

$$\sqrt{\frac{\sum(x_{\text{pred}} - x_{\text{obs}})^2}{n}},$$

where x_{pred} is the predicted value (here, model thresholds), x_{obs} is the observed data (here, human thresholds for individual observers), and n is the number of conditions (here, talkers). Although the RMSD does not control for the "complexity" or number of free parameters in the models compared, here it is an appropriate benchmark of model performance because all of the models are presumably of equal complexity (Pitt, Kim, & Myung, 2003). The models do not differ formally and were each fitted to the human threshold data using only one free parameter (the scaling parameter); only their templates and the stimuli they were presented with differed. The RMSD for

[1] The model-to-human threshold ratios for the spatially restricted models are not efficiencies in the sense of measuring the percentage of available information used, because unlike the human or ideal observers, the spatially restricted model observers did not have access to information from the whole face.

each model fit was calculated separately for each human observer. It was considered appropriate to fit the models to the data from individual observers rather than to the average of the individual data because the thresholds of different observers also varied greatly in absolute magnitude (see Fig. 2).

The model with the lowest RMSD, or smallest error, provides the best fit to the human data. The ideal observer, or whole face, model had a mean RMSD of 0.24 log units ($SD = 0.04$ log units) across human observers. The lower half model had a mean RMSD of 0.27 log units ($SD = 0.04$ log units), and the mouth-only model had a mean RMSD of 0.22 log units ($SD = 0.04$ log units). Paired t tests showed that the RMSDs for the whole face and lower half models were not significantly different ($t(7) = -1.593$, $p > .05$), but that the RMSDs for the mouth-only model were significantly lower than for either the whole face or the lower half model ($t(7) = 3.537$, $t(7) = 4.075$, respectively; $p\text{s} < .05$). These results indicate that of the ideal and spatially restricted models, the mouth-only model performed best overall in fitting the human data. When the data from individual observers were considered, the mouth-only model also performed best overall. For the individual observers, the RMSDs for the mouth only were lowest in six of eight cases and second lowest in the other two cases. The RMSDs for the lower half of the face were lowest for one observer, second lowest for one observer, and highest for the remaining six observers. The whole-face RMSDs were lowest for one observer, second lowest for five observers, and highest for two observers.

4.1.4. Summary: Spatially restricted models

The results from simulations conducted with an ideal observer and two model observers spatially restricted in their use of information suggest that the physical information differences across talkers combined with the perceptual strategy of only attending to the talker's mouth produce the best agreement with human data of the models tested. The ideal or whole face model's contrast energy thresholds showed variability across talkers, but the pattern of variability displayed by the mouth-only model was more consistent with human performance. The lower half of the face model did not improve significantly over the whole face model and in fact displayed higher overall RMSDs.

This study primarily sought to find whether cross-talker variability could be explained by physical information differences or if it could be due to human perceptual strategies. The answer so far seems to be that cross-talker variability results from both physical differences and perceptual strategies including perhaps the limiting of attention to the region around the mouth of the talker. However, although the mouth-only model does provide a better fit of the human data than the ideal observer model, it does not account for all of the variability in human thresholds across talkers.

4.2. Models of other perceptual strategies and inefficiencies

Spatially restricting information use to the lower half of the face and the mouth only are the primary perceptual strategies suggested by the visual-only speech perception literature (Bernstein et al., 2001; Campbell, 2000; Lansing & McConkie, 2003). To determine whether other simple perceptual strategies and inefficiencies could also be contributing to the remaining variability across talkers, several other simulations were conducted using model observers with other built-in constraints including spatial uncertainty, temporal uncertainty, and selective attention to the highest contrast pixels in the word movie. For all of these simulations, the mouth-only version of the stimuli was used, because the mouth-only model provided a significantly better fit to the human data than the other models tested and is also supported by eye-movement data.

4.2.1. Spatial and temporal uncertainty

One possible factor that could have limited human performance on the visual-only word-identification task is spatial uncertainty, or uncertainty about the exact spatial location of the stimulus on the computer screen. Spatial uncertainty has also been described as uncertainty about where spatially to apply a template that is to be matched to the stimulus (Eckstein & Whiting, 1996). In addition to spatial uncertainty, human observers may have experienced temporal uncertainty, or uncertainty about when the word occurred during stimulus presentation. Simulations were conducted to explore the effects of small amounts of spatial or temporal uncertainty on model observer thresholds.

In the spatial uncertainty simulation, on each trial, the stimulus was compared with templates that had been shifted to all possible positions within a certain number of pixels of the original templates, which in this case were the mouth-only versions of the word movies. The word corresponding to the template with the highest cross-correlation with the stimulus was chosen as the response. Spatial uncertainties of up to 1, 3, and 5 pixels were simulated in this way. In order to allow for uncertainty in the position of the stimulus, 10 additional zero-contrast pixels were added to each side of the mouth-only word movies, thus increasing the stimulus size from 69 pixels wide × 54 pixels high × 42 frames to 89 pixels wide × 74 pixels high × 42 -frames. (Because the last dimension in the stimulus size is temporal and represents the number of frames in the word movie, this did not need to be altered for the spatial uncertainty simulations.)

Monte Carlo simulations of 200 trials per talker[2] were conducted to estimate word-identification thresholds,

[2] Only 200 trials per talker were used in each condition for the spatial and temporal uncertainty simulations, rather than the 1000 trials per talker used in the lower half and mouth-only simulations, because the spatial and temporal uncertainty simulations were computationally much more intensive.

shown in the left panels of Fig. 6, for each of the spatial uncertainty conditions. As expected, the spatially uncertain model observer thresholds increased relative to the spatially fixed mouth-only model, and the model-to-human threshold ratios showed a corresponding increase. For spatial uncertainties of 1 pixel, thresholds increased by an average of 1.2 log units relative to the mouth-only model. Model-to-human threshold ratios were between 0.14% and 0.7% for individual human observers, with an overall average of 0.4%. For spatial uncertainties of up to 3 and 5 pixels, thresholds averaged 1.4 and 1.5 log units higher, respectively, than thresholds for the mouth-only model. In addition, average model-to-human threshold ratios for individual observers ranged from 0.2% to 1.6% (mean = 0.8%) for the 3-pixel condition, and from 0.4% to 2.1% (mean = 1.1) for the 5-pixel condition. Although the model-to-human threshold ratios were higher for all of the spatial uncertainty models, only the 1-pixel model was a significantly better fit to the human data than the mouth-only model, with a mean RMSD of 0.18 log units versus 0.22 log units for the mouth-only model ($t(7) = 3.855, p < .05$). The 3- and 5-pixel spatial uncertainty models were significantly worse than the mouth-only model, with mean RMSDs of 0.25 and 0.26 log units for the 3- and 5-pixel models, respectively ($t(7) = -3.010$, $t(7) = -3.299$, respectively; ps < .05).

In the temporal uncertainty simulations, the stimulus on each trial was compared with templates that had been shifted by up to a certain number of frames from the original templates, which again were the mouth-only versions of the word movies. As for the spatial uncertainty simulations, the response on each trial was the word corresponding to the template with the highest cross-correlation with the stimulus. Temporal uncertainties of up to 1, 3, and 5 frames were simulated in the same manner as the spatial uncertainties. To allow for temporal uncertainty in when the word began, 10 additional zero-contrast frames were added to the beginning and ending of the mouth-only word movies. This resulted in an increase of stimulus size from 69 pixels wide × 54 pixels high × 42 frames to 69 pixels wide × 54 pixels high × 62 frames.

Contrast energy thresholds in the temporal uncertainty conditions were higher than those in the mouth-only condition, by an average of 1.4, 1.5, and 1.4 log units for the 1-, 3-, and 5-frame conditions, respectively (right panels of Fig. 6). Model-to-human threshold ratios were similar in

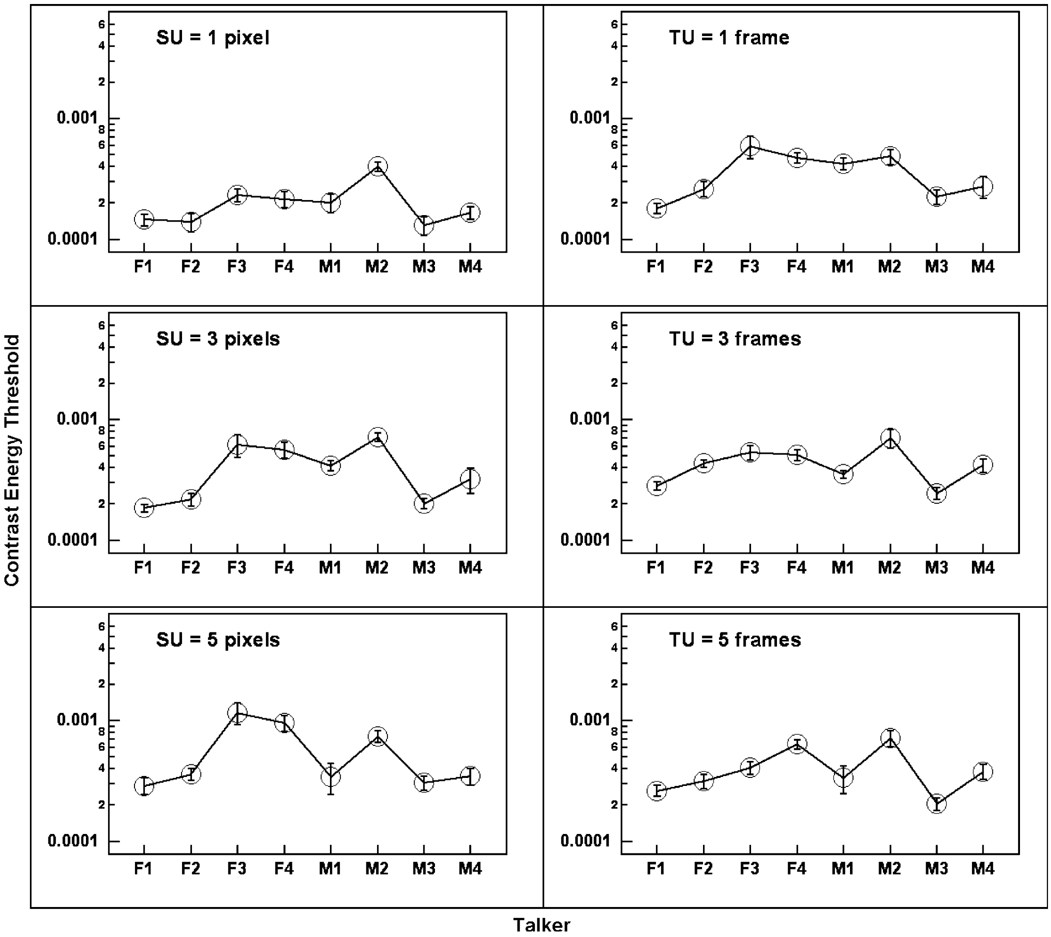

Fig. 6. Contrast energy thresholds for the spatial and temporal uncertainty model observers. Models with spatial uncertainties of up to 1, 3, and 5 pixels are shown in the left panels, and models with temporal uncertainties of up to 1, 3, and 5 frames are shown in the right panels. Error bars correspond to ±1 SD.

the three conditions, ranging from 0.3% to 1.5% (mean = 0.8%) in the 1-frame condition; from 0.3% to 1.7% (mean = 0.9%) in the 3-frame condition; and from 0.3% to 1.6% (mean = 0.8%) in the 5-frame condition. However, the mouth-only model provided a significantly better fit to the human data in terms of RMSD than the 1-frame temporal uncertainty condition, which had a mean RMSD of 0.26 log units ($t(7) = -2.869, p < .05$), and the 3- and 5-frame conditions did not differ significantly from the mouth-only condition, with mean RMSDs of 0.22 and 0.23 log units, respectively ($t(7) = -.654$, $t(7) = -.859$, respectively; $ps < .05$).

The results of simulations modeling spatial and temporal uncertainty suggest that in general, these two types of uncertainty were not able to account for the variability in performance across talkers exhibited by the human observers. Although all of the uncertainty models resulted in higher contrast energy thresholds and model-to-human threshold ratios than the mouth-only model, most of the uncertainty models did not result in lower RMSDs. The only exception was the 1-pixel spatial uncertainty model, which produced a fit to the human data with a significantly lower RMSD value than the mouth-only model. This suggests that spatial uncertainty may have made a small contribution to the pattern of variability across talkers. However, an alternative explanation for why the 1-pixel spatial uncertainty model provided a better fit to the human data is that this model would have had a negative impact on the model observer's use of high spatial frequency information, such as edges between facial features that might have differed slightly in position across templates. That is, introducing uncertainty into the model in this fashion has a similar effect as introducing spatial blur to the templates, which would serve to reduce the contrast of the highest spatial frequencies. It might be the case that human observers were also unable to take advantage of this high spatial frequency information (perhaps due to spatial uncertainty or some other source of blur), and so their pattern of thresholds was better described by a model that could not use all of the high spatial frequency information available in the stimulus.

4.3. Highest contrast pixels

Another factor that could have affected human performance on the word-identification task was the relative salience of specific pixels in the word movies. To examine the impact of pixel contrast on word-identification thresholds, a model observer using the mouth-only templates was presented with stimuli that had been thresholded to include information from only the highest contrast pixels. The lower contrast pixels were set to zero contrast in these stimuli, whereas the highest contrast pixels kept their value. Three highest-contrast pixel conditions were tested, using the top 10%, 20%, or 30% highest contrast pixels. Example still frames from the three conditions are shown in Fig. 7. The word-identification thresholds of this top contrast observer were measured using Monte Carlo simulations of 200 trials per talker, per condition.

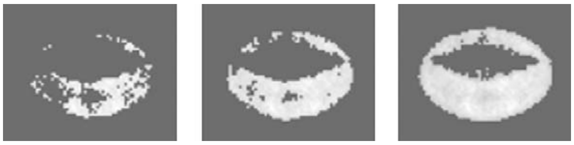

Fig. 7. Top 10% (left), 20% (middle), and 30% (right) highest contrast pixels for the static frame from Talker F4's movie shown on the right in Fig. 5.

Thresholds for the top contrast observers are shown in Fig. 8. Thresholds could be obtained for all talkers in all conditions except for talker M4 in the 10% top contrast condition. As expected, model observer thresholds rose

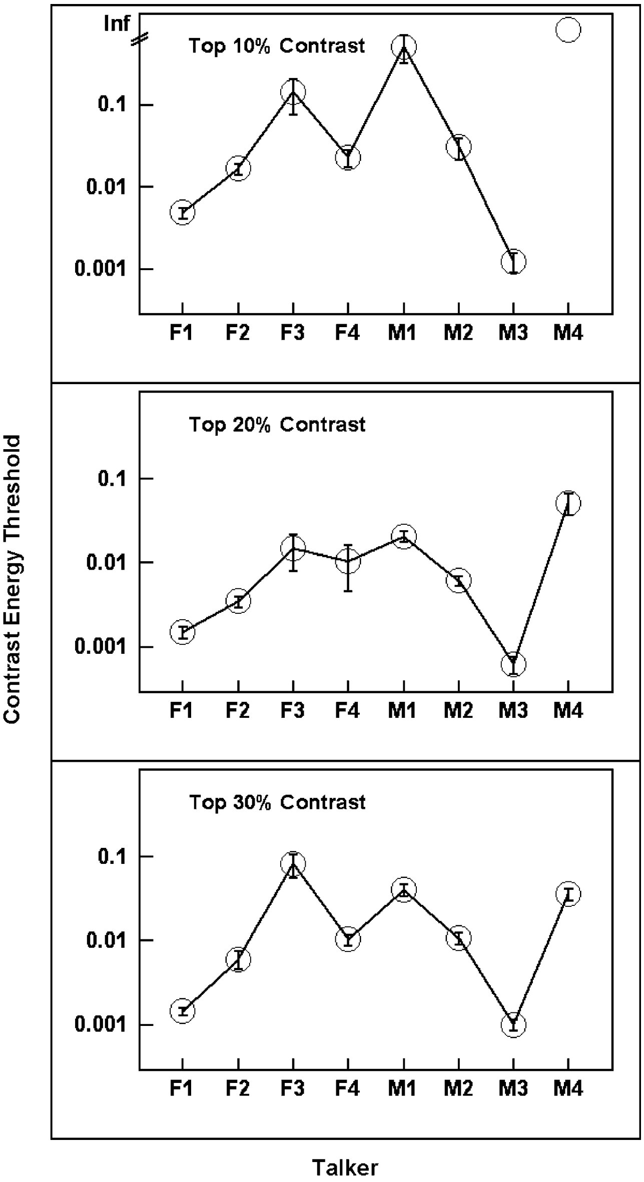

Fig. 8. Contrast energy thresholds for the top 10% (top panel), 20% (middle panel), and 30% (bottom panel) top contrast model observers. Error bars correspond to ±1 SD.

for all three conditions when compared with the mouth-only model thresholds. Thresholds increased by anywhere from 2 to 5 log units for the 10% top contrast condition and by 2–3 log units for the 20% and 30% top contrast conditions. In all three conditions, the model thresholds were sometimes higher than the human thresholds obtained in the whole-face condition. The model-to-human threshold ratios were highly variable across talkers in all three conditions, ranging from 1% to 5% to over 100% across talkers within individual observers. Given that the threshold ratios were far from constant, it is not surprising that the RMSDs of the top-contrast model fits—with means of 1.2, 0.70, and 0.73 log units for the 10%, 20%, and 30% top contrast conditions, respectively—were significantly higher than the mouth-only model RMSDs ($t(7) = -67.682$, $t(7) = -45.084$, and $t(7) = -60.223$, for the 10%, 20%, and 30% top contrast conditions, respectively; $ps < .05$).

Although the model-to-human threshold ratios increased a great deal when the model was restricted to using only the highest contrast pixels, using only the highest contrast pixels does not seem to be a likely human strategy in the word-identification task because the pattern of model observer thresholds provided a poor fit to the human thresholds.

4.3.1. Summary: Models of other perceptual strategies

Using the mouth-only stimuli, the perceptual factors of spatial and temporal uncertainty, and attention to the high contrast pixels only were simulated with model observers. Although most of the models with spatial and temporal uncertainty performed no differently from or were worse than the mouth-only model in fitting the human data, the model with 1 pixel of spatial uncertainty improved significantly on the mouth-only model's fit, suggesting that a small amount of spatial uncertainty may contribute to the variability in human thresholds across talkers. On the other hand, models that were limited to using only the top 10%, 20%, or 30% highest contrast pixels were significantly worse than the mouth-only model, indicating that paying attention to only the high contrast pixels was probably not a strategy used by the human observers.

4.4. The limiting effects of external and internal noise

In the main experiment, the level of external noise added to the stimuli was relatively low compared to many psychophysical experiments aimed at measuring efficiency. We chose to use a relatively low level of external noise in order to insure that we could present our stimuli at contrast levels that spanned observers' contrast energy thresholds. However, given this low level of external noise, it is quite possible that internal noise posed a greater limit on performance than the externally presented noise. Although unlikely, an internal noise that does not share the statistical properties of the externally added noise could potentially have differential effects on observers' performance across the different talker conditions.

One line of evidence against this idea is that the spatiotemporal spectrum of internal additive noise has been shown to be nearly white, with a small relative increase in power at very low spatiotemporal frequencies (Pelli, 1990). However, we addressed these concerns experimentally by having two of the observers who participated in the original experiment run in two additional experimental conditions: One of the additional conditions used stimuli with no added external noise, and the other condition used higher-contrast external noise than the main experiment. The higher-contrast noise had contrast variance of 0.0625 and spectral density of 1.6719×10^{-5} deg^2 at the viewing distance of 130 cm. All other aspects of the experiment were identical to those of the original experiment. We reasoned that if observers' thresholds decreased under no-noise conditions, it would be consistent with the idea that internal noise was not limiting performance in the main experiment. Similarly, if thresholds increased uniformly across talkers in the high-noise condition, it would be consistent with the idea internal noise did not produce the pattern of results that we observed across talkers in the main experiment.

The results of this experiment for both observers are shown in Fig. 9. These data show that observers' thresholds in the no external noise condition were not substantially different from those obtained in the original low noise

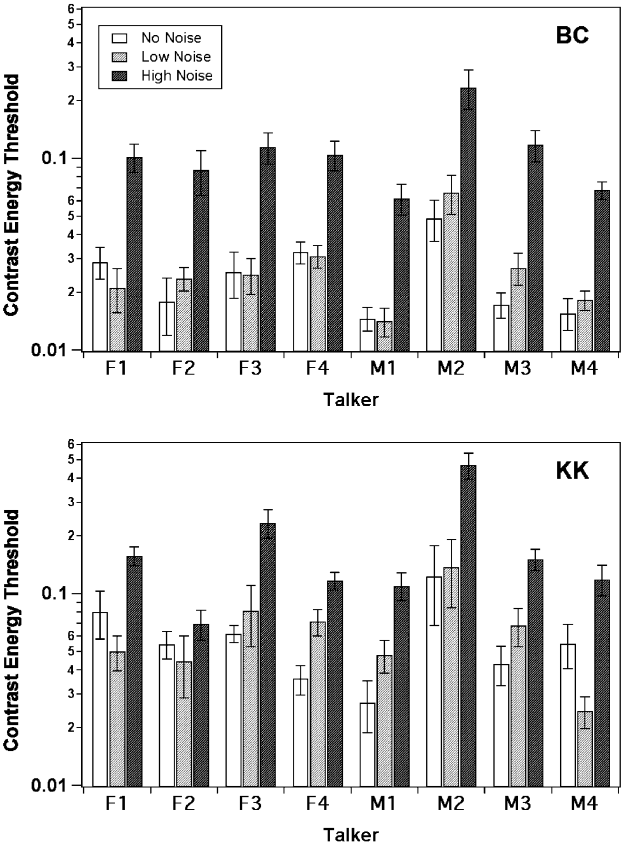

Fig. 9. Contrast energy thresholds for observers BC and KK in the no, low, and high external noise conditions. Error bars correspond to ±1 SD.

condition, suggesting that the level of external noise used in the main experiment may not have had a limiting effect on performance. However, when the observers were tested in the high external noise condition, their thresholds were much higher than in the low external noise condition, indicating that performance under these conditions was limited by the external noise. Moreover, both observers showed similar patterns of variability in thresholds across talkers as in the high, low and zero external noise conditions. The RMSD fit measuring the deviation of the high from the low external noise thresholds across talkers for observer BC was only 0.05 log units; for observer KK it was somewhat higher, at 0.16 log units. Because the pattern of variability across talkers was similar for low and high external noise conditions, it is likely that the modeling results reported above would also hold for higher external noise conditions. Although further simulations would be necessary to draw more definitive conclusions, a preliminary comparison of model fits to the low versus high external noise conditions also provides support for the generalizability of our results. For instance, for observer BC, the whole face model RMSD was 0.19 log units whether it was fit to the low or the high external noise thresholds, and the residuals from the two model fits were not significantly different from each other ($t(7) < 0.0001$, $p > .99$). For observer KK, the whole face model RMSD was 0.22 log units when fit to the low noise thresholds and 0.23 log units when fit to the high noise thresholds, and again the residuals from the two fits were not significantly different ($t(7) < 0.0001$, $p > .99$). Similar results were obtained for both observers using the other models.

5. Summary and conclusions

The primary goal of this study was to ascertain whether cross-talker variability in visual-only speech perception is due to differences in the information available across talkers or to human perceptual strategies that are better suited to some talkers than others. An ideal observer analysis of a visual-only word identification task revealed variability in the information available from the visual-only speech of different talkers. This physical variability in available information did not account for all of the cross-talker variability found in human observer thresholds, however. Thus, a secondary goal of this study was to simulate some simple perceptual strategies and inefficiencies that could account for the additional variability in human performance across talkers. A spatially restricted model observer that used information only from the area around the talker's mouth provided a better description of the human pattern of thresholds than either the ideal, whole face model or a model restricted to using information from only the lower half of the face. The better fit of the mouth-only model suggests that the perceptual strategy of looking only at a talker's mouth when trying to identify visual-only words works better for some talkers than others because of differences in the availability of information around a talker's mouth. Aside from providing a better fit to our human observer data, the mouth-only model is also consistent with eye movement data suggesting that human observers tend to look predominantly at a talker's mouth when trying to understand visual-only speech (Lansing & McConkie, 2003). It appears that human observers may primarily try to "lipread" visual-only speech, whereas in fact also attending to areas of the face other than the lips may be a useful strategy for understanding some talkers.

In an attempt to account for some of the remaining discrepancies between the patterns of thresholds produced by the human observers and the mouth-only model observer, several additional simple perceptual strategies were simulated using the mouth-only movies. These simulations included models of spatial and temporal uncertainty and models of attention to the highest contrast pixels. The only model that improved over the mouth-only model was the model with 1 pixel of spatial uncertainty. The better fit of the spatial uncertainty model to the human data suggested either that humans may have had a small amount of uncertainty about the location of the stimulus, or that they were not using the high spatial frequency information that was blurred out by the small amount of spatial uncertainty. The latter explanation is intuitively appealing because it suggests that human observers were not using minor variations in the position of the facial features across templates. This variation is intrinsic to our stimulus movies and seems on examination of the movies to vary in magnitude across talkers, but it is in general more likely to be artifactual than speech-related. Previous research has also reported that human observers tend not to use high spatial frequencies in auditory-visual speech perception (Munhall, Kroos, Jozan, & Vatikiotis-Bateson, 2004).

Although the simulations account for some of the variability across talkers in human observer thresholds, human efficiency is still quite low in this task. Several factors that are likely to contribute to this low efficiency include the use of dynamic Gaussian noise during stimulus presentation, the optimal use of information at all spatial and temporal frequencies by the ideal observer, the ability of the ideal observer to attend to all stimulus locations simultaneously, and the falloff of visual acuity for non-foveal stimuli. The ideal observer integrates perfectly over time, so the use of uncorrelated dynamic Gaussian white noise caused some of the noise to cancel out and resulted in a higher quality signal for the ideal observer (e.g., Gold, Bennett, & Sekuler, 1999). The ideal observer is also able to use information at all spatial and temporal frequencies, whereas the human observers may selectively focus on particularly spatial or temporal frequencies when perceiving visual speech (Grant & Seitz, 2000; Munhall et al., 2004). Similarly, the ideal observer "attends" simultaneously to all spatial locations, whereas humans may shift their attention during stimulus presentation. In addition, the stimuli subtended a large enough visual angle that they may have been subject to an extrafoveal falloff of visual acuity (Banks, Sekuler, & Anderson, 1991), which would also cause human thresholds to increase.

Despite the low overall levels of efficiency, the present ideal observer analysis and follow-up simulations have provided evidence that talker variability in visual-only speech perception is due both to differences in the physical availability of information across talkers and to human perceptual strategies that are better suited to some talkers than others. In addition, this study has demonstrated for the first time that the ideal observer analysis technique can be successfully applied to the study of visual speech perception. Ideal observer analysis and other quantitative modeling techniques from psychophysics have great potential for adding to our basic understanding of visual-only speech perception strategies in normal-hearing individuals and expert "lipreaders." Results from studies using these techniques could also lead to clinical applications such as specific strategies to improve visual-only speech perception in hearing-impaired individuals.

Acknowledgments

This research was supported by an NSF Graduate Research Fellowship to B.C. and by NIH research Grant EY015787-01 to J.M.G. We thank Gary Rubin and an anonymous reviewer for helpful comments on an earlier version of this manuscript. We like to thank Thad Czuba for his assistance in making the video recordings and choosing the stimuli. Parts of this research were previously presented at the 2005 Vision Sciences Society meeting.

References

Banks, M. S., Sekuler, A. B., & Anderson, S. J. (1991). Peripheral spatial vision: limits imposed by optics, photoreceptors and receptor pooling. *Journal of the Optical Society of America, 8*, 1775–1787.

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

Bernstein, L. E., Jiang, J., Alwan, A., & Auer Jr., E. T., (2001). Similarity structure in visual phonetic perception and optical phonetics. In *Proceedings of the Auditory-Visual Speech Processing (AVSP) workshop,* pp. 104–109.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Campbell, C. S. (2000). Patterns of evidence: investigating information in visible speech perception (Doctoral dissertation, University of California, Santa Cruz, 2000). *Dissertation Abstracts International-B, 61*(7), 3869, 2001.

Cox, R. M., Alexander, G. C., & Gilmore, C. (1987). Intelligibility of average talkers in typical listening environments. *Journal of the Acoustical Society of America, 8*(5), 1598–1608.

Demorest, M. E., & Bernstein, L. E. (1992). Sources of variability in speechreading sentences: a generalizability analysis. *Journal of Speech and Hearing Research, 3*(4), 876–891.

Eckstein, M. P., & Whiting, J. S. (1996). Visual signal detection in structured backgrounds: I. Effect of number of possible spatial locations and signal contrast. *Journal of the Optical Society of America, 1*(9), 1777–1787.

Gagné, J.-P., Querengesser, C., Folkeard, P., Munhall, K. G., & Masterson, V. M. (1995). Auditory, visual, and audiovisual speech intelligibility for sentence-length stimuli: An investigation of conversational and clear speech. *The Volta Review, 97*, 33–51.

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

Geisler, W. S. (2004). Ideal observer analysis. In J. S. Werner & L. M. Chalupa (Eds.), *The visual neurosciences*. Cambridge, Mass: MIT Press, Chapter 52.

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Identification of band-pass filtered letters and faces by human and ideal observers. *Vision Research, 3*(21), 3537–3560.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America, 108*, 1197–1208.

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

Kricos, P. B., & Lesner, S. A. (1982). Differences in visual intelligibility across talkers. *The Volta Review, 8*(4), 219–225.

Kricos, P. B., & Lesner, S. A. (1985). Effect of talker differences on the speechreading of hearing-impaired teenagers. *The Volta Review, 8*(1), 5–14.

Kucera, H., & Francis, W. N. (1967). *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

Lachs, L. (1999). Use of partial stimulus information in spoken word recognition without auditory stimulation. In *Research on Spoken Language Processing Progress Report No. 23* (pp. 82–118). Speech Research Laboratory, Indiana University: Bloomington, IN.

Lachs, L., & Hernandez, L. R. (1998). Update: the Hoosier audiovisual multi-talker database. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 377–388). Speech Research Laboratory, Indiana University: Bloomington, IN.

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, 6*(4), 536–552.

Lesner, S. A. (1988). The talker. *The volta review (special issue: new reflections on speechreading), 9*(5), 89–98.

Massaro, D. W., Cohen, M. M., Campbell, C. S., & Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin & Review, 8*, 1–17.

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review, 97*, 225–252.

Montgomery, A. A., & Jackson, P. L. (1983). Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America, 7*(6), 2134–2144.

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

Pelli, D. G. (1990). The quantum efficiency of vision. In C. Blakemore (Ed.), *Vision: coding and efficiency* (pp. 1–24). Cambridge: Cambridge University Press.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision, 10*, 437–442.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullenix (Eds.), *Talker variability in speech processing* (pp. 9–32). San Diego: Academic Press.

Pitt, M. A., Kim, W., & Myung, I. J. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin & Review, 10*, 29–44.

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

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye* (pp. 3–51). Hillsdale, NJ: Erlbaum.

Tjan, B. S., Braje, W. L., Legge, G. E., & Kersten, D. (1995). Human efficiency for recognizing 3-d objects in luminance noise. *Vision Research, 3*(21), 3053–3069.

Tyler, C. W., Chan, H., Liu, L., McBride, B., & Kontsevich, L. (1992). Bit-stealing: How to get 1786 or more gray levels from an 8-bit color monitor. In B. E. Rogowitz & T. N. Pappas (Eds.), *Societyof Photo-Optical Instrumentation Engineers*, *Vol. 1666*, (pp. 351–364).

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., & Munhall, K. G. (1998). Eye movements of perceivers during audiovisual speech perception. *Perception & Psychophysics, 6*(6), 926–940.

Yakel, D. A., Rosenblum, L. D., & Fortier, M. A. (2000). Effects of talker variability on speechreading. *Perception & Psychophysics, 62*(7), 1405–1412.