Atten Percept Psychophys (2016) 78:1472–1487
DOI 10.3758/s13414-016-1109-4

High visual resolution matters in audiovisual speech perception, but only for some

Agnès Alsius[1] · Rachel V. Wayne[1] · Martin Paré[2] · Kevin G. Munhall[1,2]

Published online: 5 May 2016

Abstract The basis for individual differences in the degree to which visual speech input enhances comprehension of acoustically degraded speech is largely unknown. Previous research indicates that fine facial detail is not critical for visual enhancement when auditory information is available; however, these studies did not examine individual differences in ability to make use of fine facial detail in relation to audiovisual speech perception ability. Here, we compare participants based on their ability to benefit from visual speech information in the presence of an auditory signal degraded with noise, modulating the resolution of the visual signal through low-pass spatial frequency filtering and monitoring gaze behavior. Participants who benefited most from the addition of visual information (high visual gain) were more adversely affected by the removal of high spatial frequency information, compared to participants with low visual gain, for materials with both poor and rich contextual cues (i.e., words and sentences, respectively). Differences as a function of gaze behavior between participants with the highest and lowest visual gains were observed only for words, with participants with the highest visual gain fixating longer on the mouth region. Our results indicate that the individual variance in audiovisual speech in noise performance can be accounted for, in part, by better use of fine facial detail information extracted from the visual signal and increased fixation on mouth regions for short stimuli. Thus, for some, audiovisual speech perception may suffer when the visual input (in addition to the auditory signal) is less than perfect.

Keywords Speech perception · Audiovisual integration · Speechreading skill · Spatial frequency · Eye gaze

Accurate speech comprehension is critical for efficient communication. Although the auditory signal plays an integral part in speech processing, the sight of the talker also conveys information that can be used by the perceiver to decode the message more reliably. The benefits derived from access to the speaker's facial speech information are especially evident in difficult communication environments, such as those involving acoustically degraded signals. Indeed, research has repeatedly shown that when the intelligibility of acoustic speech is impoverished by adding noise (speech-in-noise tasks; SPIN), the concurrent presentation of corresponding visual speech cues improves comprehension dramatically (Cotton, 1935; Rosenblum, Johnson, & Saldaña, 1996; Sumby & Pollack, 1954). The bimodal enhancement of speech perception is probably due, in part, to the complementary nature of the phonological information provided by each of the two sensory channels (Binnie, Montgomery, & Jackson, 1974; Summerfield, 1987). Although the ability to categorize speech signals using visual information alone is generally rather limited (Jackson, 1988; although see Bernstein, Demorest, & Tucker, 2000), visual information can support perception of the auditory stimulus, helping to disambiguate speech sounds that sound similar. Furthermore, the audiovisual speech signal is replete with time-varying features that are common to the acoustical and visual physical channels, such as speech onsets

Electronic supplementary material The online version of this article (doi:10.3758/s13414-016-1109-4) contains supplementary material, which is available to authorized users.

✉ Agnès Alsius
aalsius@gmail.com

[1] Department of Psychology, Queen's University, Humphrey Hall, 62 Arch St, Kingston, Ontario, Canada K7L 3N6

[2] Centre for Neuroscience Studies, Queen's University, Kingston, Ontario, Canada

and offsets, rhythmical patterning, duration or intensity variations (Rosenblum, 2008; Studdert-Kennedy, 1989; Summerfield, 1987). These redundant cues can be exploited by the perceiver when the auditory source is compromised. The bimodal nature of speech perception thus has the potential to be extremely useful for listeners with hearing impairments (Berger, 1972; Payton, Uchanski, & Braida, 1994; Picheny, Durlach, & Braida, 1985) and deaf individuals with cochlear implants (Rouger, Fraysse, Deguine, & Barone, 2008; Rouger et al., 2007).

An intriguing recurrent finding in audiovisual speech-in-noise (AV SPIN) studies is that the enhancement provided by the addition of the visual speech input varies substantially between participants (MacLeod & Summerfield, 1990). Several factors need to be considered when accounting for this large interindividual variability in AV SPIN tasks. One source of variance is individual differences in unimodal performance. That is, high scores in AV SPIN tasks could be simply explained by a superior ability to extract meaningful speech information solely from the auditory source when the signal is compromised by noise and/or a superior ability to extract information from the visual signal alone (which we define as 'speechreading'; MacLeod & Summerfield, 1990). The claim that speechreading ability has an impact in AV performance is supported by evidence of a positive correlation between the two tasks (Grant, Walden, & Seitz, 1998; MacLeod & Summerfield, 1990).

However, a number of studies examining individual differences in AV performance have shown that the size of the benefit afforded by the visual signal cannot solely be explained by differences in unimodal intelligibility levels (Grant et al., 1998; Middelweerd & Plomp, 1987; Smith & Bennetto, 2007; Sommers, Tye-Murray, & Spehar, 2005; Tye-Murray et al., 2008; Watson, Qiu, Chamberlain, & Li, 1996). Indeed, AV performance has been shown to be superadditive, that is, greater than the linear sum of unimodal (auditory and visual) performance scores (Foxe et al., 2013; Grant et al., 1998; Ma, Zhou, Ross, Foxe, & Parra, 2009; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007). One possible source of this variance in the magnitude of superadditive effects is the efficiency of the perceptual mechanism responsible for integrating audiovisual speech information (Grant & Seitz, 1998). This mechanism encompasses the processes by which the brain detects audiovisual correspondences across modalities and the mechanisms that, once such correspondences are detected, combine crossmodal information into a unified percept.

A large number of researchers have explored the factors that contribute to the individual differences observed in speechreading ability and the sources of variance underlying differences in the ability to combine audiovisual information. Studies of individual differences in speechreading ability have shown a number of perceptual and cognitive abilities related to a superior capacity to extract meaningful information when only the talker's face is available. Cognitive factors include spatial working memory, verbal processing speed (Feld & Sommers, 2009; Lidestam, Lyxell, & Andersson, 1999), and perceptual synthesis (the ability to generate complete representations from partial information, as shown by the strong correlations between speechreading and auditory-only SPIN performance; Watson et al., 1996). Perceptual factors, such as the ability to encode speech elements at a phonological level (Bernstein et al., 2000; although see Sommers et al., 2005), as well as the speed with which this is accomplished (Gagné, Charbonneau, & Leroux, 2011), have been shown to independently contribute to the variance in speechreading proficiency. This latter result suggests that proficient speechreaders are able to extract more information from the visual speech input and do so at a greater speed, thereby enabling more efficient use of visual speech information. In a recent study, Wilson, Alsius, Paré, & Munhall (2016) investigated the visual components of visual speech identification that are most important for optimizing speechreading. In particular, we explored the impact of fine facial detail on speechreading ability using a spatial frequency filtering technique to manipulate the resolution of silent clips of a talker pronouncing vowel–consonant–vowel (VCV) utterances. The spatial frequency filtering technique involves removing certain spatial frequencies of the image in order to determine which frequencies are necessary for visual and audiovisual speech perception (Munhall, Kroos, Jozan, & Vatikiotis-Bateson, 2004). The spatial frequency of an image is the rate of change of contrast per spatial unit, and, in general, higher frequencies represent more detailed information. We presented participants with various clips at different cutoff frequencies, along with unfiltered clips, comparing the performance of the 12 participants with the highest speechreading ability to the performance of the 18 participants with the lowest speechreading skill. We found that proficient speechreaders were more negatively affected by the removal of high spatial frequency information than were poor speechreaders. That is, the saturation point (i.e., the frequency at which the addition of increased visual resolution becomes redundant for speech intelligibility) occurred at 22.3 cycles/face (c/f) for the best speechreaders and at 15.5 c/f for the worst speechreaders. The finding that poor speechreaders gain significantly less from increasing spatial frequency information than skilled speechreaders do suggests that the variability in speechreading proficiency can be accounted for, in part, by the use of fine-detailed facial information extracted from the visual signal. The more efficient use of high-frequency visual information by some participants could stem from a superior ability to map the optical signals into visual speech representations (Wilson et al., 2016).

While it is known that individuals with better speechreading proficiency benefit more from high frequency spatial information, it is unclear whether this benefit from spatial resolution also extends to individuals who benefit most from the addition of visual speech information in the AV SPIN signal. The fact that some studies have shown a positive correlation between speechreading abilities and visual enhancement in AV SPIN could lead one to assume that audiovisual and visual speech processing rely on the same mechanisms of visual analysis. Therefore, individuals would be equally affected by the removal of high-frequency speech information in visual-only and SPIN tasks.

However, findings suggesting that speechreading might require greater visual resolution than audiovisual speech (Lansing & McConkie, 2003; Wilson et al., 2016) challenge this assumption. Higher spatial frequency information might not be as important in AV SPIN conditions, because the addition of auditory information could provide cues that are complementary to those found in the high resolution range of the image (e.g., seeing the exact position of the tongue relative to the teeth might not be as important when there are acoustic cues available that might facilitate the recognition of the linguistic unit). Indeed, a large body of evidence shows that, for the average listener, fine facial detail is not critical for visual enhancement when auditory information is available (Jordan & Sergeant, 1998, 2000; MacDonald, Andersen, & Bachmann, 2000; Munhall et al., 2004, Munhall & Vatikiotis-Bateson, 2004; Neely, 1956) or for the McGurk effect to occur (Paré et al., 2003; Wilson et al., 2016). Munhall et al. (2004) degraded the image of a talker by applying low- and band-pass filter cutoffs, showing that the addition of only very low spatial frequencies is required to observe visual enhancement for SPIN comparable to that produced by unfiltered images, as the saturation point occurred at 7.3 c/f. Similarly, other studies have shown that speech perception is reduced—but remains effective—when facial images are spatially degraded by quantization (e.g., Campbell & Massaro, 1997; MacDonald et al., 2000), visual blur (Thomas & Jordan, 2002; Thorn & Thorn, 1989), or increased stimulus distance (Erber, 1971; Jordan & Sergeant, 2000; Small & Infante, 1988).

However, none of these studies manipulating visual resolution in audiovisual speech in noise compared participants as a function of their AV speech perception abilities, thus masking potential individual differences in the visual mechanisms involved in extracting visual cues from the face. Although, on average, participants do not seem to benefit from the addition of higher frequency spatial information, it is possible that those with superior visual gains actually do rely more on this high resolution visual information (i.e., high-frequency range) in order to more accurately map optical signals onto visual speech representations or to combine the corresponding auditory and visual cues more efficiently (e.g., by providing more visual cues that indicate audiovisual correspondences).

The goal of the present study was to extend the results of Wilson et al. by examining the effects of low-pass spatial frequency filtering on performance in an AV SPIN task, comparing participants as a function of their visual gain (i.e., the contribution of the visual information to speech recognition) rather than speechreading ability. Understanding how visual information is integrated with auditory speech information requires a detailed specification of the content of the visual speech information, the mechanisms responsible for visual information processing, and the interaction between the two (how the visual information is processed as a function of its content). Regarding the former (i.e., content), this study aims at exploring the spatial frequency bands that capture the critical phonetic information for the most successful combination with auditory information. Regarding the latter (i.e., processing), the study aims at elucidating the interindividual differences in the ability to make use of this available phonetic information. Specifically, we wanted to determine if the frequency saturation point differs between those participants that most benefit from the visual information (i.e., participants with high visual gain; HVG) and those who benefit the least from the presence of visual information (i.e., participants with low visual gain; LVG). Finding similar saturation points for participants with high and low visual gains would suggest that they use the available visual information in a similar way; that is, they extract similar cues from the visual signal. Alternatively, observation of different perceptual patterns as a function of filter cutoff would provide critical information regarding the frequency range that most contributes AV SPIN proficiency.

In order to test whether the extraction of visual speech cues varies as a function of the availability of contextual cues in the speech utterance, we sampled AV SPIN performance and speechreading ability across two levels of linguistic complexity: isolated words (reduced contextual cues) and sentences (richer contextual cues). At the lowest linguistic levels (e.g., phonetic features of speech sounds, consonants, vowels, syllables), speechreading and AV SPIN performance is heavily dependent on visual-analytic processes (Grant et al., 1998). However, as the structure of the linguistic stimuli increases (i.e., in words), speechreaders may rely more on other types of cues, such as phonological and lexical information (e.g., Mattys, Bernstein, & Auer, 2002; Strand & Sommers, 2011). In continuous speech streams (i.e., sentences), however, the borders between words become less visible and speechreaders may use other sources of linguistic redundancy, such as morphosyntactic and semantic information (Bradarić-Jončić, 1998; Tye-Murray et al., 2008). A likely possibility, therefore, is that perceivers use highly detailed information of the speaking face in contexts with less contextual cues (i.e., syllables and words) but rely less on such low-level cues when presented with sentences. If this is the case, we should find different effects of filter cutoffs in words and sentences.

In order to determine the degree by which the variance in AV SPIN can be accounted by speechreading ability, we screened participants' ability to speechread words and/or sentences in an unfiltered visual-only block. We also used eye tracking, in order to assess potential differences in gaze strategy for gathering information from a talker's face, between HVG and LVG perceivers. Previous results by Lansing and McConkie (2003) showed no improvement in speechreading as people directed their gaze to the mouth. To our knowledge, however, no one has explored gaze behaviors differences between HVG and LVG groups in audiovisual contexts.

Method

Participants

Fifty-one participants completed both the word and sentence sessions. Twelve additional participants completed only the word session, and a different set of 12 participants completed only the sentence session for a total of 63 participants in each session (Words: 54 females, mean age = 19.95 years, age range: 18–31; Sentences: 55 females, mean age = 20.23 years, age range: 17–31). Eye-tracking analyses included data collected from 37 participants in the word session and 37 participants in the sentences session. All participants were native English speakers with no known hearing, speech, or language disorders as well as normal or corrected-to-normal vision.

Stimuli

Audiovisual stimuli were prepared from digital video recordings of a female, native-North American English speaker (full frontal view of the entire head) articulating a set of 161 familiar words (i.e., monosyllabic nouns) and 147 sentences (see the Supplemental materials to access the full set of materials).

The word corpus was initially generated using the MRC Psycholinguistic database. We then used SUBTLEXus to determine the frequency of each item (word count per million words; Brysbaert & New, 2009). We constructed seven sets of 22 words each, matched in the mean logarithm of word frequency (mean = 1.41, *SD* = 0.04) from the corpus. The sentence corpus was taken from Wild et al. (2012). We constructed seven sets of 20 sentences, matched for the number of key words (mean = 5.94, *SD* = 0.18) and the logarithm of the sum of word frequency (SUBTLEXus; mean = 3.74, *SD* = .05). The sets of words and sentences were assigned to one of the seven experimental conditions (five filter cutoffs, unfiltered, auditory-only). The assignment of sets to conditions was counterbalanced across subjects such that all stimuli occurred equally often in all conditions across participants, and participants were presented with each speech stimulus only once during the study. An additional set (fixed) of 22 words and 20 sentences was used for the speechreading (visual only) screener test.

We transferred the video recordings to a computer as image sequences, converting them to grayscale and low-pass filtering them with a custom MATLAB program using a second-order Butterworth filter. The following frequency cutoffs (Fc) were used to create five sets of conditions: 5, 11, 17, 23, and 29 horizontal cycles/face (c/f). We also included an unfiltered condition in which the images were not filtered and thus contained the full spatial frequency spectrum at the resolution of the monitor (i.e., 228 c/f; see Fig. 1).

The audio level was normalized using custom MATLAB software. We decreased speech intelligibility by adding a commercial multitalker babble track (Auditec, St. Louis, MO) with a signal-to-noise (S/N) ratio of -2 dB for words and -3 dB for sentences. These S/N ratios were selected because they produced auditory-only accuracy levels around 30 % in a pilot experiment, thus ensuring that performance would not reach ceiling in the audiovisual conditions. The overall intensity was 60 dB (SPL A). The stimuli were then compiled and run on a PC (Intel Core) using Experiment Builder software. A voice recorder (Edirol by Roland, R-09HR) was used to record participants' responses throughout the study.

Equipment

Participants were tested in a single-walled sound isolation booth (Eckel Model C-17), seated at a table with their head positioned in a chin rest with their eyes approximately 57 cm from a 20-inch video monitor (JVC Model TM-H1950G). The stimuli were displayed with Experiment Builder, with a resolution of 640 × 480 pixels. The auditory signal was amplified (InterM R300 reference amplifier) and played through speakers (Paradigm Reference Studio 20), located on each side of the screen. Eye position was monitored using an EyeLink II eye-tracking system (SR Research, Osgoode, Canada) using dark pupil tracking with a sampling rate of 500 Hz.

Procedure

Participants from which we collected eye-tracking data were fitted with the EyeLink II headband. The cameras were then adjusted and focused, and participants performed a nine-point calibration and validation procedure. For all subjects in the eye-tracking data collection, the maximum allowable average error was 1.0 visual degrees, and maximum error on a single point was 1.5 visual degrees in the validation phase.

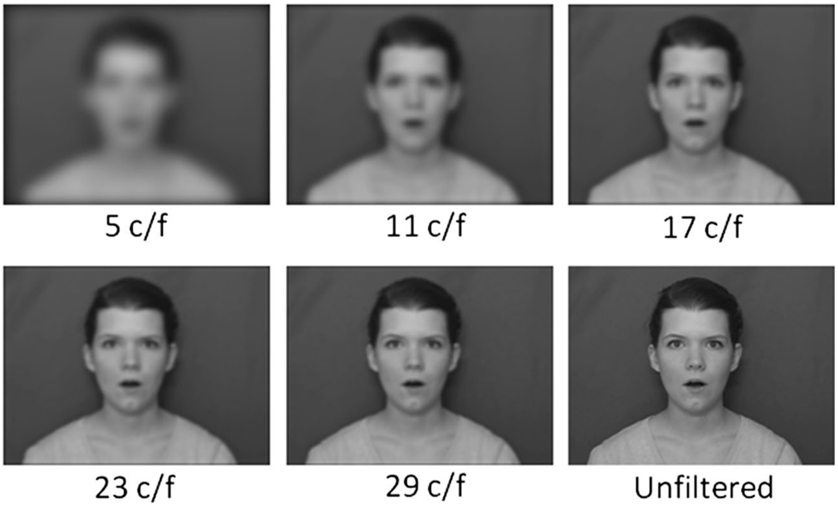

Fig. 1 Single images of the talker filtered at Fc of 5, 11, 17, 23, 29 c/f, and the same image, unfiltered (i.e., 228 c/f)

Before each session, all participants completed a practice block of seven trials with either words or sentences (in preparation for the experimental setting), with one item in each condition (i.e., five different Fc, one auditory only, one unfiltered). Participants were then immediately presented with 154 words (22 words per condition) in one session and 140 sentences (20 per condition) in another session. The words and sentences within each session were presented in pseudorandom order (i.e., seven different randomly ordered lists were created). A drift correction (i.e., the correction for the disparity between the location of the fixation point and the actual fixation maintained by the participant) was performed before each trial. The order of the sessions was counterbalanced for those participating in both sessions. Participants were instructed to watch the screen during the video clip and to repeat verbally as much information as they understood from the speech signal. The experimenter manually triggered the next stimulus presentation after each trial. At the end of each session, participants performed a visual-only screener, in which they were presented with either words or sentences (in both the word and sentence sessions).

Data analysis

Scoring Participants' responses were scored by a condition-blind native English speaker. For sentences, only keywords were considered (see the Supplemental materials). Homonyms were considered correct. Noun pluralisation was ignored, both for the sentences and words sets (Bench & Bamford, 1979).

Visual gain We assessed AV SPIN ability for each participant by calculating the average score in the unfiltered condition (i.e., the most ecologically relevant condition) for words and for sentences separately. Significant positive correlations were found between audiovisual scores in the unfiltered condition and the auditory-only score in sentences: r = .262, p = .04 (words: r = .09, p = .495). Because the goal of the study was to explore individual differences in the ability to extract linguistic content *from the image*, we decided to correct the data by using an index that accounted for interindividual differences in the ability to understand speech auditorily in noisy conditions. Specifically, the relative benefit in accuracy due to the additional presence of the visual signal in combination with the auditory signal (i.e., visual gain) was calculated by using [Audiovisual score - Auditory-only score/1 - Auditory-only score] for each condition and session separately (e.g., Sumby & Pollack, 1954; see also Grant, 2002). The visual gain score calculated here was used in subsequent analyses. Note that, given that the auditory signal-to-noise ratio (SNR) was held constant across Fcs, auditory-only abilities should have little impact on the rate of change in performance moving along the spatial frequency (Fc) visual gain function, which was the focus of the study.

Gaze behavior To analyze gaze position with respect to the talker's face in the video, three regions of interest (ROI; eyes, nose, mouth; see Fig. 2) were defined using a custom program, following frame-by-frame coding of all stimuli with MatchMovie software (Autodesk Maya 2008). The eyes and the nose ROIs remained fairly constant in size across the frames, but the mouth ROI varied in size as the talker opened and closed her mouth throughout the utterance. Eye-movement raw data were transformed with the custom program to produce individual data files that provided the location of participant's gaze for each sample time (sampling frequency 500 Hz) for each utterance. The duration was computed by summing up the number of samples in which participant's gazed at one specific ROI and dividing it by the total sampling time of the utterance.

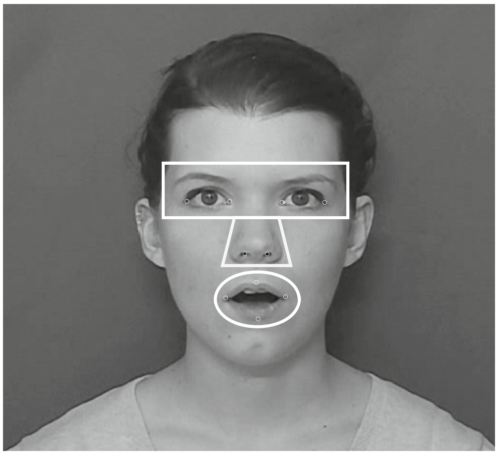

Fig. 2 Depiction of the ROIs

Results

Response distributions for the audiovisual unfiltered, auditory-only condition, the visual-only screener, along with the distributions for the corrected AV scores (i.e., visual gain) can be seen in Fig. 3.

Average visual gain as a function of filter cutoff In general, performance significantly increased with increasing Fc (i.e., improving resolution), both for words and sentences, as confirmed in a repeated-measures ANOVA with Fc (5, 11, 17, 23, 29 c/f and unfiltered) as a factor, $F(5, 310) = 81.18$, $p < .001$, $\eta_p^2 = .57$; $F(4.12, 255.21) = 163.03$, $p < .001$, $\eta_p^2 = .72$, respectively. Paired *t*-tests using a Holm-Bonferroni correction revealed significant differences between the unfiltered condition and all the other cutoffs (all *p*s < .006) in both words and sentences. This shows that the addition of the high-frequency information globally increased participants' performance. Pairwise comparisons were also computed for adjacent Fc. In words, significant differences were found between 5 c/f and 11 c/f, and between 17 c/f and 23 c/f. In sentences, all the contrasts were significant, except for between 23 c/f and 29 c/f.

Visual gain as a function of filter cutoffs: Individual differences We tested how individual differences in visual gain were reflected across spatial frequency by correlating participants' mean correct responses in the unfiltered condition with the mean of correct responses in each of the filtered conditions. Pearson correlations were found to be significant for all Fc, both in words and sentences (see Table 1; *r* range for words: from .38 to .55; for sentences: from .42 to .76). These results indicate that the relative performance of participants was consistent across spatial frequency cutoffs, even at a low visual resolution (i.e., 5 c/f).

Separate analyses aimed at comparing participants as a function of their ability level in the unfiltered condition in order to explore their rate of change in performance as moving along the spatial frequency (Fc) visual gain function. This analysis confirmed the existence of a group of individuals who were responding to the higher visual resolution information in a distinct manner. In these analyses, we estimated the number of clusters in the dataset using both their visual gain and their respective difference in visual gain between the two highest filters (i.e., 29 c/f and unfiltered stimuli) with the gap statistic method (Tibshirani, Walther, & Hastie, 2001). Note that these two parameters were positively correlated (Pearson *r* = .58 and .5 for word and sentence data, respectively). This method compares the actual dispersion of the data sample within putative clusters identified with the *k*-means algorithm with a reference dispersion; the optimal number of clusters is taken as that for which the difference between actual and reference dispersion is maximal. For this study, the actual dispersion was calculated using compactness as the measure of dispersion, and the reference dispersion was estimated with 5,000 bootstrap iterations. As in Tibshirani et al. (2001), the optimal number of clusters was taken as the mode of the distribution of the optimal number of clusters obtained after 50 trials of the gap statistic. The clusters and their centroids were identified with the *k*-means algorithm (Hartigan & Wong, 1979), which was set to minimize the squared Euclidean distance between data samples. Critically, these analyses identified the optimal number of clusters as two for both the word ($N_{HVG} = 37$, $N_{LVG} = 26$) and the sentence ($N_{HVG} = 22$, $N_{LVG} = 41$) data (see Fig. 4). The number of clusters identified with the gap statistic method was validated with the silhouette partitioning technique (Rousseeuw, 1987): The average silhouette width was maximal when considering the two-cluster solution, compared to a larger number of clusters. These results supports the view that the participants' performances were not homogeneous across the high spatial frequency content of the data: A subgroup of participants continue to improve performance when they are given additional visual information (i.e., HVG group) but others do not seem to be able to take advantage of the subtle cues provided by high spatial frequency information (i.e., LVG group).

Figure 5 plots group performances (i.e., HVG, LVG) in the AV unfiltered condition and auditory-only and visual-only conditions. A mixed ANOVA with visual gain group (HVG, LVG) and condition (AV, auditory only, visual only) as factors revealed significant main effects of condition and visual gain group and significant interactions, for words and sentences (all *p*s < .001). As expected, the percentage of correct responses in the audiovisual condition was significantly reduced in the LVG group as compared to the HVG group, both in words, $t(61) = -5.93$, $p < .001$, $d = 1.52$, and sentences, $t(61) = -9.08$, $p < .001$, $d = 2.4$. No significant differences were found in the auditory-only condition in words or sentences, $t(61) = 1.72$, $p = .09$, $d = .43$; $t(61) = .31$, $p =.76$, $d = .09$, respectively. Differences between the two groups in the visual condition

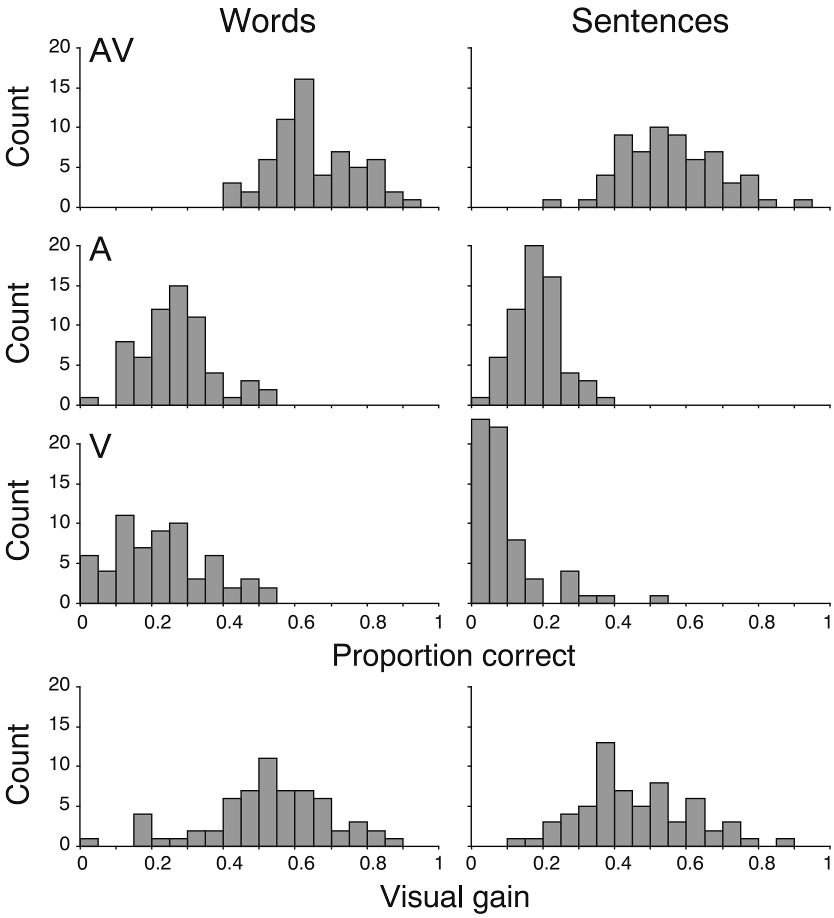

Fig. 3 Histograms showing distribution of participants' correct responses in the audiovisual (uncorrected) unfiltered condition, auditory-only condition, and visual-only screener test, for words and sentences. The bottom panels show the distribution for visual gain (i.e., corrected scores, [AV – A]/[1 – A]). The *x*-axis shows the bins for the proportion of correct responses and the *y*-axis the count (number of participants) in each bin

was found in sentences, $t(61) = -2.2$, $p = .04$, $d = .63$, but not in words, $t(61) = -.48$, $p = .63$, $d = 1.25$.

Figure 5 also shows participants' performance as a function of their visual gain, across the different Fcs. A 2 × 6 split-plot ANOVA showed significant main effects of visual gain group, both for words, $F(1, 61) = 10.91$, $p = .002$, $\eta_p^2 = .15$, and for sentences, $F(1, 61) = 28.29$, $p < .001$, $\eta_p^2 = .32$. Thus, as seen in Fig. 5, the HVG group outperformed the LVG group at all filter cutoffs (except for the 29 c/f Fc in words). There was also a significant interaction of visual gain group × Fc, words: $F(5, 305) = 9.52$, $p < .001$, $\eta_p^2 = .135$; sentences: $F(3.93, 239.72) = 11.87$, $p < .001$, $\eta_p^2 = .163$, suggesting differences in benefit from spatial frequency information by the HVG and the LVG groups. Pairwise comparisons of performance of the two groups in the unfiltered versus each filter condition (corrected with Holm-Bonferroni) revealed significant

Table 1 Pearson correlation coefficients between the visual gain in the unfiltered condition and each filter cutoff, between speechreading and visual gain in the unfiltered condition, and between speechreading and auditory-only condition, for words and sentences

		Words	Sentences
Visual Gain	5 c/f–Unfiltered	$r = .426$, $p = .001$	$r = .417$, $p = .001$
	11 c/f–Unfiltered	$r = .536$, $p < .001$	$r = .631$, $p < .001$
	17 c/f–Unfiltered	$r = .441$, $p < .001$	$r = .524$, $p < .001$
	23 c/f–Unfiltered	$r = .512$, $p < .001$	$r = .762$, $p < .001$
	29 c/f–Unfiltered	$r = .382$, $p = .002$	$r = .732$, $p < .001$
Speechreading–Visual Gain (Unfiltered)		$r = .335$, $p = .007$	$r = .585$, $p < .001$
Speechreading–Auditory only		$r = -.064$, $p = .62$	$r = .215$, $p = .09$

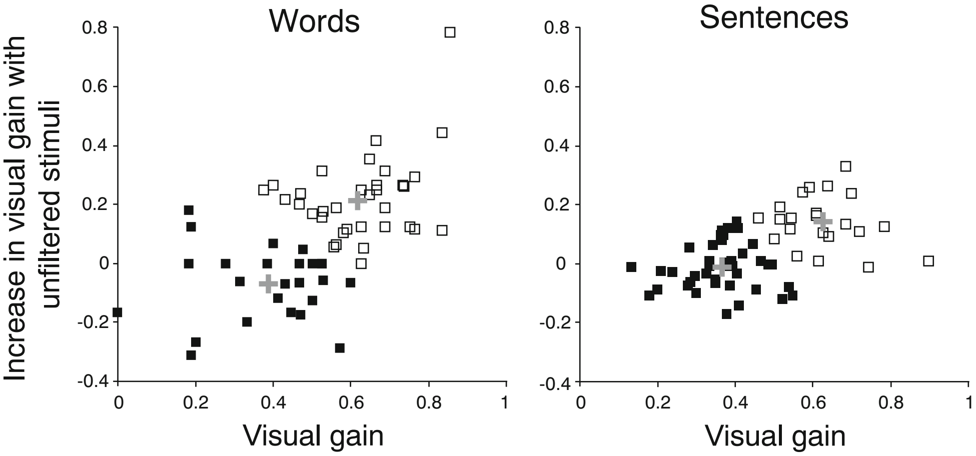

Fig. 4 Scatter plot with participants' visual gain and their respective difference in visual gain between the two highest filters, 29 c/f, and unfiltered stimuli. The gap statistic method identified the optimal number of clusters as 2 for both the word ($N_{HVG} = 37$, $N_{LVG} = 26$) and the sentence ($N_{HVG} = 22$, $N_{LVG} = 41$). The empty squares correspond to the HVG group, and the filled squares to the LVG group. The crosses correspond to the centroids

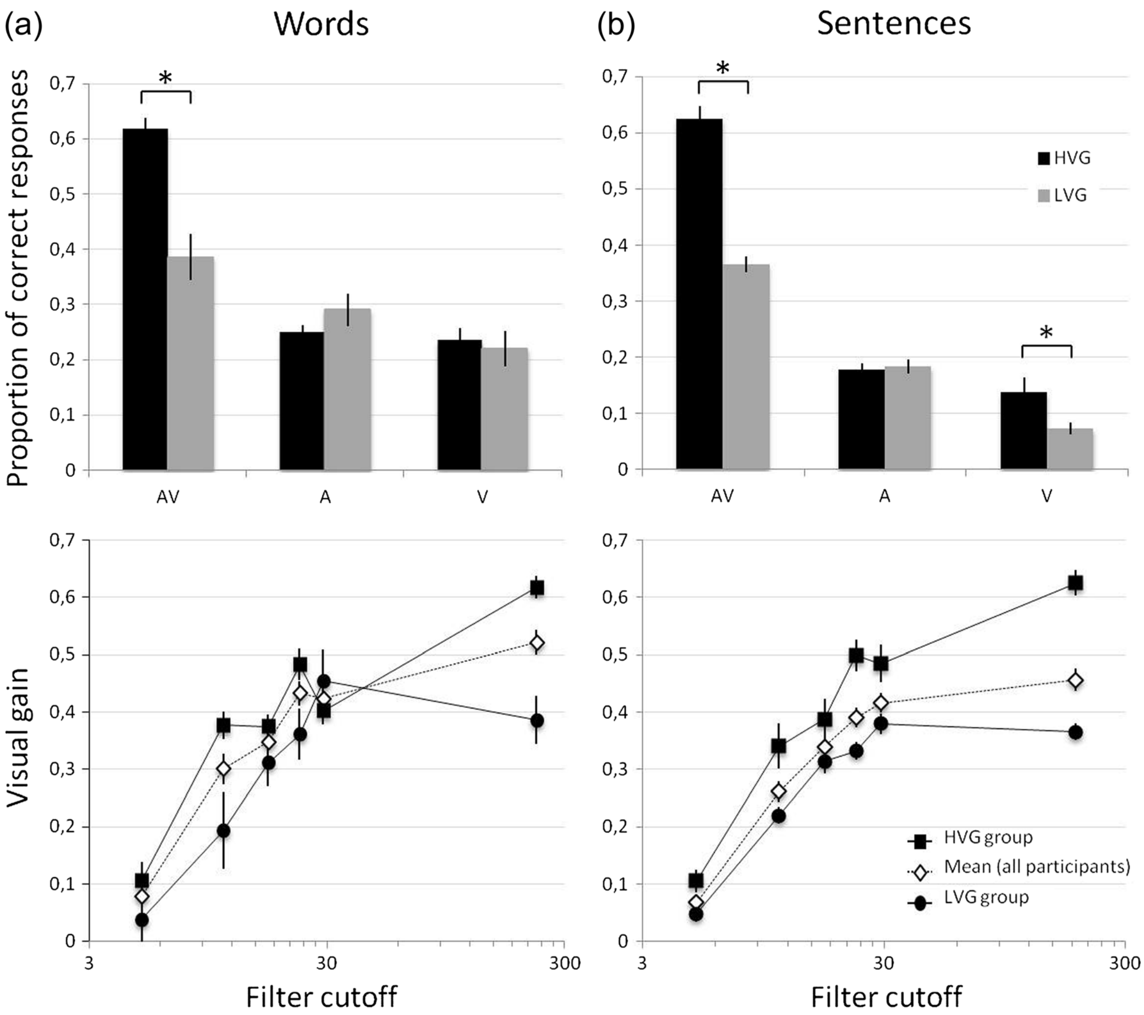

Fig. 5 Upper panels: Proportion of correct scores for the AV unfiltered condition (uncorrected), auditory-only (A) and visual-only (V) condition of the two groups, for (**a**) words and (**b**) sentences. Bottom panels: Performance of participants as a function of their visual gain, across the filter cutoffs. The performance of participants with the highest visual gain (i.e., HVG) was contrasted with the performance of participants with the lowest visual gain (i.e., LVG), for (**a**) words and (**b**) sentences

differences between the unfiltered and all the filter cutoffs in the HVG group, both for words and sentences (all *p*s < .001). For the LVG group, word performance in the unfiltered condition was significantly higher than performance in 5 c/f Fcs, $t(25) = 10.17$, $p <. 001$, 11 c/f, $t(25) = 4.61$, $p < .001$, and 17 c/f, $t(25) = 2.6$, $p = .015$, but not for performance in 23 c/f, $t(25) = .98$, $p = .33$, and 29 c/f, which was actually found to be significantly higher, $t(25) = -2.9$, $p = .008$. For sentences, performance in the unfiltered condition was significantly higher than performance in 5 c/f, $t(40) = 19.04$, $p < .001$, 11 c/f, $t(40) = 9.43$, $p < .001$, 17 c/f, $t(40) = 3.05$, $p = .004$, but not for performance in 23 c/f, $t(40) = 2.61$, $p = .013$; nonsignificant after Holm-Bonferroni correction, and 29 c/f, $t(40) = -1.15$, $p = .26$. These results suggest that performance in this group reached plateau at a much lower spatial frequency filtering. Note that the same pattern of results (with a boost in performance in the unfiltered vs. the highest level of filter cutoff; i.e., 29 c/f for some participants but not for others) is observed when using the uncorrected audiovisual scores in the analyses.

Although the material sets were matched for word frequency, the possibility still remains that some sets were more visually distinct (and thus easier to speechread) than others. In order to confirm that differences in visual gain group could not be explained by uncontrolled differences in level of difficulty in specific sets of materials, we ran a control study with a new set of participants ($N = 24$) in which they were presented with all the words of each set. Because the observed individual differences with increasing spatial frequency were largest with word stimuli, we tested the possibility of stimulus set distinctiveness with only this stimuli set. Half of the words in each set were presented audiovisually and the other half auditorily. We calculated the visual gain obtained by each participant in each set and the visual gain (across participants) for each word. The within-participants ANOVA for individual visual gain did not significantly differ between the different word sets, $F(3.36, 77.24) = 1.84$, $p = .14$, $\eta p^2 = .07$, nor did we observe significant differences for a one-way between-words ANOVA analysis, $F(6, 153) = 1.14$, $p = .34$, $\eta^2 = .04$. These results suggest that word sets were not different in terms of difficulty.

Correlations between visual gain and speechreading In order to further establish the degree to which speechreading ability determines the degree to which listeners benefit from the addition of visual speech information in audiovisual contexts, we computed correlations of performance in the visual-only screener and the mean proportion of visual gain in the unfiltered audiovisual condition (words, $N = 63$; sentences $N = 63$). Significant correlations were observed for words ($r = .335$, $p = .007$) and for sentences ($r = .585$, $p < .001$; see Table 1).

Correlation between speechreading and auditory-only performance In order to test whether speech recognition is, in part, explained by the existence of a modality-independent source of variance in speech-recognition abilities, as previously suggested (see Watson et al., 1996), we correlated participants' performance in the unfiltered visual-only condition to the mean proportion of correct identifications in the auditory-only condition (words, $N = 63$; sentences, $N = 63$; see Table 1). The Pearson correlation between performance in the visual-only screener and in the auditory-only condition was nonsignificant for both words ($r = -.064$, $p = .62$) and for sentences ($r = .215$, $p = .09$).

Correlation between words and sentences In order to determine whether participant's visual gain was stable across materials of different complexity, we computed the correlation between participants' performance for words and sentences in the unfiltered condition. Note that only a subset of participants—those that took part in both sessions—were used for these analyses ($N = 51$). A Pearson correlation revealed a significant correlation between words and sentences in the visual gain scores for the AV SPIN unfiltered condition ($r = .52$, $p < .001$). Similarly, significant correlations were observed between words and sentences in the visual-only screener ($r = .58$, $p < .001$).

Gaze behavior as a function of filter cutoff in AV SPIN In order to determine which parts of the talker's face participants gaze at to maximize speech intelligibility as a function of Fc, we measured the time participants spent gazing at each ROI for both words (Fig. 6a, $N = 37$) and sentences (Fig. 6b, $N = 37$). As can be seen in Fig. 6, and just as in Wilson et al. (2016), when the resolution of the image was high (i.e., there was high spatial frequency information), there was an increased tendency to focus on the mouth and a decreased tendency to focus on the nose, both in words and in sentences. Very little time was spent gazing at the eye region. A within-subjects one-way ANOVA was conducted for each ROI to analyze the effect of filtering on the percentage of time spent in each ROI. For words, a significant effect of Fc was found for every ROI—eyes: $F(2.53, 91.04) = 9.29$, $p < .001$, $\eta_p^2 = .20$; mouth: $F(2.67, 96.19) = 67.09$, $p <.001$, $\eta_p^2 = .65$; nose: $F(2.57, 92.69) = 64.14$, $p < .001$, $\eta_p^2 = .64$. In sentences, significant effects of Fc were also found in every ROI—eyes: $F(2.42, 82.31) = 10.50$, $p < .001$, $\eta_p^2 = .24$; mouth: $F(2.92, 99.30) = 34.83$, $p < .001$, $\eta_p^2 = .51$; nose: $F(3.04, 103.31) = 45.30$, $p < .001$, $\eta_p^2 = .57$, again, with participants gazing more at the mouth and eye regions and less at the nose when the resolution increased.

Visual gain group and gaze behavior In order to compare eye-gaze behavior as a function of visual gain group, we measured the amount of time participants in the two groups spent gazing at each ROI (see Fig. 7). Note that only those participants in the groups whose gaze was monitored (HVG group: words, $N = 20$; sentences, $N = 13$; LVG groups: words, $N = 17$; sentences, $N = 22$) were included in these analyses.

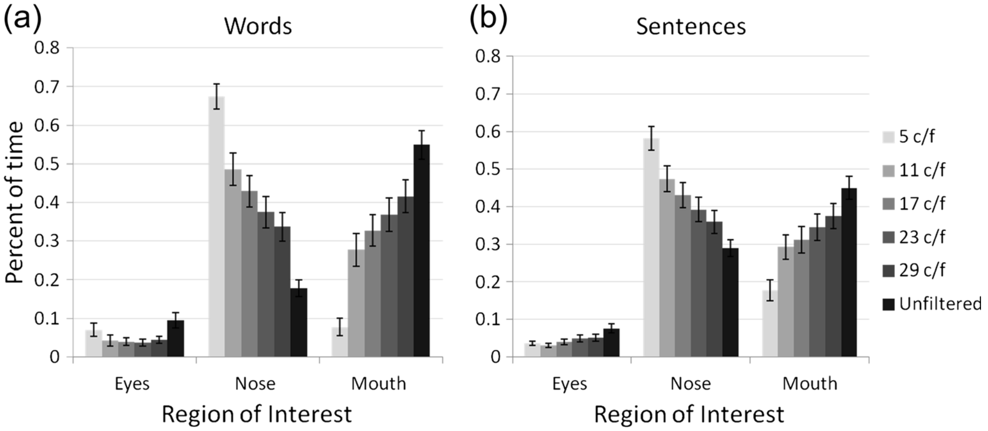

Fig. 6 Percentage of time (±*SE*) spent with eye gaze directed to each region of interest (ROI), as a function of Fc, for words and sentences

Words A mixed ANOVA with visual gain group as the between-subjects factor and Fc as the within-participants factor conducted for each ROI revealed a main effect of visual gain group in the mouth, $F(1, 35) = 10.67$, $p = .002$, $\eta_p^2 = .23$, and the nose, $F(1, 35) = 9.05$, $p = .005$, $\eta_p^2 = .205$, regions. The HVG group spent more time on the mouth region (62 % of the time, overall) and less time on the nose (14 % of the time, overall) compared to the LVG group (47 % and 23 %, respectively). The Visual Gain Group × Fc interaction was significant, mouth, $F(2.75, 96.13) = 3.55$, $p = .02$, $\eta_p^2 = .09$; nose, $F(2.81, 98.37) = 5.09$, $p = .003$, $\eta_p^2 = .13$. The main effect of visual gain group at the eyes ROI was not significant ($p = .07$), and neither was the Visual Gain Group × Fc interaction in this region ($p = .15$).

Post hoc analyses were conducted by calculating the respective difference in the time of gaze between adjacent FCs (e.g., [unfiltered stimuli - 29 c/f], [29c/f – 23 c/f]) for the two groups. We then ran five two-sample *t*-tests, in the mouth and the nose ROIs. The only contrast that reached significance after correcting for multiple comparisons (with Holm-Bonferroni) was in the [unfiltered-29 c/f] contrast of the Nose ROI, with participants in the LVG group spending less time on that region than participants in the HVG group.[1]

The relationship between the proportion of correct responses produced across subjects and time spent gazing at the mouth region for each subject was also assessed for each filter condition by using a Pearson product-moment correlation. For words, significant correlations were found in the unfiltered condition, $r(36) = .34$, $p = .04$. Thus, individuals who scored higher in visual gain performance were more likely to fixate on the mouth region. Significant correlations were also found in the 11 c/f, $r(36) = .365$, $p = .03$. The 5 c/f, $r(36) = -.074$, $p = .66$, 17c/f, $r(36) = .29$, $p = .08$, 23 c/f, $r(36) = .31$, $p = .056$, and the 29 c/f, $r(36) = .09$, $p = .58$, were not significant.

Sentences The ANOVA conducted in each ROI revealed no main effect of visual gain group or significant Visual Gain Group × Fc interaction (all *p*s > .1). This suggests that, in the case of long speech materials, the differences between the HVG and the LVG groups do not seem to be explained by differences in visual strategies. None of the correlations between the proportion of correct responses produced across subjects and time spent gazing at the mouth region were significant (all *p*s > .3).

HVG and LVG groups' speechreading ability and gaze behavior On average, when only visual information was available (words, $N = 33$; sentences, $N = 31$), participants spent more time gazing at the mouth region than any other region (71 % of the time in words and 61 % in sentences). In order to compare HVG group's and LVG group's eye behavior in visual-only conditions, we measured the time HVG (words, $N = 18$; sentences, $N = 12$) and LVG participants (words, $N = 15$; sentences, $N = 19$) spent gazing at each ROI. Between-groups *t*-tests (corrected with Holm-Bonferroni) were run for each ROI. These did not reveal any significant differences between the HVG and the LVG groups in any of the ROIs, either for words or sentences (all *p*s > .18, except for fixations to the eye in sentences, $p = .01$, which was not significant when corrected for multiple comparisons). In general, HVG and LVG participants fixated the mouth the majority of the time.

[1] Three other contrasts reached significance before correcting for multiple comparisons. In the nose ROI, the contrast [11c/f - 5c/f] was also significant ($p = .02$) with participants in the HVG group gazing on this region less than participants in the LVG group. In the mouth ROIs, both the contrasts [11 c/f – 5 c/f] and [unfiltered - 29 c/f] were significant ($p = .015$ and $p = .02$, respectively), with participants in the HVG group gazing more at this region at the lowest visual resolution and the LVG group gazing more at this region at the highest resolution.

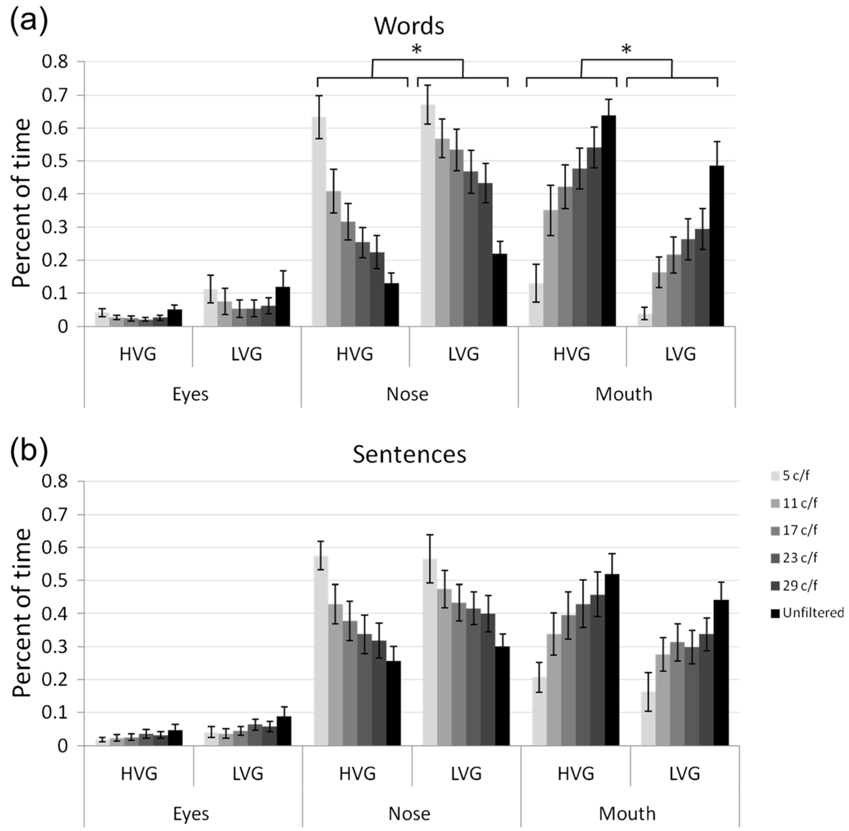

Fig. 7 Percentage of time (±*SE*) spent with eye gaze directed to each region of interest (ROI), as a function of skill level, for words and sentences

Discussion

The main result of the present study is that the participant population was not homogeneous with respect to the benefit from high-resolution components of images. The existence of a distinct high-performing group of participants that could gain more from high-frequency information was assessed with the gap statistic method, which identified the optimal number of clusters as two for both the word and the sentence data. Using this method, we found observers with HVG benefit more from high-frequency spatial information (i.e., good image resolution) than LVG participants in audiovisual conditions where the auditory signal is degraded. While the HVG group generally outperformed the LVG group across Fc, their relative performance increased dramatically when high-frequency information of the talker's face was available (i.e., unfiltered condition). The relative gain obtained by the addition of high-frequency spatial information was 42 % of the overall visual gain in words and 27 % in sentences (see Fig. 5). This suggests that fine facial detail is critical for HVG perceivers to achieve optimal performance. The LVG group, on the other hand, plateaued at lower cutoffs (~23 c/f), suggesting that the information provided by the high spatial frequency spectrum did not confer any additional improvement in performance.

Although the benefit obtained by the addition of higher frequency spatial information was more prominent for words than for sentences, a clear peak in performance in the unfiltered condition for the HVG group is observed across both sets of materials. This suggests that the use of high spatial frequencies by the HVG perceivers is not specific to conditions in which contextual information is more limited (i.e., words). Finally, the HVG group spent more time gazing at the mouth region, but only when presented with shorter stimuli (i.e., words).

The novel finding that higher visual resolution leads to a larger visual gain seems to be initially at odds with previous studies, which have suggested that a rather crude visual signal is sufficient to equate performance to clear image (Munhall et al., 2004; although see Dickinson & Taylor, 2011; Wilson et al., 2016). However, our findings suggest that the provision of higher frequency spatial information benefits a specific subset of the overall population; it is possible that, in previous studies, collapsing across all participants obscured the effects found here. It should be noted, however, that even when averaging our data across all participants, performance in the unfiltered condition was significantly higher compared to remaining filter cutoffs. The different outcomes between our results and previous studies could be related to the use of different talkers; it could be the case that the talkers used in previous studies articulated more clearly than the one used here, thus allowing the extraction of visual cues at a lower visual resolution.

Participants with HVG benefit more from high image resolution: Visual-only versus visual gain

The finding that HVG perceivers require high visual resolution in order to achieve optimal performance is consistent with Wilson et al. (2016), who showed a similar pattern of results between proficient and poor speechreaders using *silent* CVC materials. Our results, however, extend Wilson et al.'s findings—that HVG perceivers make use of fine facial detail in order to make subtle distinctions in visual speech—to more naturalistic conditions (i.e., when residual auditory information is available, as well as with complex linguistic materials, such as sentences). This result is important because it shows that good speechreaders use higher frequency information of the image not only when the visual information is rather limited but also when auditory information could potentially provide complementary cues to those found in the high-resolution range of the image. The similar pattern of results found in Wilson et al. and here might suggest that speechreading and audiovisual speech perception (i.e., visual gain) share the same processes of visual analysis.

It is important to point out however, that the origin of the visual gain for the different participants here is unknown. Higher scores in visual gain could stem from a more efficient ability to understand speech using the visual information alone (i.e., better speechreading skills) and/or a more efficient ability to combine multisensory integration (i.e., better integration skills). The fact that we found a (modest) positive correlation between speechreading ability and visual gain suggests that those participants with higher visual gains are also better speechreaders. These two possibilities cannot be conclusively disentangled without a visual-only baseline for each filter cutoff, which our design did not include. Note that speechreading words and sentences under filtered conditions would likely have been difficult (floor effects), and would have possibly discouraged participants from engaging in the experimental task.

Visual processing mechanisms supporting AV SPIN ability

The finding that participants' visual gain for sentences correlated with their visual gain for isolated monosyllabic words, in which the contextual information is limited to constraints from the lexicon, suggests that part of the variance in visual gain is accounted by individual differences in the ability to extract visual cues from the signal. What is not clear, however, is the specific mechanism that enables participants with HVG to benefit from the higher spatial frequency information. The finding that the HVG group spent more time gazing at the mouth region when presented with words might suggest that they use a more effective gaze strategy, by fixating on regions of the face offering greater benefit from the high definition provided by high spatial frequencies (i.e. mouth region). Note, however, that the HVG group spent more time in this region overall, independent of the quality of the visual image. Trials across filter cutoffs were presented in random order here, and thus this result could stem from a strategy adopted by participants in the HVG group in order to optimize the extraction of visual information in ever-changing contexts. Importantly, no differences in terms of the time spent on the mouth region were observed between the HVG and the LVG groups in longer stimuli materials (i.e., sentences), even though we observed a boost in performance in the unfiltered condition in the HVG group. This suggests that the peak in performance observed in the unfiltered condition is not explained only by a difference in gaze behavior between the HVG and the LVG groups when the resolution is high.

Another possibility is that participants with HVG have an underlying processing ability that enables them to extract more information from the visual image, especially when the resolution is high. Because of their high degree of visual confusability, previous research has identified clusters of phonemes that cannot be visually distinguished from each other. For example, the phonemes /p/, /b/, and /m/ (bilabial group) are articulated at the same place (lips) and have been claimed to appear the same visually (Auer & Bernstein, 1997; Massaro, 1998; Summerfield, 1987). However, very subtle differences in jaw, lip, tongue, and cheek movements of phonemes usually grouped under the same viseme cluster can be detected when examining high-speed video recordings (around 120 frames/second) of a talking face in slow motion (Abel, Barbosa, Black, Mayer, & Vatikiotis-Bateson, 2011). For example, some participants can pick up the differences in optical flow that accompany a plosive bilabial /p/ ('path') and a non-plosive /m/ ('math'; see Jiang et al., 2007). Bernstein (2012) presented deaf and normal-hearing adults with spoken pairs of words that differed only in phonemes belonging to the same viseme clusters and found that, even when the visemes were highly confusable, all listeners could reliably identify above chance which of the spoken words corresponded to an orthographic target word (i.e., 65 %–80 % of correct identification in normal-hearing participants and 80 %–100 % in deaf participants). This suggests that, despite the reduction in phonetic information afforded by visible speech stimuli, perceivers are able to pick up on fine visual phonetic cues present in the visual image in closed sets contexts. It is likely that the fast, subtle microactions that allow the perceptual distinction between phonemes belonging to the same visemic cluster can be fully observed only in the high-frequency range of the image (i.e., good definition). This conclusion is consistent with the effects found here. It is also likely that some of the subtle features that participants with HVG use to distinguish between phonemes become progressively available as the visual resolution increases. Indeed, by visually inspecting the data of the HVG group, one might observe discrete peaks in performance at 11c/f and 23 c/f, both for words and in sentences. These peaks could be related to the availability of specific facial features (e.g., tongue, teeth) as the resolution of the image improves, which again would only be picked up by participants with HVG.

The question still arises as to why such articulatory cues of high detail are not apparent when presented at a natural rate and in open responses sets for some people but can indeed be picked up by participants with HVG. From the perspective of the current visual neural processing models in speech perception (Bernstein & Liebenthal, 2014), the interindividual differences in visual gain found here could be occurring at high levels in the visual pathway, where the system determines whether the crossmodal sensory information belong to the same external event (i.e., detects correspondences) and/or where the optical signals are mapped into its corresponding auditory speech representations. However, the mechanisms responsible for such interindividual differences could be found at lower levels of visual processing. Previous literature suggests that observers that benefit the most from the visual speech information may have superior processing speed (Auer & Berstein, 1997; Feld & Sommers, 2009; Gagné et al., 2011) or superior sensitivity to visual movement (Armstrong, Neville, Hillyard, & Mitchell, 2002; Mohammed et al., 2005). Mohammed et al. showed that the detection of coherent visible motion in random-dot kinematogram displays was related to speechreading skill in the best speechreaders (i.e., deaf participants). Visual speech relies critically on information carried in its dynamic properties. It is possible, thus, that HVG could retrieve subtle motion information from the speaking face that would allow the integration mechanisms to pick up the time-varying features that are common to the acoustical and visual physical channels. More studies are needed to determine what accounts for enhanced phonetic perception in some individuals (see Bernstein et al., 2000).

HVG perceivers generally outperformed LVG perceivers

Even though the HVG group showed a greater relative boost in performance in the unfiltered condition, they were better overall than the LVG group, even at the lowest cutoffs. The fact that the HVG group outperformed the LVG group at even the lowest Fc suggests that they were better able to extract visual speech information in the low spatial frequency range as well, compared to the LVG group. Because the local motion information is severely degraded at the lowest Fc, we believe this group may be better at extracting dynamic configural information from the face (Calvert & Campbell, 2003; Campbell, Zihl, Massaro, Munhall, & Cohen, 1997) that could be used efficiently to integrate audiovisual information. Another possibility is that, as previously suggested with proficient speechreaders (Lidestam et al., 1999), participants with HVG have a superior verbal working memory. If this was the case, however, one would expect to see superior performance overall for the HVG group in sentences compared to words, which was not the case. Another likely possibility is that the HVG group was better even at lowest Fc because they have a superior ability to perceive linguistic wholes on the basis of linguistic fragments (perceptual synthesis; i.e., Lyxell & Rönnberg, 1989; Sanders & Coscarelli, 1970; Sharp, 1972; Watson et al., 1996). However, contrary to Watson et al., we did not find significant correlations between speechreading and auditory-only SPIN for either words or sentences. This challenges the hypothesis that individuals in the HVG group were better overall at perceiving linguistic 'wholes' on the basis of segments. Finally, the higher level of performance at the lowest Fc by the HVG group could be explained by motivation effects. Better performance at higher cutoffs could potentially encourage HVG participants to be more engaged in deciphering speech at lower cutoffs.

Eye gaze and visual resolution

With respect to gaze behavior, the trends observed in eye movements as the resolution increased were similar to that of Wilson et al. (2016): When provided with higher spatial frequency information, participants increasingly gazed at the mouth region, with less time spent gazing at the nose. This is consistent with a strategy where listeners look at the mouth region when there is more information to be gained in doing so. That is, higher resolution results in a greater degree of high acuity information in the mouth region, which cannot be gathered though peripheral vision, whereas sufficient information from the mouth region can be gained through peripheral vision when high spatial frequency information is filtered out.

HVG and LVG group differences in eye gaze

Group differences in eye gaze, however, only arose in word AV SPIN, with the HVG group fixating longer at the mouth (and shorter on the nose) region than the LVG group. The two groups did not show significant differences in eye gaze in sentence AV SPIN, suggesting that they adopted a similar gaze strategy. However, close inspection of the data reveals that, in the unfiltered condition, the LVG group gazed at the mouth a similar percentage of time in words and sentences, whereas the HVG group spent less time on the mouth in the sentence condition. A likely explanation for the decrease of time spent on the mouth by HVG in longer stimuli materials is that they opt to rely more on other types of visual cues beyond those found in the mouth. Besides providing information for the phonemes, visual speech cues have been shown to improve the recognition of prosodic aspects of the message. Prosodic information (i.e., suprasegmental features such as rhythm, stress, and intonation of speech) is also processed to define word, phrase, and sentence boundaries; stress patterns; and syntactic structure (Cutler & Butterfield, 1992; Soto-Faraco, Sebastián-Gallés, & Cutler, 2001). Besides the linguistic structure of the message, suprasegmental features also provide information about certain pragmatic aspects of the conversational situation, such as the emotional state or communicative intent. Such cues extend beyond the mouth region and can be extracted even when the lower part of the face is occluded (Cvejic, Kim, & Davis, 2010; Davis & Kim, 2006). For instance, raising eyebrow movements (Granström, House, & Lunderberg, 1999) or eye widening (Massaro & Beskow, 2002) can serve as an independent prosodic cue to prominence. Indeed, observers preferentially look at these regions in prosody-related judgments (Buchan, Paré, & Munhall, 2007; Lansing & McConkie, 1999).

It is possible, therefore, that participants with HVG use a strategy where they can benefit the most both from the phonemic and from these other segmental cues, in order to parse and interpret the longer utterances optimally. The use of this strategy would be limited to those contexts in which perceivers have access to residual auditory information because the same strategy was not observed in the visual-only screener. That is, our results showed, in line with previous eye-tracking studies (Lansing & McConkie, 2003; Wilson et al., 2016), that the time spent gazing at each ROI was unrelated to speechreading ability.

Limitation of eye gaze behavior in the present study

It is important to mention a clear limitation of the present (and all) studies of eye gaze behavior during speech perception. In natural face-to-face conversations, gaze towards the face serves a number of social purposes. Indeed, in a normal interaction, the listener does not gaze at the mouth of the speaker all the time, because social conventions require eye contact between the two and the observer needs to grasp other cues provided by general body language besides the mere linguistic information (Mirenda, Donnellan, & Yoder, 1983). It is possible, therefore, that the unnatural characteristics of the task might have masked potential differences in terms of gaze strategy between observers with HVG and LVG in the long stimulus materials (i.e., sentences). What is clear, however, is that even in these unnatural contexts, we found large interindividual differences in the amount of visual gain for sentences, but we did not observe differences in eye gaze. Further studies are required to determine whether differences between HVG and LVG observers in eye behavior would emerge in more natural contexts.

Implications

The finding that fine facial detail is critical to some individuals for achieving optimal performance in a SPIN task has several important implications. Whereas a considerable amount of research has demonstrated deleterious effects of degraded auditory input on speech comprehension, our research, along with that of Wilson et al. (2016), highlights the importance of the clarity of the visual signal in speech understanding. Our finding that speech comprehension can suffer from even the mildest forms of visual degradation—for at least a subset of the population–indicates that speech understanding may be negatively impacted when the resolution of the visual signal is less than perfect, as it often is in the case of video-conferencing systems or online streaming.

It is presently unclear why some subjects demonstrate higher visual gain (or speechreading skill) than others. Replication of these results, utilizing measurements of working memory and processing speed, for example, may provide a better understanding of the sources of individual variability in the ability to benefit from the addition of the visual signal. This research may hold important rehabilitative potential; for example, training software dedicated to improving speechreading or audiovisual speech perception skills might focus on helping users to identify the subtle differences that occur in the high-frequency range of the image, enabling users to more quickly recognize visemes or audiovisual speech tokens.

It would also be valuable to extend the study to populations with age-related or congenital hearing loss, where participants may rely more heavily on the visual signal. It is possible that subjects with hearing loss (particularly congenital) may be more attuned to high resolution visual information, perhaps more so than the normal hearing population tested here. Such findings would highlight the need for stronger attention to the clarity of the visual signal in populations with hearing loss.

Acknowledgments We would like to thank Celina Wong for her support in the coding of the videos and in scoring participant responses, and Paul Plante for the development of the eye-tracking analyses program. We would also like to thank the reviewers for their insightful comments. AA is supported by CIHR and NSERC grants.

References

Abel, J., Barbosa, A.V., Black, A., Mayer, C., Vatikiotis-Bateson, E. (2011). The labial viseme reconsidered: Evidence from production and perception. In Y. Laprie & I. Steiner (Eds.), *9th International Seminar on Speech Production* (pp. 337–344). Montreal: PQ.

Armstrong, B. A., Neville, H. J., Hillyard, S. A., & Mitchell, T. V. (2002). Auditory deprivation affects processing of motion, but not color. *Cognitive Brain Research, 14,* 422–434.

Auer, E. T., & Bernstein, L. E. (1997). Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *Journal of the Acoustical Society of America, 102*(6), 3704–3710.

Autodesk Maya [Computer software] (2008). Retrieved from http://www.autodesk.com

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

Berger, K. W. (1972). Visemes and homophenous words. *Teacher of the Deaf, 70,* 396–399.

Bernstein, L. E. (2012). Visual speech perception. In E. Vatikiotis-Bateson, G. Bailly, & P. Perrier (Eds.), *Audiovisual speech processing* (pp. 21–39). Cambridge, England: Cambridge University Press.

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception and Psychophysics, 62*(2), 233–252.

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in Neuroscience, 1*(8), 386.

Binnie, C., Montgomery, A. A., & Jackson, P. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research, 17,* 619–630.

Bradarić-Jončić, S. (1998). The structure of lip-reading different linguistic stimuli in prelingually deaf children. *Hrvatska revija za rehabilitacijska istraživanja, 34*(2), 147–155.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

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

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.

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.

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

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory & Language, 31,* 218–236.

Cvejic, E., Kim, J., & Davis, C. (2010). Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication, 52*(6), 555–564.

Davis, C., & Kim, J. (2006). Audio-visual speech perception off the top of the head. *Cognition, 100,* B21–B31.

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

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

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

Foxe, J.J., Molholm, S., Del Bene, V.A., Frey, H.., Russo, N.N., Blanco, D., . . . Ross, L.A. (2013). Severe multisensory speech integration deficits in high-functioning school-aged children with autism spectrum disorder (ASD) and their resolution during early adolescence. *Cerebral Cortex, 25*(2), 298–312.

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

Granström, B., House, D., Lundeberg, M. (1999). Prosodic cues in multimodal speech perception. *Proceedings of the International Congress of Phonetic Sciences (ICPhS99)* (pp. 655–658). San Francisco, CA: University of California.

Grant, K. W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective. *Journal of the Acoustical Society of America, 112,* 30–33.

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

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America, 103,* 2677–2690.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics, 28*(1), 100–108.

Jackson, P. L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. In C. Lee De Filippo & D. G. Sims (Eds.), *New reflections on speechreading* (pp. 99–115). Washington: The Volta Review.

Jiang, J., Auer, E. T., Jr., Alwan, A., Keating, P. A., & Bernstein, L. E. (2007). Similarity structure in visual speech perception and optical phonetic signals. *Perception and Psychophysics, 69,* 1070–1083.

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

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

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

Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics, 65*(4), 536–552.

Lidestam, B., Lyxell, B., & Andersson, G. (1999). Speech-reading: Cognitive predictors and displayed emotion. *Scandinavian Audiology, 28,* 211–217.

Lyxell, B., & Holmberg, I. (2000). Visual speechreading and cognitive performance in hearing-impaired and normal hearing children (11–14 years). *British Journal of Educational Psychology, 70*(4), 505–518.

Ma, W.J., Zhou, X., Ross, L.A., Foxe, J.J., Parra, L.C. (2009). Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space. *PLOS ONE, 4*(3), e4638.

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

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

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

Massaro, D. W., & Beskow, J. (2002). Multimodal speech perception: A paradigm for speech science. In B. Granström, D. House, & I. Karlsson (Eds.), *Multimodality in language and speech systems* (pp. 45–71). Dordrecht: Kluwer Academic.

Mattys, S. L., Bernstein, L. E., & Auer, E. T., Jr. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perception & Psychophysics, 64*(4), 667–679.

Middelweerd, M. J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *Journal of the Acoustical Society of America, 82*(6), 2145–2147.

Mirenda, P. L., Donnellan, A. M., & Yoder, D. E. (1983). Gaze behavior: A new look at an old problem. *Journal of Autism and Developmental Disorders, 13,* 397–409.

Mohammed, T., Campbell, R., MacSweeney, M., Milne, E., Hansen, P., & Coleman, M. (2005). Speechreading skill and visual movement sensitivity are related in deaf speechreaders. *Perception, 34*(2), 205–216.

Munhall, K. G., Kroos, C., Jozan, G., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perception and Psychophysics, 66*(4), 574–583.

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

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

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,* 55-567.

Payton, K. L., Uchanski, R. M., & Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *Journal of the Acoustical Society of America, 95,* 1581–1592.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research, 28,* 96–103.

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science, 17*(6), 405–409.

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

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex, 17*(5), 1147–1153.

Rouger, J., Fraysse, B., Deguine, O., & Barone, P. (2008). McGurk effects in cochlear-implanted deaf subjects. *Brain Research, 1188,* 87–99.

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., & Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proceedings of the National Academy of Sciences of the United States of America, 104*(17), 7295–7300.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics, 20,* 53–65.

Sanders, J. W., & Coscarelli, J. E. (1970). The relationship of visual synthesis skill to lipreading. *American Annals of the Deaf, 115,* 23–26.

Sharp, E. Y. (1972). The relationship of visual closure to speechreading. *Exceptional Children, 38,* 729–734.

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

Smith, E. G., & Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. *Journal Child Psychology Psychiatry, 48*(8), 813–821.

Sommers, M., Tye-Murray, N., & Spehar, B. (2005). Audiovisual integration and aging. *Ear Hearing, 26,* 263–275.

Soto-Faraco, S., Sebastián-Gallés, N., & Cutler, A. (2001). Segmental and suprasegmental mismatch in lexical access. *Journal of Memory and Language, 45,* 412–432.

Strand, J. F., & Sommers, M. S. (2011). Sizing up the competition: Quantifying the influence of the mental lexicon on auditory and visual spoken word recognition. *Journal of the Acoustical Society of America, 130,* 1663–1672.

Studdert-Kennedy, M. (1989). Feature fitting: A comment on K. N. Stevens' "On the quantal nature of speech.". *Journal of Phonetics, 17,* 135–144.

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: The psychology of lipreading* (pp. 3–51). Hillsdale: Erlbaum.

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

Thorn, F., & Thorn, S. (1989). Speechreading with reduced vision: A problem of aging. *Journal of the Optical Society of America, 6,* 491–499.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistic Society: Series B, 63*(2), 411–423.

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., Hale, S., & Rose, N. S. (2008). Auditory-visual discourse comprehension by older and young adults in favorable and unfavorable conditions. *International Journal of Audiology, 47*(Suppl. 2), S31–S37.

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

Wild, C. J., Yusuf, A., Wilson, D., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful listening: The processing of degraded speech depends critically on attention. *Journal of Neuroscience, 32,* 14010–14021.

Wilson, A., Alsius, A., Paré, M., Munhall, K. (2016). Spatial frequency requirements and gaze strategy in visual-only and audiovisual speech perception. *Journal of Speech, Language, and Hearing Research* (in press).