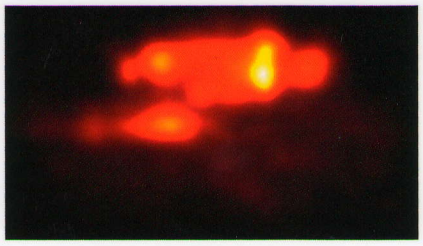
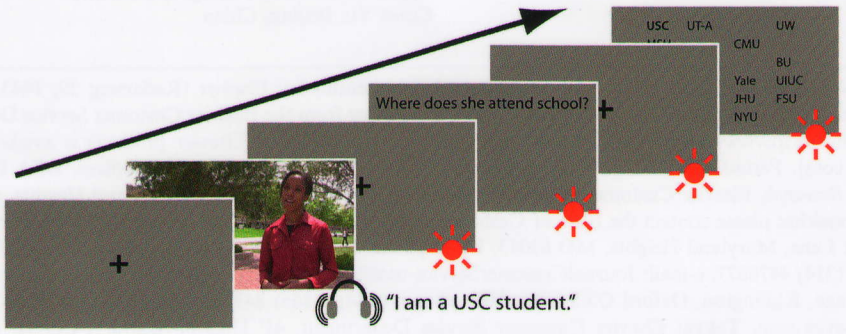
ELSEVIER

VISION RESEARCH

An International Journal for Functional Aspects of Vision





Biochemistry & Cell Biology • Molecular Biology & Genetics
Anatomy, Physiology, Pathology & Pharmacology • Optics, Accommodation & Refractive Error
Circuitry & Pathways • Psychophysics • Perception • Attention & Cognition
Computational Vision • Eye Movements & Visuomotor Control

ISSN 0042-6989 | Volume 65 | 15 July 2012

Vision Research 65 (2012) 62–76

Contents lists available at SciVerse ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Top-down influences on visual attention during listening are modulated by observer sex

John Shen [a,*], Laurent Itti [a,b]

[a] Neuroscience Graduate Program, University of Southern California, 3641 Watt Way, Los Angeles, CA 90089, United States
[b] Department of Computer Science, University of Southern California, 3641 Watt Way, Los Angeles, CA 90089, United States

ARTICLE INFO

Article history:
Received 26 July 2011
Received in revised form 15 May 2012
Available online 20 June 2012

Keywords:
Visual attention
Sex differences
Faces
Saliency
Eye tracking
Task relevance

ABSTRACT

In conversation, women have a small advantage in decoding non-verbal communication compared to men. In light of these findings, we sought to determine whether sex differences also existed in visual attention during a related listening task, and if so, if the differences existed among attention to high-level aspects of the scene or to conspicuous visual features. Using eye-tracking and computational techniques, we present direct evidence that men and women orient attention differently during conversational listening. We tracked the eyes of 15 men and 19 women who watched and listened to 84 clips featuring 12 different speakers in various outdoor settings. At the fixation following each saccadic eye movement, we analyzed the type of object that was fixated. Men gazed more often at the mouth and women at the eyes of the speaker. Women more often exhibited "distracted" saccades directed away from the speaker and towards a background scene element. Examining the multi-scale center-surround variation in low-level visual features (static: color, intensity, orientation, and dynamic: motion energy), we found that men consistently selected regions which expressed more variation in dynamic features, which can be attributed to a male preference for motion and a female preference for areas that may contain nonverbal information about the speaker. In sum, significant differences were observed, which we speculate arise from different integration strategies of visual cues in selecting the final target of attention. Our findings have implications for studies of sex in nonverbal communication, as well as for more predictive models of visual attention.

Published by Elsevier Ltd.

1. Introduction

"Look at me when I'm talking to you!" is a familiar complaint issued by women to their husbands, and a possible illustration that men and women pay attention in different ways. As listeners, women maintain a small but consistent advantage over men at interpreting nonverbal cues during communication (Hall, 1978; McClure, 2000). Women are also popularly believed to be stronger non-verbal communicators than men, not only in interpreting non-verbal communication but also in maintaining gaze at the other person (Briton & Hall, 1995). Since eye gaze is also predominantly steered by visual attention, especially in task contexts (Land & Hayhoe, 2001), eye movements serve both as social indicators of attention and as locations of interest for detailed visual processing. Therefore, sex differences in nonverbal communication ability may be affected by visual attentional selection mechanisms.

In conversation, early studies showed that men make less eye contact than women when listening (Argyle, Alkema, & Gilmour, 1971; Frances, 1979); however, measures of eye-contact were made by third-party observation and not eye-tracking. Other studies suggest that listeners alternate between fixating the eyes and the mouth in stereotypical patterns (Vatikiotis-Bateson et al., 1998). In addition, male observers fixate longer at the nose compared to female observers when faces are displayed in an emotion discrimination task (Vassallo, Cooper, & Douglas, 2009). Since previous eye-movement research on sex differences has predominantly focused on passive viewing of static images, and because sex differences in communication are already known, we chose to study sex differences through a listening task with long clips of interviews to elicit these differences.

At first glance, any sex differences in gaze during listening seem likely to be induced by the nature of the listening task. For example, in Vassallo, Cooper, and Douglas (2009), women attended more to visual cues that were task-relevant. Corresponding with findings in nonverbal communication, sex differences in gaze led to faster response times for women, although they did not lead to significant differences in task accuracy. On the other hand, conspicuous visual stimuli may also steer attention in different tasks (Itti & Koch, 2001; Itti, Koch, & Niebur, 1998; Parkhurst, Law, & Niebur, 2002). Several studies have assessed the contribution of saliency to attention to facial expressions or other stimuli containing humans as minimal

* Corresponding author.
E-mail addresses: shenjohn@usc.edu (J. Shen), itti@pollux.usc.edu (L. Itti).

0042-6989/$ - see front matter Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.visres.2012.06.001

(Birmingham, Bischof, & Kingstone, 2009; Fletcher-Watson et al., 2009; Neumann et al., 2006). However, to the best of our knowledge, no study has examined the differential engagement of men and women to stimulus conspicuity in these tasks. If the stimulus-driven attentional component in men is different than in women during a task involving one-way communication, then this may contribute to the sex differences in sensitivity to nonverbal cues.

For this reason, visual attention should be studied both from a task-centric and stimulus-centric perspective in more natural social environments. Over several different tasks, task-dependent motivations interact with both target feature knowledge and the visual contents of the scene to guide shifts of attention (Kollmorgen et al., 2010). Vincent, Baddeley, and Correani (2009) suggested that stimulus conspicuity plays a minimal role in directing attention using mixture modeling. However, using similar techniques, Mital et al. (2010) have noted that coherent motion, which is a conspicuous cue, tends to correlate with cognitively relevant objects, suggesting a confound between objects and motion. Therefore, we studied how fixations correlate to bottom-up features, as defined by the Itti video saliency model (Itti, 2004), as well as to task-relevant entities such as the face or eyes. We did this in order to fully describe fixation patterns during listening, especially those that distinguish male from female behavior.

Our goal is to quantitatively capture how men and women attend to dynamic scenes in a task relevant to everyday conversation. In the present study, participants watch pre-recorded first-person interviews and answer questions based on the interview content, while we retrieve both object-level information and feature-level statistics at their point of gaze. Admittedly, our task is not equivalent to face-to-face communication, as real-life people elicit a different pattern of eye movements compared to recordings of people (Laidlaw et al., 2011). However, our task is similar to conversation because we ask participants to actively listen to the interviews and give them a followup question after each trial.

Based on the aforementioned observations about conversation (Argyle, Alkema, & Gilmour, 1971; Frances, 1979), we hypothesize that women, more than men, engage in eye contact, and that more generally women fixate areas of the speaker that reveal possible nonverbal signals. We also hypothesize, based on shared experience, that men and women have differing patterns of gazing away from the face of the speaker, a behavior that could be called "distracted" away from the speaker. Furthermore, by examining static and dynamic feature statistics, we measure the degree to which men and women might weigh bottom-up features differently in attentional selection. In other words, the gaze of men and women should be more easily captured by different stimuli, which should be reflected in the predictive performance of saliency feature models. These results will clarify the nature of different gaze behavior between men and women and also illustrate the interplay between task-driven and stimulus-driven modes of attention.

2. Methods

All experimental procedures were approved by the University of Southern California Institutional Review Board (IRB).

2.1. Stimuli

Twelve volunteers, six male and six female, were filmed and interviewed at various locations in Los Angeles (Fig. 1A). Locations were selected such that the scene background would contain interesting objects, i.e., pedestrians, vehicles, etc. which might attract attention. The video camera (Canon HF-11) was set immobile and each speaker was placed in the foreground of the shot, about 1.5–2 m away. The interviews were then cut into 84 clips (MPEG-4, 1080p, non-interlaced) of varying duration (6–60 s, mean 24 s). In each clip, the volunteer spoke in answer to a question posed by the camera operator. Instead of positioning each speaker at the center of the shot, the speaker was filmed at various locations in the shot to guard against explicit center bias due to photography (Tatler, 2007; Tseng et al., 2009). This is reflected in the uncentered distribution of fixation endpoints over all clips (Fig. 1B). All speakers gave verbal consent for the recorded materials to be used.

In the experiment, a question was presented following each clip, testing the recall of either the spoken content of the clip or a visual detail about the speaker (Fig. 1C). Sixteen possible answers were then presented at pre-chosen random locations in a 4×6 grid (choice array), with one answer always being "I don't know".

2.2. Participants

Thirty-four participants (age 18–53, 15 male, 19 female), all distinct from the 12 volunteers, took part in the main experiment. All participants were compensated and naïve to the purpose of the experiment, and gave written consent in compliance with IRB regulations. Participants were told to "listen to and watch the video" and informed that they would have to answer a question about the clip's contents immediately after each clip, by searching for and fixating the correct answer within the choice array. Ten participants underwent only one of the two one-hour sessions.

2.3. Experimental procedures

Stimuli were displayed on a 107-cm color monitor 98 cm in front of the participant, corresponding to a $54.8^\circ \times 32.7^\circ$ field of view. The audio track of the stimuli was played over headphones. Participants' heads were stabilized with a chinrest. Clips were played in random order with the constraint that the clips from each interview were kept in sequence, i.e., an interview clip would appear only after clips which preceded it in the interview had been presented. Participants were allowed to rest after every 15 clips, which comprised one block. Three blocks comprised one of two 1-h sessions, and blocks of clips were evenly divided such that only six speakers, three male and three female, were seen in each block.

Data were acquired with an infrared video-based eye tracker (ISCAN RK-464). The pupil and corneal reflections of the right eye were tracked at a nominal frequency of 240 Hz. Where runs were tracked at an actual confirmed rate slower than 240 Hz, all calculations were adjusted for the slower sample rate: this occurs in less than 2% of the data. A nine-point calibration was performed at the start of each session, and a fixation cross for drift correction was presented before each video clip, before each question, and before each choice array. Calibration data were used to fit a thin-plate-spline transformation in the same manner as (Itti, 2004), with a target mean calibration error of 0.5° and no more than 1° for a given calibration point. The fixation cross that appeared at the start of every trial was used to verify calibration. Data were discarded where initial calibration was off by more than 1°; this occurred for about 5% of the data.

Participants were told that eye-tracking took place only during the presentation of the question testing recall and the choice array of possible answers, but was turned off during the presentation of the video clip. While the question and answer arrays were presented, a red LED turned on below the eye-tracker. Participants were told that the LED indicated that the eye-tracker was recording their eye-movements. This measure of deception was taken to attempt to minimize any possible self-monitoring; participants may have altered their gaze behavior if they knew that they were being recorded while watching the video clips. At the end of the

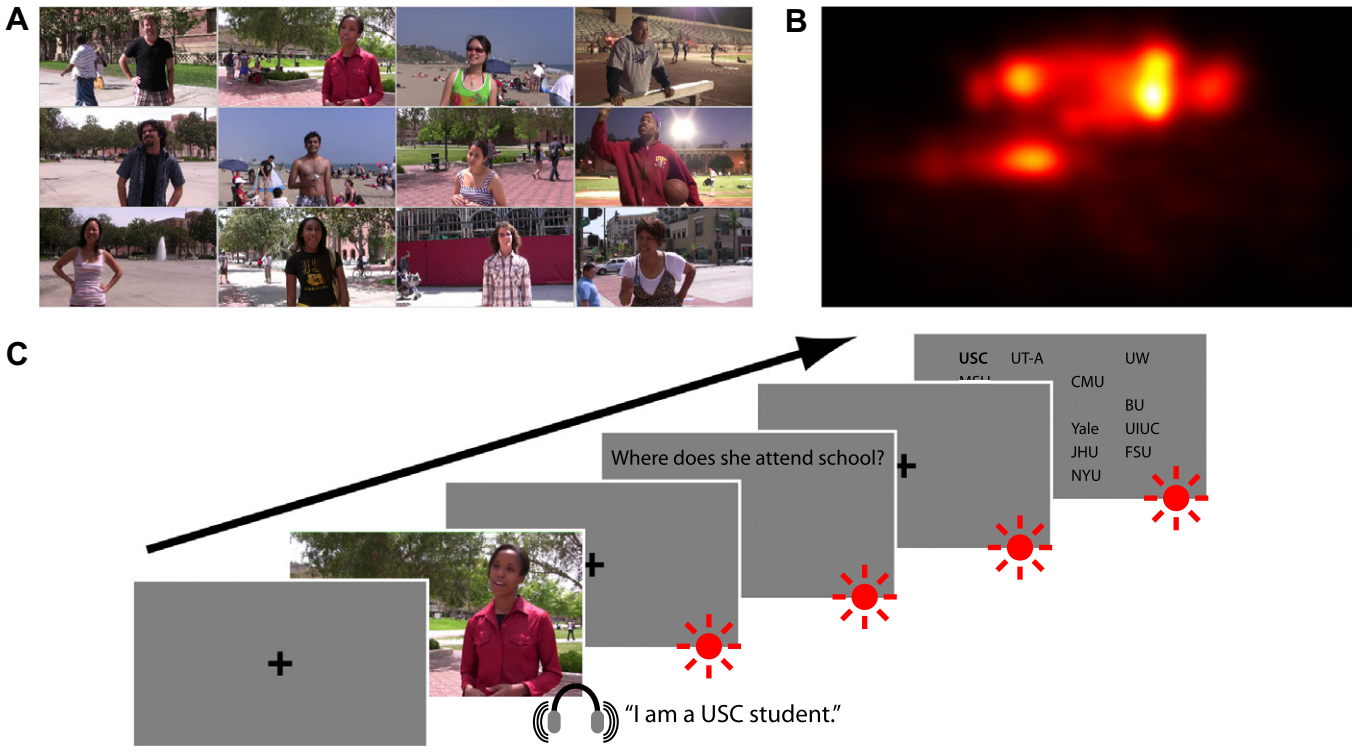


Fig. 1. (A) Stimuli were clips of pre-recorded interviews of speakers. (B) A fixation heat map showing the distribution of fixations across all videos. Note that peaks correspond to locations of the face in the videos, not the center of the screen. (C) Following each interview, a question was presented in order to validate the listening task. Participants were told that eye-tracking only occurred during presentation of a red LED seated on the eye-tracker as a deceptive measure, in order to remove awareness of eye-tracking as a possible confound during video watching. Fixation crosses were interposed between viewing periods to verify calibration. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

experiment, each participant was given full disclosure of the deception component, and was given the option to withdraw their collected data, in accordance with IRB guidelines. No participant exercised this option.

During the presentation of the answer array, participants were asked to press the space bar when holding fixation at the correct choice; in addition to verifying the listening behavior, this procedure was also used to motivate the eye-tracking apparatus for the participants, since they had been told that the eye-tracker was turned off during the video clips. The success rate of male ($M = 52.7\% \pm 7.01\%$) and female ($M = 53.2\% \pm 5.14\%$) participants did not differ significantly (t-test: $t_{32} = -0.17$, $p > 0.05$), and was well above chance ($p = 1/15 \approx 6.67\%$) for all participants.

3. Data analysis

3.1. Marking saccades

In-house MATLAB software was used to parse the eye position on the screen after calibration; the algorithm is described in Berg et al. (2009). Calibrated data were smoothed at 63 Hz with a Butterworth filter in order to attenuate any possible line noise. Afterwards, saccade regions were marked as the minimal eye-trace sub-sequences satisfying the following criteria. First, the ratio of minimum variance to maximum variance in a principal component analysis (PCA) of the samples in 2D coordinates (known as the *condition*), r_{PCA} had to be less than 0.1. Second, the peak velocity on the screen must be greater than 13 deg/s, and the minimum amplitude had to be 0.7°. Regions with greater r_{PCA} or lesser amplitude were designated as fixations, and regions with lesser peak velocity were designated as smooth pursuit.

3.2. Region labeling

To obtain an object-level representation of the stimuli, LabelMe online video software was used to manually annotate regions of interest (ROI) in every clip (Yuen et al., 2009) (Fig. 2A)[1]. The face, eyes, and body were labeled in all clips. All moving objects and persons in the background were also labeled, as well as other objects that the authors thought might be fixated, and all major segments of the background frame, such as trees, road, grass, and sky. Objects were allowed to overlap in the annotation. Less than 1% of fixations landed in between background segments due to limitations in the labeling; these were put into a background category.

Because the eye tracker calibration has a mean accuracy of 0.5°, ROIs were loosely fit to their objects with a correspondingly wide margin (which was approximately 15–20 pixels on screen). In many videos, ROIs moved to track interviewees and other actors in the scene. Moving targets in the background were annotated as rectangles in order to save labor and aid in interpolation. Objects in the foreground were annotated as more general polygons. Annotations were saved as XML files.

We collected saccade targets and subsequent fixations over all eyetraces recorded while watching videos and for each participant (Fig. 3). For each fixation, we recorded the regions of interest (ROIs) at the target location, e.g. mouth, body, eyes, etc. These fixations are marked as circles in Fig. 3A. In other words, a region was considered fixated if the the saccade endpoint was within the ROI at the time of the saccade initiation. Afterwards, these ROIs were categorized into a small set of regions, or ROI types, as follows:

[1] For interpretation of color in Figs. 1–9, the reader is referred to the web version of this article.

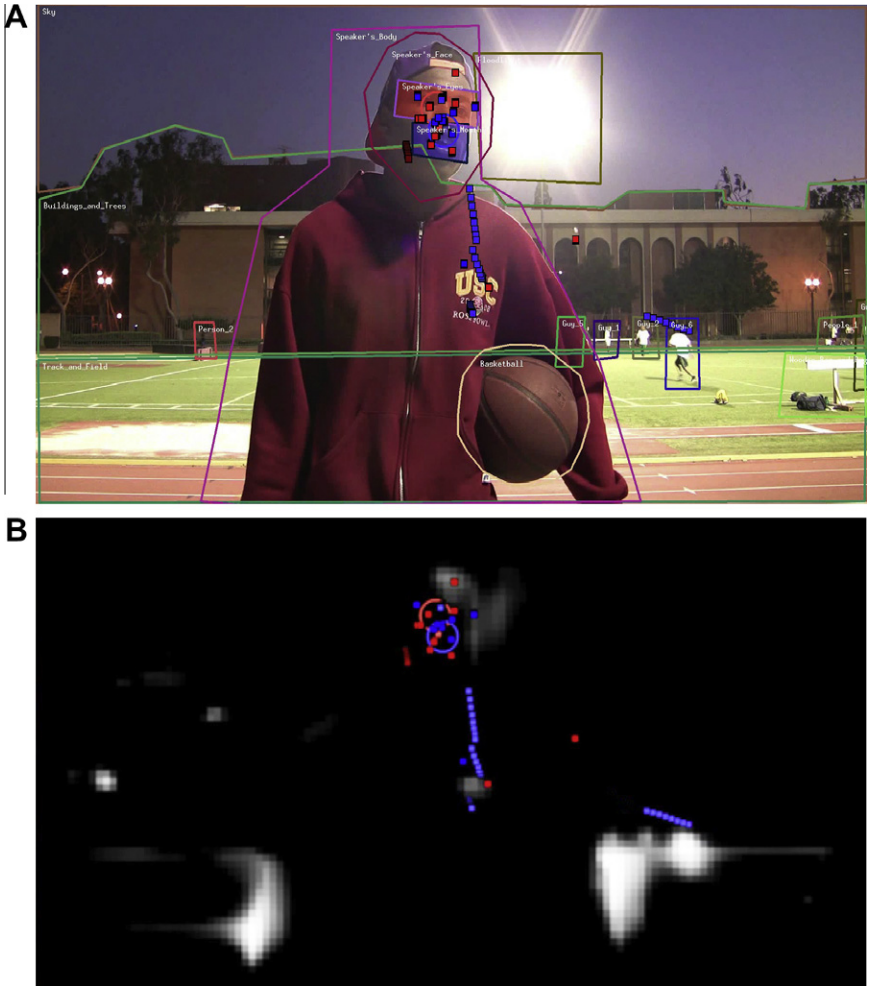


Fig. 2. (A) Calibrated eye-traces are shown super-imposed on a sample frame of the clips. Red traces represent female participants and blue traces represent male participants. For every clip, moving region-of-interest (ROI) outlines around important areas (face, eyes, mouth, body, other people in the scene, and background regions) were drawn, using LabelMe Video online software (Yuen et al., 2009). Highlighted ROIs are regions that are being targeted for a saccade within the display time of this frame. A 1° circular window is drawn around each upcoming saccade target in the appropriate color. (B) Feature maps from both static cues (i.e. color, intensity, and orientation variance) and dynamic cues (i.e. motion energy) were calculated by frame for all clips (SI Movie 1). The combined feature map is drawn here. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

- *face:* the face ROI, consisting of
  - *mouth:* the mouth ROI
  - *eyes:* the eyes ROI
  - *mask:* the face ROI, excluding the mouth and eyes ROIs
- *scene:* the total video excluded by the face ROI, consisting of
  - *body:* the region outlined by the body ROI, excluding the face ROI
  - *people:* the regions outlined by any person ROIs, excluding that of the main speaker
  - *background:* everything else, including any objects but excluding the ROIs of any person including the speaker

In the case that ROIs overlapped, the region that was fixated was determined by the depth ordering of the ROIs, which is listed in the above list of ROI types from front to back. This order was consistent because no objects or people moved between the camera and the speaker.

For the answer arrays, ROIs were defined as equally sized adjacent rectangles, centered on the 4×6 grid on which the choices were superimposed. An answer was labeled as correct if the last point of eye gaze was inside the box containing the correct answer.

3.3. Group fixation maps

Fixation maps were computed with the method described in Blais et al. (2008). Individual fixation maps were generated and registered for each participant for each set of clips belonging to a given interview. Fixations inside the face ROI were selected, Gaussian smoothed at a FWHM of 1°, and re-centered in a parallel coordinate frame, where the origin was always the midpoint between the center of mass of the mouth ROI and that of the eyes ROI for that frame. These formed individual participant fixation maps. Group fixation maps were then calculated for each participant sex and each of the 12 speakers by averaging across participants and z-normalizing. The difference map is the renormalized difference of the male and female group fixation maps. We used the *Pixel test* (Chauvin et al., 2005) to determine the appropriate threshold Z_{crit} such that, from a difference map generated from such a smoothed difference field containing random points, the

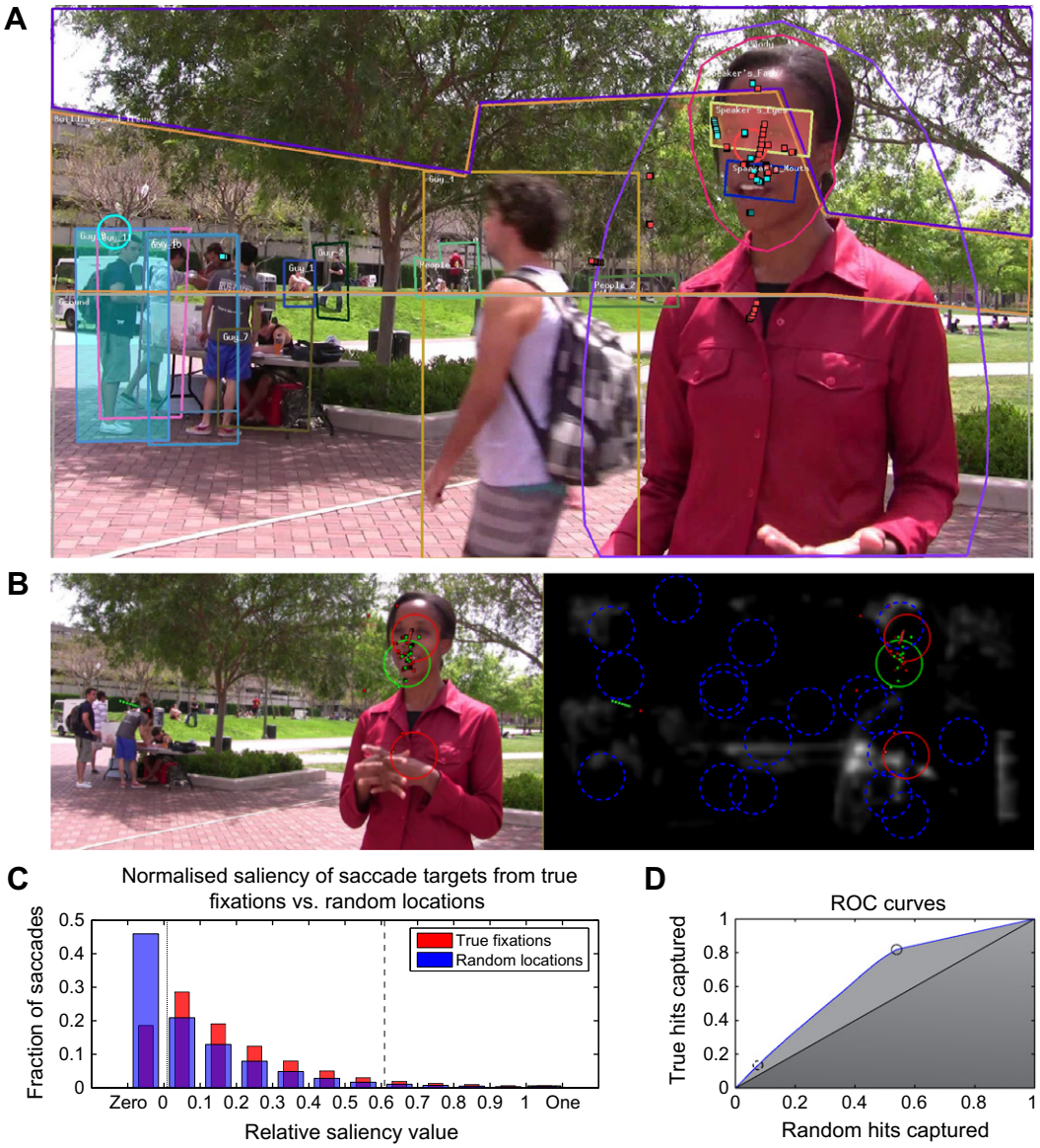


Fig. 3. (A) We examine video frames at the time of saccadic initiation. The fixation following each saccade is represented as a circular window, which we take as the fovea. Another random fixation is taken from the set of saccade targets made by participants watching other videos. (B) The normalized saliency map value is drawn over the maximum of both the true foveated fixation and the random fixations (also foveated). (C) A histogram showing the distribution of saliency values for true and random locations. Brackets near Zero and One are exact 0s and 1s, respectively. Two sample thresholds are shown, the dotted line representing a case capturing a large number of true hits, and the dashed line representing a case rejecting a large proportion of random hits. (D) The reconstructed ordinal dominance curve, with the resultant coordinates of the previous thresholds in corresponding outline. The shaded area under the curve is the AUC score.

probability of finding a pixel intensity greater than Z_{crit} is the Type I error ($\alpha = 0.05$). In our case, $Z_{crit} = 3.37$.

3.4. Feature correlations to the model

We employ a well-used model of visual attention (Itti, 2004; Itti, Koch, & Niebur, 1998) that is inspired from neurophysiology studies of the primate visual system. We use this model in its completely bottom-up form, that is, without providing any contextual guidance or feature biasing. In early visual cortical areas, specific features are calculated in retinotopic maps; these features include intensity, orientation, color, and motion. Instead of encoding a purely local value, such as a pixel value in a graphics file, each location in cortex encodes a local contrast between a small central area and a larger surround. The scale of both the center and surround vary in different cortical areas. Analogously, we calculate the pixel-wise values for each of the aforementioned features at multiple scales, and then subtract filtered features at a small scale with filtered features at larger scales to create scaled center-surround neural feature maps.

Once these maps are generated, we find the locations that are most conspicuous among multiple scales and features by allowing the maps to compete for representation (Itti & Koch, 2001). To do so, the scaled feature maps are first normalized to a common dynamic range, and then nonlinearly combined across scales but within a feature. This non-linear combination mirrors lateral competition for representation in cortical maps. These multi-scale feature maps are again non-linearly combined in similar fashion to create a saliency map. Details of the scaling and combination algorithms can be found in Itti, Koch, and Niebur (1998) and Itti and Koch (2001).

Therefore, for example, in the static condition, the initial scaled center-surround feature maps are calculated for luminance, red–green opponency, blue–yellow opponency, and orientation at

0°, 45°, 90°, and 135°. These maps are first combined within features, but across scales, and then combined a second time across features to create the final map (Fig. 2B).

We use a common measure of correlation related to signal detection theory, the ordinal dominance curve, sometimes referred to as the receiver operating characteristic (ROC), to determine how well feature maps correlate to human fixations (Fig. 3A–C). While the ROC measure is often used to evaluate a model's ability to predict human fixations, we use the model for descriptive purposes; here we presume that the model accurately represents the image's feature expression, and evaluate how well human fixations correlate to those features. The resulting AUC score is calculated by comparing hit rates for normalized saliency map values at fixated locations with those of other randomly selected fixations from other videos (Fig. 3B). In our work, both human fixations and control random fixations are foveated across a 1° radius window to extract the maximum feature value within the window. The feature values for human fixations are compiled into a histogram, as are the feature values for control random fixations (Fig. 3C). Instead of directly calculating a difference measure between histograms, we calculate the ordinal dominance curve by plotting hit rates for human fixations against those of random fixations at all possible thresholds, where fixations with feature values above the threshold constitute hits (Fig. 3D). The area under the ordinal domimnance curve, or AUC score, is the final measure of correlation. In practice, the AUC score is an average of AUC scores over 100 iterations of randomly matched control fixation sets.

All measures of feature correlation, as well as oculomotor and region-based measures of eye movements, were subject to an unbalanced, fixed-effect, 3-factor ANOVA (MATLAB) of participant sex, speaker sex, and saccade target location unless otherwise stated. The saccade target location refers to whether or not the saccade landed on the face or the scene. When post hoc tests are used at each ROI type, the significance level is adjusted with the Šidák–Bonferroni correction for multiple comparisons.

4. Results

4.1. Saccade metrics

Before examining sex differences towards the image regions under gaze, we investigated if any differences in basic eye movement metrics existed (Fig. 4A). Specifically, we measured the average peak velocity, amplitude, and saccade duration for all saccades of each participant (Fig. 4B–D). We found no difference in the medians of average peak velocity between male and female participants (Mann–Whitney U test, $p = 0.47$), amplitude ($p = 0.78$) or saccade duration ($p = 0.32$). We also compared fixation durations for all fixations following a saccade, discarding the first fixation and fixations occurring within or after blinks, and excluding smooth pursuit sections from the fixation (Fig. 4E). A sex difference is significant in the medians for fixation durations ($p < 0.001$), suggesting that fixations are longer for men than for women in the listening task.

4.2. Fixation analysis

To determine whether the inequality in fixation durations was due to longer fixations at a particular region, we measured average fixation durations to both the face and scene after saccades were

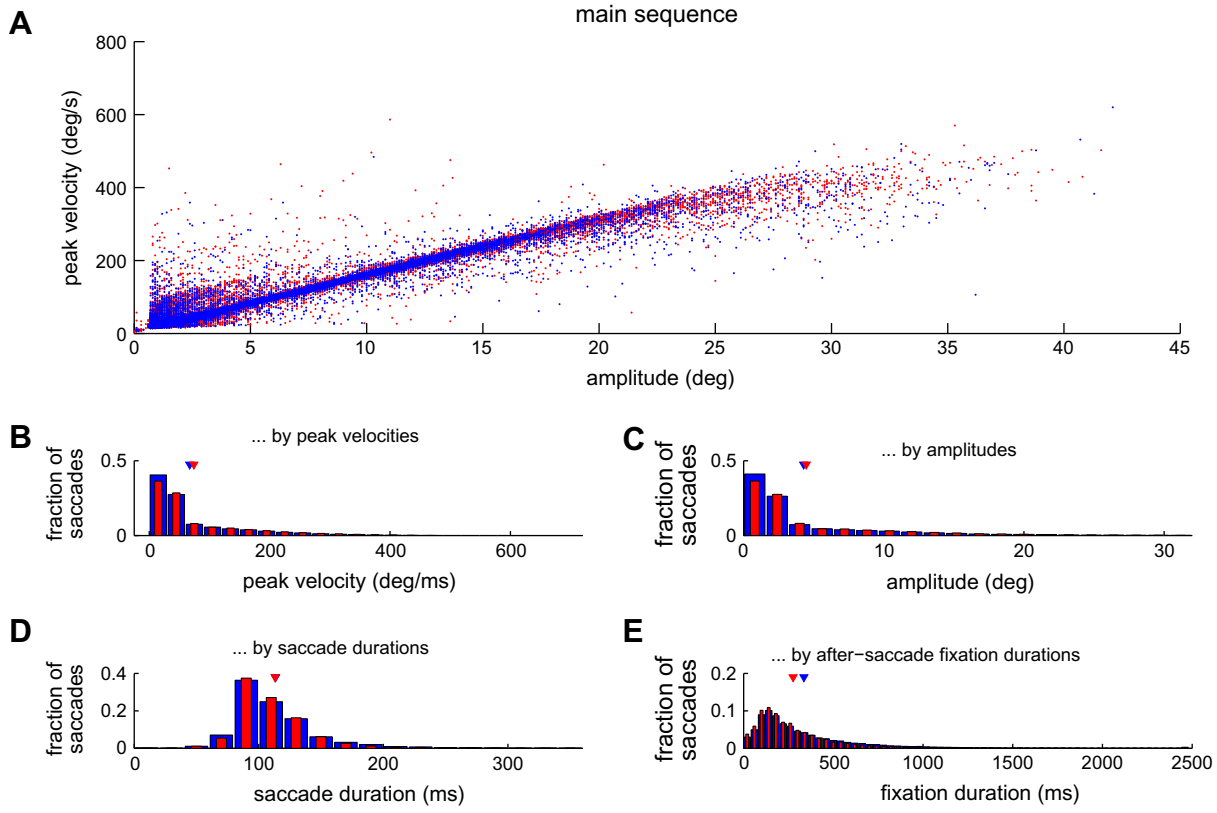


Fig. 4. Examining the main sequence (A), basic metrics of saccades (B-D) and fixations (E) made by men and women. Triangles in (B–E) indicate the median of either male participants (blue) or female participants (red). (A) Main sequence of all saccades made by either men or women. (B) Histogram of the peak velocity of all saccades made by either men or women. (C) Histogram of saccade amplitudes. (D) Histogram of saccade durations. (E) Histogram of the length of fixations made after a saccade. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

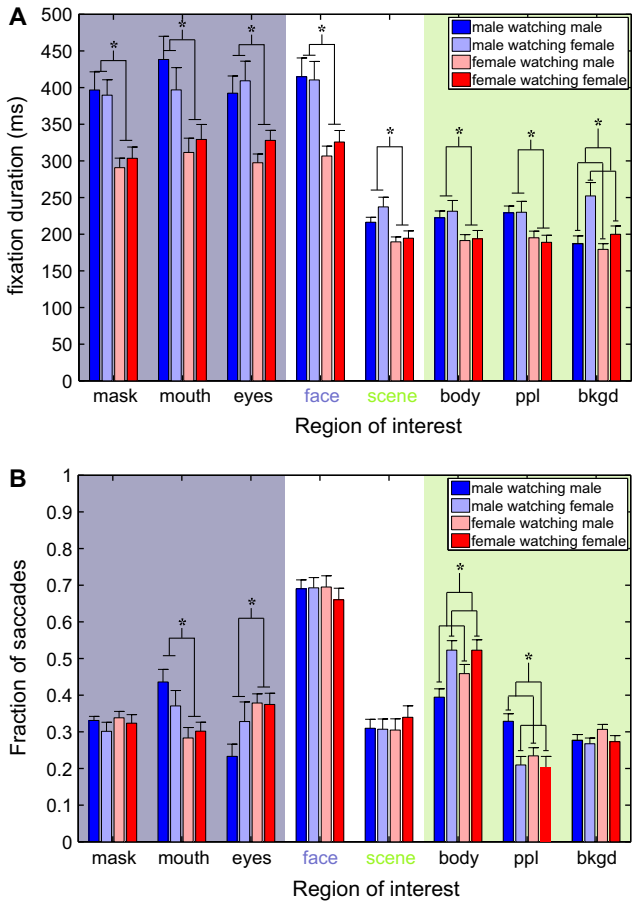


Fig. 5. (A) Average fixation durations at different ROI types for male and female participants. Men look longer than women at all non-background regions. (B) Average estimated likelihood of fixation of different ROI types for male (blue) and female (red) participants. Fixation likelihoods of sub-face (blue background) and sub-scene (green background) regions are calculated conditionally, given the likelihood of the face and scene, respectively. Face and scene bars (white background) represent summary statistics and are not included as ROI types. While listening, men and women look at the eyes and mouth with different frequency. When a man is speaking, the body is fixated less often, and the background more often. Error bars represent between-participant SEMs in all graphs. *: $p < 0.05$, Šidák–Bonferroni corrected. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
All fixation durations.

2 × 2 × 2 ANOVA	$F(1,133)$	p
Main effects		
Listener sex	38	7×10^{-9}*
Speaker sex	0.91	0.34
Face/scene	210	4×10^{-29}*
Interactions		
Listener × speaker	0.03	0.86
Listener × face/scene	8.5	0.004*
Speaker × face/scene	0.024	0.88

* Significant at $p = 0.05$.

made to those regions (Fig. 5A). We employed a 2 (listener sex) by 2 (speaker sex) by 2 (face/scene) between-participant fixed-effect ANOVA with interactions measure any effects on fixation duration. The sex of the participant interacted with whether the participant fixated a scene or face region ($F(1,133) = 8.5$, $p = 0.0043$). There were also strong main effects of both participant sex ($F(1,133) = 38$, $p < 0.0001$) and face/scene ($F(1,133) = 210$, $p < 0.0001$), but no other effects (Table 1). This indicated that men tended to make longer fixations to both the face and the scene, but especially to the face. Also, fixation durations were longer to the face than to the scene, independent of any other factor.

We expanded the analysis by examining fixation durations at each ROI type. To do so, we ran 2 (participant sex) × 2 (speaker sex) between-participant ANOVAs for each ROI type, which included the mask, mouth, eyes, body, people, and background (see Section 2). We used Šidák–Bonferroni correction for multiple comparisions at the $p = 0.05$ confidence level. As listeners, men fixated all regions longer, except the background (Table 2). This confirms that male gaze dwelt longer than female gaze at nearly all times while listening, especially at the face.

In addition to the average fixation duration, the frequency that an ROI is fixated also indicates relative attentional priority. Therefore, for a given ROI type, we calculated its likelihood of fixation for each participant when the participant was watching both males and females (Fig. 5B). The likelihood that sub-regions of the face were fixated is reported conditionally over the likelihood of fixation to the face, and similarly for sub-regions of the scene. This reflects the view that attention to the face is governed, although not wholly driven by, distinct cognitive pathways from those governing other influences on attention (Kanwisher, McDermott, & Chun, 1997; Sergent, Ohta, & MacDonald, 1992).

We performed the same 2 × 2 × 2 between-participant ANOVA as reported in Table 1 on saccade frequency to the face and scene, but found no significant results. Nevertheless, we performed a 2 × 2 ANOVA separately on saccade frequency for each ROI type, and for the face and the scene, correcting for multiple comparisons as before. We found differences in the likelihood that men and women fixated different regions of the face (Table 3). Specifically, males were more likely to fixate the mouth (2-way ANOVA, $F(1,66) = 13$, $p = 0.00069$) and less likely to fixate the eyes ($F(1,66) = 7.7$, $p = 0.0073$). This finding supported our initial hypothesis that men preferentially gazed at the mouth while listening, while women preferentially targeted the eyes. Also, men were not more likely to fixate the face than women were regardless of the sex being viewed, but the face was clearly fixated more often than the scene (Fig. 5B).

We also noted that the likelihood that scene regions were fixated depended upon the sex of the speaker. When the speaker was female, the body of the speaker was more likely to be fixated ($F(1,66) = 14$, $p = 0.00035$), and consequently, the background was less likely to be fixated ($F(1,66) = 9.7$, $p = 0.0028$). However, we found no interactions between participant sex and speaker sex on saccade frequency.

4.3. Listening engagement

We sought to determine whether or not visual gaze behaviors were related to performance on the listening task by computing, for each trial, the fraction of time that participants fixated the face. This percentage we name the *face engagement* (Fig. 6A and B). If face or mouth engagement was lower on incorrect trials, we could infer that face engagement was correlated with listening. We also made an analogous measure of the percent of time that participants looked at the mouth, which we named the *mouth engagement*. For this analysis, we estimate that once a person fixates a region, they are looking at that region until the next fixation. We treated each trial as a separate block, with either a correct or incorrect result.

We analyzed all trials under a 3-way mixed-effect ANOVA with interaction, testing for effects of trial outcome, participant sex, and speaker sex on face engagement. We found that men maintained more face gaze than women (3-way ANOVA, main effect of participant sex, $F(1,1) = 14.68$, $p = 0.0001$), with no other effect reaching

Table 2
Fixation durations at each ROI type.

2 × 2 ANOVA Region	Listener sex		Speaker sex		Interaction	
	$F(1,66)$	p	$F(1,66)$	p	$F(1,66)$	p
Mask	29	1×10^{-6}*	0.026	0.87	0.31	0.58
Mouth	16	0.0002*	0.24	0.63	1.5	0.23
Eyes	23	8×10^{-6}*	1.7	0.20	0.14	0.71
Body	10	0.002*	0.28	0.60	0.096	0.76
People	13	0.0006*	0.065	0.80	0.11	0.74
Background	6.6	0.012	13	0.0005*	3.6	0.063

* Significant at $p \approx 0.0085$, Šidák–Bonferroni corrected for $\alpha = 0.05$.

Table 3
Saccade frequency at each ROI type.

2 × 2 ANOVA Region	Participant sex		Speaker sex		Interaction	
	$F(1,66)$	p	$F(1,66)$	p	$F(1,66)$	p
Face/scene	0.23	0.63	0.31	0.58	0.42	0.52
Mask	0.55	0.42	1.2	0.27	0.14	0.71
Mouth	13	0.0007*	0.57	0.45	1.8	0.18
Eyes	7.7	0.007*	1.7	0.19	2.0	0.16
Body	1.6	0.21	14	0.0004*	1.6	0.21
People	4.3	0.041	9.7	0.003*	3.4	0.069
Background	1.4	0.24	2.0	0.16	0.65	0.042

* Significant at $p \approx 0.0073$, Šidák–Bonferroni corrected for $\alpha = 0.05$.

significance (all $p > 0.1$). When examining the 2-way ANOVA on correct trials only, we found that men maintained more face gaze on correct trials (2-way ANOVA, main effect, $F(1,1) = 16.32$, $p = 0.0001$). This effect did not persist on incorrect trials (2-way ANOVA, main effect, $F(1,1) = 2.24$, $p = 0.13$). Therefore, men looked at the face for a longer time than women did while listening, partially due to their longer fixations. However, it is not necessarily true that men had greater face engagement when listening versus when not listening.

It may be that participants were listening for only a portion of the video, and our result would be confounded by whether or not they had listened when the answer was given. In order to examine this possibility, we measured the face engagement only during a 2-s window where an answer was spoken, also called the *relevant face engagement* (Fig. 6B). Again, we analyzed all relevant trials in a 3-way ANOVA with interaction, testing across factors of trial outcome, participant sex, and speaker sex. We found again that men maintained more face gaze than women (main effect of participant sex, $F(1,1) = 6.51$, $p = 0.011$). No other effect was significant (all $p > 0.05$). Therefore, although visual behavior is still indicative of visual attention, we cannot confirm the hypothesis that looking at the face correlates with better listening.

We performed the same analysis on mouth engagement, defined as the fraction of time spent fixating the mouth over the length of the clip. We tested a 3-way ANOVA with interaction for effects of trial outcome, participant sex, and speaker sex (Fig. 6C). We found a main effect of participant sex, males showing greater mouth engagement ($F(1,1) = 56.2$, $p < 0.0001$), of speaker sex, female speakers eliciting more mouth engagement ($F(1,1) = 10.22$,

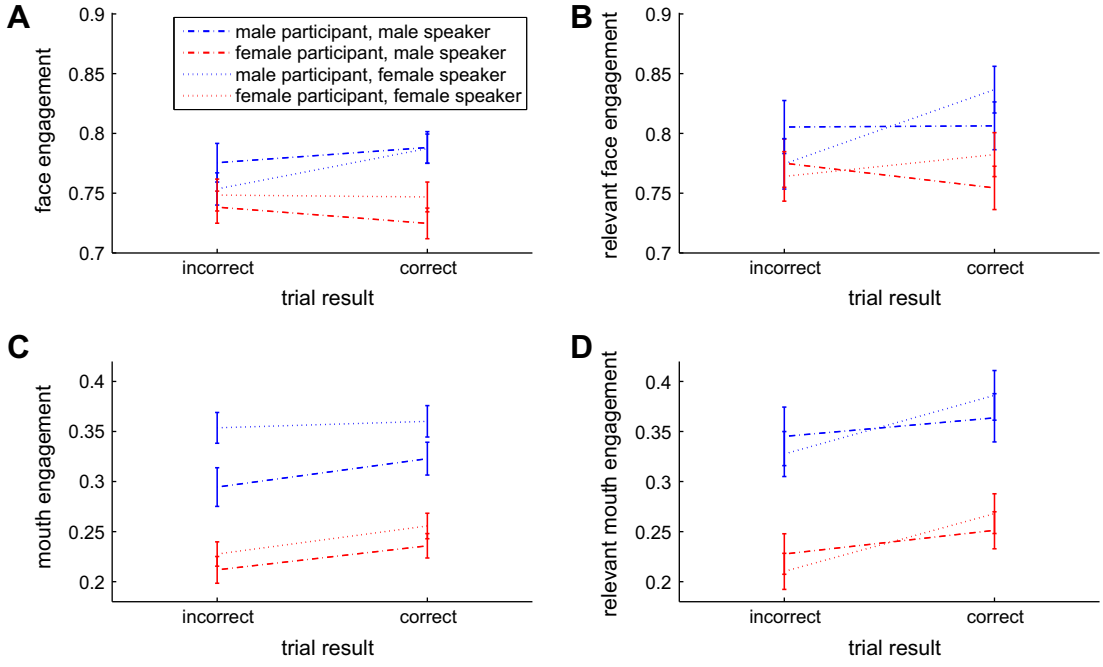


Fig. 6. (A) The fraction of time spent gazing at the face during the whole video for correct and incorrect trials. (B) The fraction of time spent gazing at the face during the 2 s around when a question was answered for correct and incorrect trials. Men look at the face for a greater part of the time than women, but time spent gazing at the face has no outcome on the listening task. (C) The fraction of time spent gazing at the mouth during the whole video for correct or incorrect trials. (D) The fraction of time spent gazing at the mouth during the 2 s around when a question was answered for correct or incorrect trials. Unlike gaze to the face, greater total gaze durations to the mouth improved performance on the listening task.

Table 4
Two-state transition rates of distracted and recovered saccades.

2-sample t-test	Saccade rates (#/min)		t_{32}	p
Saccade type	Male	Female		
Distracted	7.83 ± 0.66	10.22 ± 0.88	−2.218	0.021
Recovered	7.17 ± 0.51	9.47 ± 0.78	−2.406	0.011

$p = 0.001$), and of trial outcome, correct trials showing greater engagement ($F(1,1) = 6.5$, $p = 0.011$). There were no interactions between any of the factors. A similar pattern was found when we examined mouth engagement over only the relevant portion of the clip, at the time the answer was being given (Fig. 6D). We found a main effect of participant gender on mouth engagement ($F(1,1) = 96.1$, $p < 0.0001$) and of incorrect vs. correct trials ($F(1,1) = 4.4$, $p = 0.037$). There were no interactions in the ANOVA. This additionally shows that participants more often gave a correct answer to the listening task when they increased fixation to the mouth, regardless of the sex of the participant.

4.4. Saccade transition analysis

Gaze durations and saccade rates are complementary measures of visual attentional priority. However, we wanted to examine more specifically how men and women also transitioned from observing the speaker to observing the background. It may be that men or women move more frequently from fixating the face to fixating the scene, or vice versa. This would indicate a different viewing strategy that could not be captured by looking at fixations alone. In colloquial terms, we called the transition from the face to the scene a "distracted" saccade, and a transition from the scene to the face a "recovered" saccade.

We measured how many times a minute the average male and female participant made distracted and recovered saccades (Table 4). We ignored saccades that were made either within the face or within the scene. Women made more saccades between the face and the scene compared to men ($t_{32} = -2.27$, $p = 0.015$), and in both directions (distracted, $t_{32} = -2.22$, $p = 0.021$, and recovered, $t_{32} = -2.41$, $p = 0.011$). This may reflect a different approach to listening where men maintain more fixed visual attention and women employ more fluid visual attention.

In addition to transitions between the face and the scene, we reasoned that saccades between subregions of the face and the scene may indicate more specific viewing strategies. Instead of analyzing the full scanpath of each participant, we treated the successive ROI types targeted by each saccade as part of a Markov process, where saccades play the natural role of transitions. Our goal was to see if the estimated transition rates of men and women differed across various ROI types.

For this analysis we included transition rates made from one ROI type to the same ROI type. We then found the average transition rate matrix across male participants (Fig. 7A) and female participants (Fig. 7B). The pairwise differences are reported as t-test statistics for each entry in the transition matrix (Fig. 7C). Significance was determined by a two-tailed t-test, Šidák–Bonferroni corrected for six comparisons across each different saccade source at $\alpha = 0.05$.

We found that men usually made more saccades from the mouth region to the mouth region, which is in line with the main conclusion that men fixate the mouth more often ($t_{32} = -2.962$, $p = 0.017$, Šidák–Bonferroni corrected). More interestingly, women made more saccades between the eyes and the body (to the eyes, $t_{32} = 2.906$, $p = 0.020$, to the body, $t_{32} = 3.089$, $p = 0.012$). Women

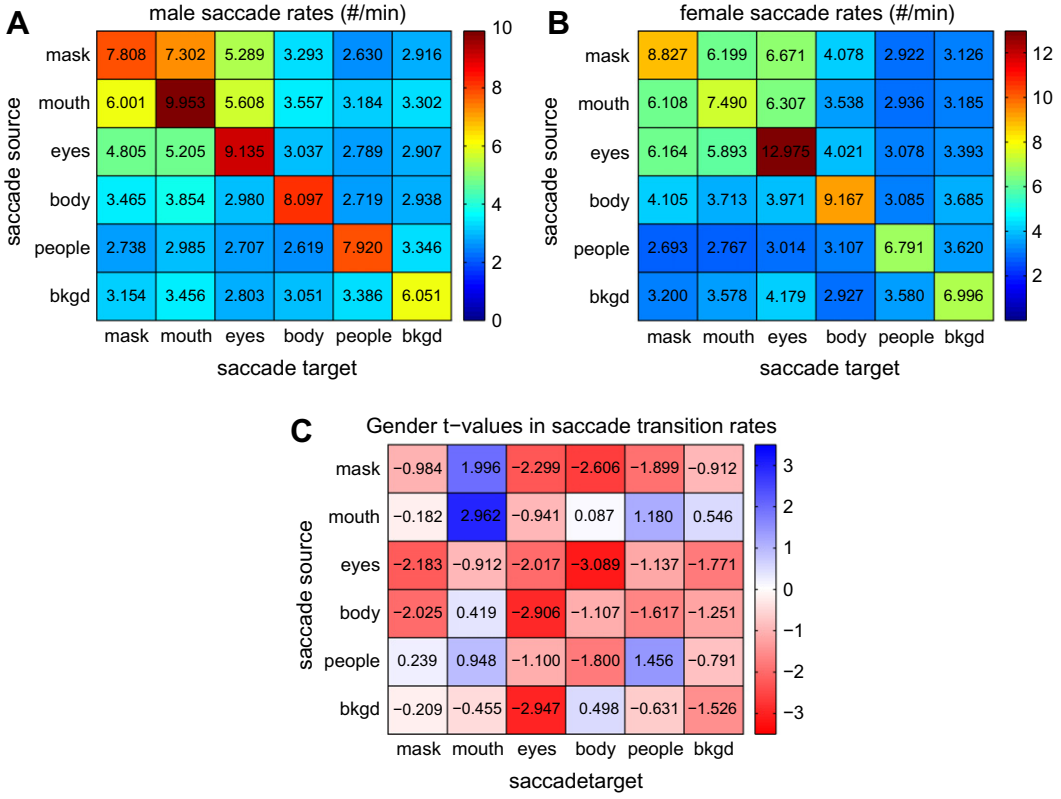


Fig. 7. Saccade transition rates in counts per minute between ROI types, for (A) men and (B) women. (C) The sex difference, as a t-statistic, between transition rates made between men and women. Differences are significant at the 0.05 confidence level when $|t_{32}| \geqslant 1.69$, and significant for multiple comparisons along each transition row when $|t_{32}| \geqslant 2.53$.

also saccaded more to the eyes from the background, (t_{32} = 2.947, p = 0.018). Therefore, we include the body of the speaker as a region of interest that women preferentially fixate over men. Whether these saccades to the eyes or the body follow a more specific scanpath, either to different sub-regions of the face or body, or at different times, remains to be investigated.

4.5. Direct fixation comparisons

While our ROI data revealed significant differences in gaze behavior across the sexes, we also confirmed our hypotheses independently of object-based annotations. To do this, we determined sex differences in attended locations within the face via direct comparison of fixation locations. We prepared group fixation maps to compare under the *Pixel test* for each speaker (Blais et al., 2008; Chauvin et al., 2005) (Section 2). Then we calculated the difference between male and female group maps, re-normalized, and superpose the difference map over a typical frame of the speaker's face (Fig. 8A). In the difference maps, the regions that significantly differ by frequency of fixation (*Pixel test*, $|Z_{crit}| > 3.37$, $p < 0.05$) recover a consistent pattern of male gaze towards the mouth and female gaze towards the eyes. This is most clearly seen when the difference between male and female participant fixations is composed over all stimuli (Fig. 8B).

4.6. Feature-based analysis

While we have shown that men and women look at face sub-regions differently, suggesting that the visual information within those sub-regions influences visual attention, we now ask the question, do men and women fixate feature-level attributes with different specificity? If this is true, then, at the point of gaze, correlations to stimulus features should vary between men and women. We calculated correlation to stimulus features as AUC scores (Section 2, Fig. 3), and then used a 2 × 2 × 2 between-participant ANOVA (participant sex × speaker sex × face/scene ROI) with interaction to determine if male or female participants orient gaze to static features differently. We noted briefly that almost all AUC scores in each category were significantly above chance, which is 0.5 (Fig. 9).

Participant sex did not modulate the AUC score for correlations to static features (Fig. 9A) (main effect, $F(1,133) = 1.7$, $p = 0.193$), but there was an interaction between participant sex and saccades to the face versus the scene (interaction, $F(1,133) = 9.1$, $p = 0.003$) (Table 5). Specifically, saccades to the female face tended to correlate with static features more strongly. In order to examine which differences existed in individual ROIs, we performed a 2 × 2 participant sex × speaker sex ANOVA, post hoc, for each ROI type separately (Table 6). Male gaze correlated to static features more than female gaze did for the body only ($F(1,66) = 7.6$, $p = 0.008$), which contributes to the interaction.

Turning to correlations between AUC score and dynamic features (Fig. 9B), the 2 × 2 × 2 ANOVA revealed a main effect of participant sex (main effect, $F(1,133) = 16.06$, $p = 0.0001$). When looking more specifically at the ANOVAs by region, male gaze correlated to dynamic features more for the body ($F(1,66) = 9.67$, $p = 0.0028$) and weakly for other people ($F(1,66) = 7.19$, $p = 0.0093$). However, there was no interaction between ROI type and participant sex ($F(1,133) = 1.118$, $p = 0.35$). This suggested that motion or its correlates was a greater predictor of male fixations in listening compared to female fixations, regardless of ROI type.

We also note from the three-way ANOVA results on dynamic features that the face was more dynamically salient than the scene

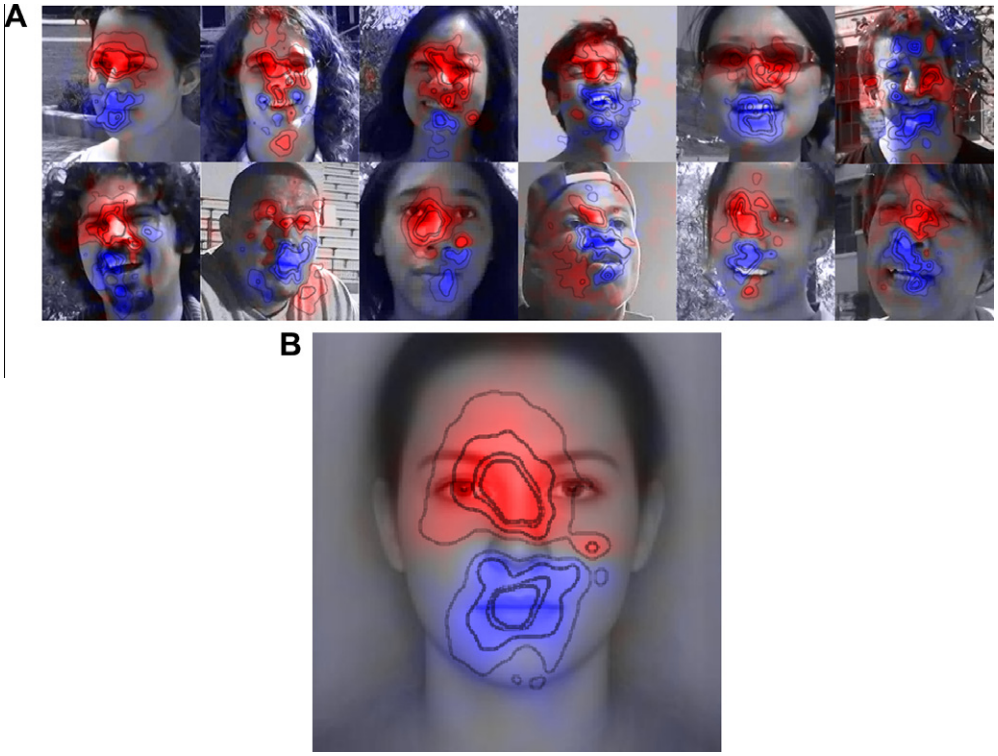


Fig. 8. (A) Fixation difference maps for each speaker. Saccade targets to the face were re-centered relative to the midpoint between the eyes and the mouth, and then aggregated into group fixation maps for male and female participants (see Section 2). Locations that were more frequently fixated by male and female participants are shown in blue and red, respectively. Level contours are drawn where the difference field is equal to ±1, ±2, ±3, or $\pm Z_{crit}$, where Z_{crit} indicates significance at the 0.05 level according to the *Pixel test*. Pixels exceeding Z_{crit} are shown at maximum color saturation. (B) When the fixation difference map is composed across different speakers, the mouth emerges as a male-biased region of fixation, while the eyes emerge as a female-biased region. Figure best seen in color. Average face constructed from faceresearch.org (Tiddeman, Burt, & Perrett, 2001).

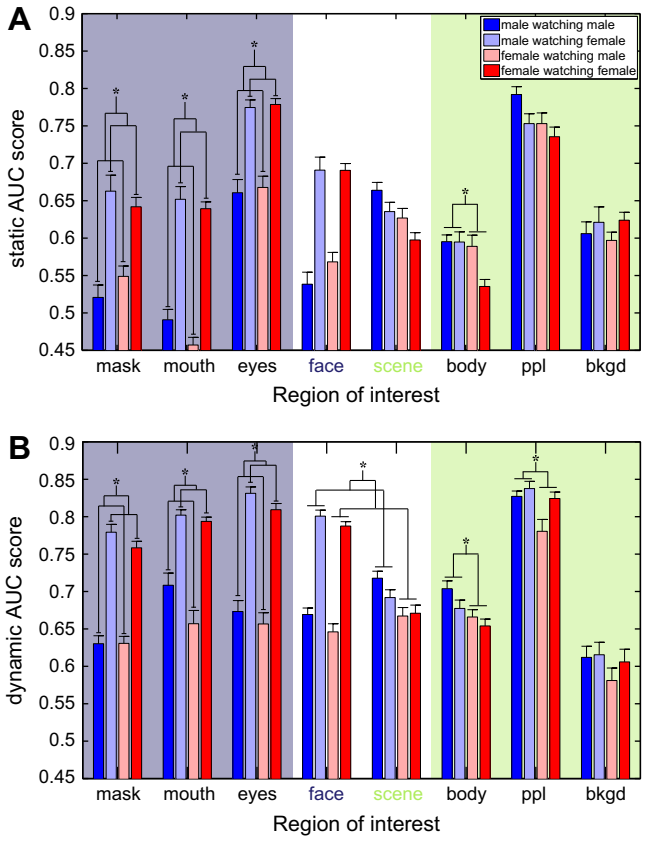


Fig. 9. Average correlations of saccade targets to specific ROIs for male (blue) and female participants (red) ± SEMs for (A) static and (B) dynamic features. Saccades to female faces were more strongly correlated with salient features for female speakers compared to male speakers. Also, male listeners made saccades to both the face and the scene that were more strongly correlated with dynamic features, compared to female listeners. In particular, male saccade targets to the body of the speaker and other people correlate more strongly with motion. *: $p < 0.05$, Šidák–Bonferroni corrected. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Correlations to low-level visual features.

3-way ANOVA	Static AUC score		Dynamic AUC score	
	$F(1,133)$	p	$F(1,133)$	p
Main effects				
Listener sex	1.7	0.19	16	1×10^{-4}*
Speaker sex	39	6×10^{-9}*	87	4×10^{-16}*
Face or scene	1.1	0.36	140	3×10^{-9}*
Interactions				
Listener × speaker	0.8	0.44	2.2	0.15
Listener × face/scene	9.1	0.003*	1.7	0.20
Speaker × face/scene	91	1×10^{-16}*	120	4×10^{-20}*

* Significant at $p = 0.05$.

($F(1,133) = 142.5$, $p < 0.0001$), although this is confounded by a significant interaction with the sex of the speaker ($F(1,133) = 120$, $p < 0.0001$). Specifically, saccades to the female face tended to correlate with dynamic features more strongly, as they also did for static features.

5. Discussion

We have explored the interaction between sex, viewing behavior during a listening task, and the strength of static or dynamic feature expression in a visual scene. Our results demonstrate that, for our one-way communication task, eye movements may be directed differently among men and women. First, women fixate the eyes more when listening, and men fixate the mouth more. This result is confirmed by both an ROI analysis and a direct comparison of fixations. Second, men have longer fixation durations on all regions, and longer fixation times at the mouth facilitates listening performance for both men and women. This illustrates that attention to the visual input is a relevant component of the listening task, and viewers were not passively viewing while listening. Third, women make more saccades between the face and the body of the speaker than men do. Finally, motion energy is a better predictor of male fixations than of female fixations, while static feature saliency is not a better predictor for the fixations of either sex. We conclude that sex differences in fixation patterns to a talking person exist in both the object of fixation and the types of visual features being fixated.

Table 6
ROI-specific correlations to low-level visual features.

2 × 2 ANOVA	Participant sex		Speaker sex		Interaction	
Region	F	p	F	p	F	p
Static feature						
Mask	0.056	0.81	58	1×10^{-10}*	2.5	0.12
Mouth	3.7	0.059	200	1×10^{-21}*	0.78	0.38
Eyes	0.19	0.66	77	1×10^{-12}*	0.013	0.91
Body	7.6	0.008*	5.1	0.027	5	0.030
People	4.9	0.031	4.9	0.031	0.7	0.41
Background	0.05	0.82	2.3	0.14	0.18	0.67
Dynamic feature						
Mask	1.15	0.29	211.9	6×10^{-22}*	1.2	0.27
Mouth	5.48	0.022	81.66	5×10^{-13}*	2.8	0.097
Eyes	2.65	0.11	172.5	8×10^{-13}*	0.05	0.82
Body	9.67	0.003*	3.79	0.056	0.53	0.47
People	7.19	0.009	5.89	0.018	2.2	0.14
Background	1.56	0.22	0.764	0.38	0.43	0.51

* Significant at $p \approx 0.0085$, Šidák corrected for $\alpha = 0.05$.

5.1. Sex differences on viewing a person

We find several region-based sex differences in eye gaze when viewing a talking person. For instance, women made more saccades between the face of the speaker and the surrounding scene, especially the body of the speaker. Because we found no significant increases to feature correlations to female gaze to the eyes or body, but in fact a decrease in dynamic feature correlations to the body, feature conspicuity cannot be a reason for increased saccades to the body. In the wider literature on sex differences, there are very few possible underlying explanations for this. The most plausible connection we found in the literature on sex differences is the finding that women display higher accuracy in decoding nonverbal cues than men (Hall, 1984; Hall & Matsumoto, 2004; McClure, 2000). If women are more able to decode nonverbal cues, then fixating regions outside the mouth would yield more information about the speaker for women than men. In this case, women would assign greater task relevance to the eyes and body of the person during a listening task.

An alternate possibility is that men and women vary on the analytic/holistic spectrum when listening. In this case, women would be considered more analytic as they made more saccades between different regions of the speaker. However, this is inconsistent with the findings made by Hall and Matsumoto (2004) and Vassallo, Cooper, and Douglas (2009) on the sex differences in emotion recognition during picture viewing. Using eye-tracking, Vassallo, Cooper, and Douglas (2009) found that men were found to fixate the

nose more frequently. To explain this they also appealed to an analytic/holistic contrast, but in their work it was men who used a more analytic, unitized approach. Given the evidence, we suggest that women fixate the eyes and body more frequently due to their advantage in nonverbal communication processing, which can be interpreted as part of listening.

We also found that men make fewer transition saccades and have longer fixation durations at all ROIs (Fig. 5). In general, fixation durations have been associated with more extensive processing at fixation (Nuthmann & Henderson, 2010; Rayner, 1998). For example, during a virtual block-stacking task (Droll et al., 2005), fixations were longer when participants detected a change in the object they were currently manipulating. Since longer fixations to the mouth were associated with better listening trial outcomes, men may have adopted a more visual strategy for performing the listening task. This may also contribute to the perception that men appear less attentive during communication (Briton & Hall, 1995); since fixation durations are longer to all ROIs during focused listening, off-task fixations are more noticeable to the speaker.

Finally, we observed various effects of the sex of the speaker on our results. First, saccades were more frequently made to the body of the speaker when the speaker was female, and to other people in the scene when the speaker was male. Furthermore, feature correlations to the face were stronger when participants viewed female faces specifically as opposed to male faces. It would be interesting to see whether these results hold when controlling for the speaker in visual stimuli more specifically.

5.2. Are sex differences more relatable to task relevance or salience?

Turning now to the positive results in the analysis on feature correlations, we found that a motion model can predict visual attention to both the face and the scene. Male attention more reliably correlated with motion compared to female attention among both scene and face saccades (Table 5), and thus men also fixate moving stimuli more frequently. Specifically, men fixate the body of the speaker and of other people more when they are moving (Table 6). This finding suggests that the variation between men and women in attention during listening is not completely explained by an endogenous change in the task-relevance of regions such as the face or body. Even if men and women fixated the same object, such as the body, motion still plays a role in eliciting more attention in men.

Without manipulating the stimuli or task beyond the natural parameters of listening, the question of why men more often fixate motion cannot be completely answered here. To the best of our knowledge, no one has reported a sex-based difference in feature-based correlations to eye movements in any other task domain. For instance, we have analyzed the free-viewing data on movies originally from Carmi and Itti (2006) (available under the CRCNS collaboration) and have not found any sex differences in correlations of gaze to motion features, suggesting that the listening task at hand has a role in sex differences to motion. There are other limitations to our study due to the laboratory setting of viewing. Eye-tracking during real-world outdoor activity, such as walking outdoors, highlights differences in viewing strategy compared to laboratory conditions; people appear to be fixated when on a collision course (Jovancevic-Misic & Hayhoe, 2009) and are less fixated when close by Foulsham, Walker, and Kingstone (2011), neither of which are predicted by motion conspicuity.

Since motion is a ubiquitious bottom up cue that correlates with situational changes, such as the entrance or presence of other objects (Dorr et al., 2010), a difference in exogenous selection due to motion in the scene is only one possibility. Salient features often correlate with objects that are of cognitive interest (Elazary & Itti, 2008; Henderson et al., 2007, chap. 25), and motion is no exception. A complementary possibility is that motion is more often selected because it correlates to objects that men consider more task-relevant. Rather than holding either possibility to be exclusively true, we highlight the need to consider an integrated approach.

The role of motion as an exogenous visual attentional cue is still under debate. It has been acknowledged that motion correlates with eye gaze (Carmi & Itti, 2006; Itti, 2006; Le Meur, Le Callet, & Barba, 2007; Mital et al., 2010). Some authors have objected that stimulus onsets, jump cuts (such as those employed in Hollywood movies) (Dorr et al., 2010) and central fixation bias (Tatler, 2007; Vincent, Baddeley, & Correani, 2009) can artificially induce fixations consistent with attracting motion. In addition, it is possible that longer clips reveal changing top-down attentional control that can idiosyncratically drive gaze, supplanting any possible role of motion (Tatler, Baddeley, & Gilchrist, 2005). We sought to minimize any possible confounds of stimuli composition on the exogenous capture of attention by motion by using long shots of uncentered participants (Fig. 1). In support of exogenous effects of motion on attention, other studies demonstrate that when motion is not a task-relevant cue, motion onset can still cause attentional capture (Franconeri, Simons, & Junge, 2004; Hillstrom & Yantis, 1994), which is the slowing of reaction times to a search task when distractors begin to move. Even when motion would be an irrelevant cue, as in Franconeri, Simons, and Junge (2004), its onset can still elicit attentional capture. Therefore, motion appears to be sufficient in some settings to attract attention, but its effects may be modulated by the nature of the stimuli and the task in other settings.

5.3. Integrated models of visual attention

As an illustration, we can consider the possible hypotheses of pure stimulus-driven and task-relevant models when trying to explaining the sex differences in gaze we observe. The pure *stimulus-driven prediction* states the following: correlations to feature conspicuity are necessary to drive fixation. This means that we should only see an increase in the frequency of fixation of a given ROI when there is a concurrent increase in correlations to feature values in that ROI. *If we see an increase in fixation frequency to a given region without a concurrent increase in the correlation to feature values, we cannot explain that increase in fixation frequency* and our prediction is incomplete. Under this prediction, we would not be able to explain why women more frequently fixate the eyes than do men, as discussed above. Neither static nor dynamic feature correlations of gaze to the eyes are greater for female fixations than for male fixations. This rules out the pure stimulus-driven model.

On the other hand, the pure *task-relevant prediction* of eye gaze states the following: the recognition of a task-relevant object's identity is necessary to drive fixation. We should only see an increase in correlations to features of a given ROI if the ROI commonly exhibits those features and if there is a concurrent increase in fixation frequency to that ROI. *If we see an increase in feature correlations of gaze to an ROI without a concurrent increase in the frequency of fixation, then we cannot explain that increase in feature correlations* and our prediction is incomplete. Under this prediction, we would not be able to explain why correlations of motion saliency to gaze in the scene ROI are significantly larger for men than for women, while the fixation frequency is the same. We note that it does not help us to subdivide the scene ROI because correlations of motion saliency to gaze in the body ROI are also significantly greater for men than for women, while fixation frequency is equal. Thus, we also cannot completely hold the either pure task-relevant model to be true, and consider a third option,

which is that task-relevant attention itself selects directly for motion.

We speculate that motion is more fixated by men because, in the listening task, men have a task-relevant attentional setting during listening which is more attuned to motion. This bias, though not related to the demands of the listening task, may nevertheless arise during this and other tasks. Along these lines, many current models of visual attention incorporate some idea of similarity to target features (Kanan et al., 2009; Navalpakkam & Itti, 2007; Wischnewski et al., 2010; Zelinsky, 2008) and associations to global scene context (Peters & Itti, 2007; Torralba et al., 2006) to enhance gaze prediction. For example, in the SNR model (Navalpakkam & Itti, 2007), a top-down signal may bias or highlight feature streams that provide a greater signal over noise for the particular sought-after target. In the SUN model (Kanan et al., 2009), a Bayesian model biases those features which are predicted to carry the most information about target location. Other proposals have suggested that attention is directed towards proto-objects that exist at an intermediate level of representation (Walther & Koch, 2006) with coherent features that are then selected by task specifications, such as search for small red objects (Bundesen, 1990; Wischnewski et al., 2010). All of these constructions allow modelers to inject task-relevant knowledge into the weighting of what would otherwise be a pure bottom-up set of maps, which subsequently more closely match real-world fixations. However, these models could not identify moving objects as targets for attention without first calculating motion as a low-level feature.

5.4. Attention to faces and their parts

Now we address what relation center-surround, feature-based attention has to fixations to the face. In this study we found that feature correlations of fixations to the face and the eyes were significantly greater than chance, contradicting earlier studies using static images where saliency did not correlate with fixations to faces in free-viewing (Birmingham, Bischof, & Kingstone, 2009; Nyström & Holmqvist, 2008). Specifically, Nyström and Holmqvist (2008) found that lowering contrast of faces reduced saliency but did not affect fixations to the face. Likewise, Birmingham, Bischof, and Kingstone (2009) found that saliency did not correlate with fixations to the eyes, and that the eyes were not more salient than other image locations. In our listening task, correlations to static and dynamic saliency may have elicited increased eye fixation, especially since fixations to the eyes were among the regions displaying the greatest static feature correlation. However, other more task-driven influences, such as attention to non-verbal communication, may also elicit fixation. Along these lines Cerf, Frady, and Koch (2009) showed that the addition of a "feature" channel implementing a dedicated, contrast-invariant face detector (Viola & Jones, 2001) dramatically increased the predictive power of the saliency framework to predicting saccades to images containing faces. Regarding the work of Nyström and Holmqvist (2008), such a contrast-invariant face detector would predict saccades to contrast-reduced faces as well. However, such a model performs more poorly for images without faces because the face detector (Viola & Jones, 2001) is prone to false positives. Our study is distinctive from these studies in that it does not use static images as stimuli, which preclude conspicuity due to motion. This would explain that, while static salience may be an unnecessary factor in saccading faces in accordance with Nyström and Holmqvist (2008) and Birmingham, Bischof, and Kingstone (2009), motion may still be included as an additional influence attracting gaze to faces.

Nevertheless, the literature indicates that the mere presence of a static face is still sufficient for its fixation. Moreover, attention to the face is involuntary in some cases (Bindemann et al., 2005) and fast, occurring within 100–110 ms in forced-choice tasks (Crouzet, Kirchner, & Thorpe, 2010). One promising line of work suggests that the amplitude spectrum can account for such speeded, automatic saccades; even phase-scrambled images of faces are more frequently saccaded than those of houses or other objects (Honey, Kirchner, & VanRullen, 2008). Using ERP evidence, Rossion and Caharel (2011) demonstrated that P1 positivity at 90–100 ms, but not N170 negativity, was stronger for phase-scrambled images of faces over similarly treated images of cars, while both were stronger for intact faces over cars. This suggests that the faster P1 response to faces can be triggered without a corresponding N170 face effect, and may be responsible for the fast orienting to faces. In more general research on visual attention, the amplitude spectrum, as captured in a summary statistic over features such as gist (Peters & Itti, 2008; Torralba et al., 2006), has been shown to have a strong top-down influence in directing gaze.

While the automatic fixation of faces as a whole can be traced to global statistics, it is more difficult to explain for what reasons men more frequently fixate the mouth over the eyes. Specifically, we see from the AUC correlations that the eyes carry more static salience than the mouth, but there is no indication that men are less attracted to static salience (Fig. 9A). On the other hand, the mouth and eyes do not significantly differ in dynamic salience; even if men gaze more often at dynamic stimuli, this in itself does not explain the increased mouth fixation (Fig. 9B). One possibility is that the amplitude spectrum that reveals the location of a face also reveals the specific locations of the mouth and eyes, and that male fixation is biased to mouths in the amplitude spectrum space. However, we also suggest some task-relevant influences that may be present to explain the male bias towards mouth fixation.

One possibility is that men specifically fixated the mouth to improve speech recognition and listening, as we showed mouth engagement correlated with better listening performance. While both men and women took advantage of the motion of the mouth in order to perform the specific listening task, men were no more prone to fixate the mouth than women more often when correctly performing the listening task. Buchan, Pare, and Munhall (2008) also observed that increasing the signal to noise ratio of the audio signal elicited greater mouth fixation in a test of speech perception. A second possibility that has been raised is that men avoid eye gaze because they perceive it as a competitive social cue which hinders communication, while women pursue eye gaze because it facilitates mutual understanding (Swaab & Swaab, 2009). We cannot isolate this hypothesis in our data because we did not use live participants with two way discussion (Bailly, Raidt, & Elisei, 2010).

While we have looked primarily at perceptual or psychological theories so far, more directly biological factors may also play a significant role. Intriguingly, similar fixation biases towards the mouth and away from the eyes are found in high-functioning autistic participants (Klin et al., 2002; Pelphrey et al., 2002; Riby, Doherty-Sneddon, & Bruce, 2009) as well as amygdala-lesioned participants (Spezio et al., 2007). In particular, the duration of eye gaze in autistic participants correlates with amygdala activation (Dalton et al., 2005). A clever fMRI experiment on normal participants carried out by Gamer and Büchel (2009) demonstrates possible causation from amygdala activity. In this experiment, emotional faces were shifted upwards or downwards during a brief initial presentation such that either the mouth or the eyes was located at initial fixation. During presentation of fearful faces, amygdala activation was higher when the mouth was present at fixation rather than the eyes; furthermore, this activation correlated with the viewer's tendency to gaze towards the eyes. This supports the idea that amygdala activation quickly modulates attention in visual cortices, speeding eye movements for analyzing emotion when faces, especially fearful ones, are presented (Adolphs & Spezio, 2006). Therefore, the greater female frequency of eye

movements may be due to differences in lateralized amygdala activation in females (Cahill et al., 2004; Mechelli et al., 2005).

Even though orienting to the parts of faces may be due to limbic, subcortical structures, it is still unclear whether or not they should be classified as exogenous or endogenous, or possibly a hybrid of both. For instance, for autistic viewers, more frequent saccades to the mouth have been attributed to impaired top-down processing (Neumann et al., 2006). However, motion was not taken into account as a feature of interest. On the other hand, for amygdala-lesioned viewers, saccades to the eyes were restored to levels of those of normal controls by using a gaze-contingent paradigm. Because the gaze-contingent manipulation only displayed the region of a face within a Gaussian mask of width 3°, it blocked bottom-up processing of peripheral regions (Kennedy & Adolphs, 2010), suggesting that in normal viewing for amygdala-lesioned viewers, stimulus-driven attention to the eyes is impaired. It would be interesting to use gaze-contingent methods to study the relationship between modes of attention in both saccades to the mouth and distracted saccades exhibited away from the face, to examine whether or not they can be explained solely by stimulus-driven capture.

6. Conclusion

This study examines the interaction between engagement in a conversational task, bottom-up visual features, and individual differences in sex in order to explain observed differences in gaze during conversation. We demonstrate in naturalistic, dynamic settings that, when listening to another person, men are more likely to gaze at the mouth and less likely to gaze at the eyes compared to women. In addition, we find that, while looking to the face, men attend more to motion cues than do women, reflecting a sex-specific feature influence on attention. We interpret this with a mixture of task-relevant and stimulus-driven effects. Our current explanation is that women rely more on features relating to the social nature of the scene to direct attention, whereas men rely more on motion features, possibly indicating that men and women maintain awareness of their social surroundings with different task priorities in mind. We surmise that future models of task-dependent modulation in visual attention can leverage these individual differences to enhance predictive performance. Our research also allows us to understand in a new light how visual design might be tailored to attract attention depending on the sex of its target audience.

Acknowledgments

Thanks to David Berg, Po-he Tseng and Farhan Baluch for helpful discussion. Dan Parks and Sophie Marat read earlier drafts of this paper. We especially thank the editor, Ben Tatler and one anonymous reviewer for many constructive suggestions. Hassan Sarwar outlined the annotations for all the videos. This work was supported by the National Science Foundation (CRCNS Grant No. BCS-0827764), the Army Research Office (W911NF-11-1-0046), and the U.S. Army (W81XWH-10-2-0076).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.visres.2012.06.001.

References

Adolphs, R., & Spezio, M. (2006). Role of the amygdala in processing visual socials stimuli. *Progress in Brain Research, 156*, 363–378.

Argyle, M., Alkema, F., & Gilmour, R. (1971). The communication of friendly and hostile attitudes by verbal and non-verbal signals. *European Journal of Social Psychology, 1*, 385–402.

Bailly, G., Raidt, S., & Elisei, F. (2010). Gaze, conversational agents and face-to-face communication. *Speech Communication, 52*, 598–612.

Berg, D. J., Boehnke, S. E., Marino, R. A., Munoz, D. P., & Itti, L. (2009). Free viewing of dynamic stimuli by humans and monkeys. *Journal of Vision, 9*, 1–15.

Bindemann, M., Burton, M. A., Hooge, I. T. C., Jenkins, E. H., & De Haan, Rob (2005). Faces retain attention. *Psychonomic Bulletin & Review, 12*, 1048–1053.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2009). Get real! resolving the debate about equivalent social stimuli. *Visual Cognition, 17*, 904–924.

Blais, C., Jack, R. E., Scheepers, C., Fiset, D., & Caldara, R. (2008). Culture shapes how we look at faces. *PLoS ONE, 3*, e3022. http://dx.doi.org/10.1371/journal.pone.0003022.

Briton, N. J., & Hall, J. A. (1995). Beliefs about female and male nonverbal communication. *Sex Roles, 32*, 79–90.

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

Bundesen, C. (1990). A theory of visual attention. *Psychological Review, 97*, 523–547.

Cahill, L., Uncapher, M., Kilpatrick, L., Alkire, M. T., & Turner, J. (2004). Sex-related hemispheric lateralization of amygdala function in emotionally influenced memory: An fMRI investigation. *Learning & Memory, 11*, 261–266.

Carmi, R., & Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research, 46*, 4333–4345.

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision, 9*, 1–15.

Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., & Gosselin, F. (2005). Accurate statistical tests for smooth classification images. *Journal of Vision, 5*, 1–8. http://dx.doi.org/10.1167/5.9.1.

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision, 10*, 16.1–17.

Dalton, K. M., Nacewicz, B. M., Johnstone, T., Schaefer, H. S., Gernsbacher, M. A., Goldsmith, H. H., et al. (2005). Gaze fixation and the neural circuitry of face processing in autism. *Nature Neuroscience, 8*, 519–526.

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision, 10*, 28.

Droll, J. A., Hayhoe, M. M., Triesch, J., & Sullivan, B. T. (2005). Task demands control acquisition and storage of visual information. *Journal of Experimental Psychology: Human Perception and Performance, 31*, 1416–1438.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision, 8*, 3.1–15.

Fletcher-Watson, S., Leekam, S., Benson, V., Frank, M., & Findlay, J. (2009). Eye-movements reveal attention to social information in autism spectrum disorder. *Neuropsychologia, 47*, 248–257.

Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research, 51*, 1920–1931.

Frances, S. J. (1979). Sex differences in nonverbal behavior. *Sex Roles, 5*, 519–535.

Franconeri, S. L., Simons, D. J., & Junge, J. A. (2004). Searching for stimulus-driven shifts of attention. *Psychonomic Bulletin & Review, 11*, 876–881.

Gamer, M., & Büchel, C. (2009). Amygdala activation predicts gaze toward fearful eyes. *Journal of Neuroscience, 29*, 9123–9126.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin, 85*, 845–857.

Hall, J. A. (1984). *Nonverbal sex differences: Communication accuracy and expressive style*. Baltimore, Maryland: The Johns Hopkins University Press.

Hall, J. A., & Matsumoto, D. (2004). Gender differences in judgments of multiple emotions from facial expressions. *Emotion, 4*, 201–206.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In Roger P. G. van Gompel, Martin H. Fischer, Wayne S. Murray, & Robin L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537–562). Amsterdam, Netherlands: Elsevier.

Hillstrom, A. P., & Yantis, S. (1994). Visual motion and attentional capture. *Perception & Psychophysics, 55*, 399–411.

Honey, C., Kirchner, H., & VanRullen, R. (2008). Faces in the cloud: Fourier power spectrum biases ultrarapid face detection. *Journal of vision, 8*(9), 1–13.

Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing, 13*, 1304–1318.

Itti, L. (2006). Quantitative modeling of perceptual salience at human eye position. *Visual Cognition, 14*, 959–984.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews. Neuroscience, 2*, 194–203.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 1254–1259.

Jovancevic-Misic, J., & Hayhoe, M. (2009). Adaptive gaze control in natural environments. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 29*, 6234–6238.

Kanan, C., Tong, M. H., Zhang, L., & Cottrell, G. W. (2009). Sun: Top-down saliency using natural statistics. *Visual Cognition, 17*, 979–1003.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*, 4302–4311.

Kennedy, D. P., & Adolphs, R. (2010). Impaired fixation to eyes following amygdala damage arises from abnormal bottom-up attention. *Neuropsychologia, 48*, 3392–3398.

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry, 59*, 809–816.

Kollmorgen, S., Nortmann, N., Schrder, S., & Knig, P. (2010). Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLoS Computational Biology, 6*, e1000791.

Laidlaw, K. E. W., Foulsham, T., Kuhn, G., & Kingstone, A. (2011). Potential social interactions are important to social attention. *Proceedings of the National Academy of Sciences, 108*, 5548–5553.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research, 41*, 3559–3565.

Le Meur, O., Le Callet, P., & Barba, D. (2007). Predicting visual fixations on video based on low-level visual features. *Vision Research, 47*, 2483–2498.

McClure, E. B. (2000). A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents. *Psychological Bulletin, 126*, 424–453.

Mechelli, A., Friston, K. J., Frackowiak, R. S., & Price, C. J. (2005). Structural covariance in the human cortex. *Journal of Neuroscience, 25*, 8303–8310.

Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2010). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation, 3*, 5–24.

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron, 53*, 605–617.

Neumann, D., Spezio, M. L., Piven, J., & Adolphs, R. (2006). Looking you in the mouth: Abnormal gaze in autism resulting from impaired top-down modulation of visual attention. *Social Cognitive and Affective Neuroscience, 1*, 194–202.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision, 10*, 20.

Nyström, M., & Holmqvist, K. (2008). Semantic override of low-level features in image viewing both initially and overall. *Journal of Eye Movement Research, 2*, 1–11.

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

Pelphrey, K. A., Sasson, N. J., Reznick, J. S., Paul, G., Goldman, B. D., & Piven, J. (2002). Visual scanning of faces in autism. *Journal of Autism and Developmental Disorders, 32*, 249–261.

Peters, R.J., & Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *IEEE conference on computer vision and pattern recognition* (pp. 1–8).

Peters, R. J., & Itti, L. (2008). Applying computational tools to predict gaze direction in interactive visual environments. *ACM Transactions on Applied Perception, 5*, 1–19.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372–422.

Riby, D. M., Doherty-Sneddon, G., & Bruce, V. (2009). The eyes or the mouth? Feature salience and unfamiliar face processing in Williams syndrome and autism. *The Quarterly Journal of Experimental Psychology, 62*, 189–203.

Rossion, B., & Caharel, S. (2011). ERP evidence for the speed of face categorization in the human brain: Disentangling the contribution of low-level visual cues from face perception. *Vision Research, 51*, 1297–1311.

Sergent, J., Ohta, S., & MacDonald, B. (1992). Functional neuroanatomy of face and object processing: A positron emission tomography study. *Brain, 115*, 15–36.

Spezio, M. L., Huang, P.-Y. S., Castelli, F., & Adolphs, R. (2007). Amygdala damage impairs eye contact during conversations with real people. *Journal of Neuroscience, 27*, 3994–3997.

Swaab, R. I., & Swaab, D. F. (2009). Sex differences in the effects of visual contact and eye contact in negotiations. *Journal of Experimental Social Psychology, 45*, 129–136.

Tatler, B. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*, 1–17.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45*, 643–659.

Tiddeman, B., Burt, D. M., & Perrett, D. (2001). Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications, 21*, 42–50.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*, 766–786.

Tseng, P.-H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9*, 4.

Vassallo, S., Cooper, S. L., & Douglas, J. M. (2009). Visual scanning in the recognition of facial affect: Is there an observer sex difference? *Journal of Vision, 9*, 11.1–11.10.

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

Vincent, B., Baddeley, R., & Correani, A. (2009). Do we look at lights? Using mixture modelling to distinguish between low-and high-level factors in natural image viewing. *Visual Cognition, 17*, 856–879.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. IEEE Comput. Soc. I-511–I-518.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*, 1395–1407.

Wischnewski, M., Belardinelli, A., Schneider, W., & Steil, J. (2010). Where to look next? Combining static and dynamic proto-objects in a TVA-based model of visual attention. *Cognitive Computation, 2*, 1–18.

Yuen, J., Russell, B., Liu, C., & Torralba, A. (2009). Labelme video: Building a video database with human annotations. In *Proceedings of the 2009 IEEE 12th international conference on computer vision* (pp. 1451–1458).

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review, 115*, 787–835.