How saliency, faces, and sound influence gaze in dynamic social scenes

Antoine Coutrot, Nathalie Guyader

How saliency, faces, and sound influence gaze in dynamic social scenes

Antoine Coutrot

Gipsa-lab, CNRS, & Grenoble-Alpes University, Grenoble, France

Nathalie Guyader

Gipsa-lab, CNRS, & Grenoble-Alpes University, Grenoble, France

Conversation scenes are a typical example in which classical models of visual attention dramatically fail to predict eye positions. Indeed, these models rarely consider faces as particular gaze attractors and never take into account the important auditory information that always accompanies dynamic social scenes. We recorded the eye movements of participants viewing dynamic conversations taking place in various contexts. Conversations were seen either with their original soundtracks or with unrelated soundtracks (unrelated speech and abrupt or continuous natural sounds). First, we analyze how auditory conditions influence the eye movement parameters of participants. Then, we model the probability distribution of eye positions across each video frame with a statistical method (Expectation-Maximization), allowing the relative contribution of different visual features such as static low-level visual saliency (based on luminance contrast), dynamic low-level visual saliency (based on motion amplitude), faces, and center bias to be quantified. Through experimental and modeling results, we show that regardless of the auditory condition, participants look more at faces, and especially at talking faces. Hearing the original soundtrack makes participants follow the speech turn-taking more closely. However, we do not find any difference between the different types of unrelated soundtracks. These eye-tracking results are confirmed by our model that shows that faces, and particularly talking faces, are the features that best explain the gazes recorded, especially in the original soundtrack condition. Low-level saliency is not a relevant feature to explain eye positions made on social scenes, even dynamic ones. Finally, we propose groundwork for an audiovisual saliency model.

Introduction

From the beginning of eye tracking, we know that faces attract gaze and capture visual attention more than any other visual feature (Buswell, 1935; Yarbus, 1967). When present in a scene, faces invariably draw gazes, even if observers are explicitly asked to look at a competing object (Bindemann, Burton, Hooge, Jenkins, & de Haan, 2005; Theeuwes & Van der Stigchel, 2006). Many studies have established that face perception is holistic (Boremanse, Norcia, & Rossion, 2013; Farah, Wilson, Drain, & Tanaka, 1998; Hershler & Hochstein, 2005) and pre-attentive (Bindemann, Burton, Langton, Schweinberger, & Doherty, 2007; Crouzet, Kirchner, & Thorpe, 2010), and the brain structures specifically involved in face perception have been pointed out (Haxby, Hoffman, & Gobbini, 2000; Kanwisher, McDermott, & Chun, 1997). Despite their leading role in attention allocation, faces have rarely been considered in visual attention modeling. Over the past 30 years, numerous computational saliency models have been proposed to predict where gaze lands (see Borji & Itti, 2012, for a taxonomy of 65 models). Most of them are based on Treisman and Gelade's (1980) Feature Integration Theory, stating that low-level features (edges, intensity, color, etc.) are extracted from the visual scene and combined to direct visual attention (Itti, Koch, & Niebur, 1998; Koch & Ullman, 1985; Le Meur, Le Callet, & Barba, 2007; Marat et al., 2009). However, these models cannot be generalized to many experimental contexts, since the dynamic and social nature of visual perception are not taken into account (Tatler, Hayhoe, Land, & Ballard, 2011). Typical examples in which they fail dramatically are visual scenes involving faces (Birmingham & Kingstone, 2009). More recently, visual saliency models combining face detection with classical low-level feature extraction have been developed and have significantly outperformed the classical ones (Cerf, Harel, Einhäuser, & Koch, 2008; Marat, Rahman, Pellerin, Guyader, & Houzet, 2013).

Citation: Coutrot, A., & Guyader, N. (2014). How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of Vision*, *14*(8):5, 1–17, http://www.journalofvision.org/content/14/8/5, doi:10.1167/14.8.5.

doi: 10.1167/14.8.5 Received October 24, 2013; published July 3, 2014 ISSN 1534-7362

Despite these significant efforts focused on attention modeling, auditory attention in general, and audiovisual attention in particular, has been left aside. Visual saliency models do not consider sound, even when dealing with dynamic scenes. When running eye-tracking experiments with videos, authors never mention the soundtracks or explicitly remove them, making participants look at silent movies, which is, of course, not an ecologically valid situation. Indeed, we live in a multimodal world and our attention is constantly guided by the fusion between auditory and visual information. Film directors offer a good illustration of this by using soundtrack to strengthen their hold on the audience. They manipulate the score to modulate the tension and tempo in scenes or to highlight important events in the story (Branigan, 2010; Zeppelzauer, Mitrovic, & Breiteneder, 2011; Chion, 1994). Research confirms that music may, in some cases, exert a significant impact upon the perception, interpretation, and remembering of film information (Boltz, 2004; Cohen, 2005). Not only music, but auditory information in general affects eye movements. In a previous study, we showed that removing the original soundtrack from videos featuring various visual content impacts eye positions, increasing the dispersion between the eye positions of different observers and shortening saccade amplitudes (Coutrot, Guyader, Ionescu, & Caplier, 2012).

Thus, what we hear has an impact on what we see. This is particularly true for speech and faces, which are known to strongly interact, as evidenced by the huge literature on audiovisual speech integration (Bailly, Perrier, & Vatikiotis-Bateson, 2012; Schwartz, Robert-Ribes, & Escudier, 1998; Summerfield, 1987). To investigate audiovisual integration, most of these studies presented talking faces to observers and measured how visual or auditory modifications impacted observers' eye movements or speech comprehension (Bailly, Raidt, & Elisei, 2010; Lansing & McConkie, 2003; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). They identified the eyes and the mouth as two strong gaze attractors during audiovisual speech processing, and showed that the degree to which gaze is directed toward the mouth depends on the difficulty of the speech identification task. Yet, results emanating from experimental set-ups using isolated close-ups of faces might not be generally applied to the real world, where everything is continuously moving and embedded in a complex social and dynamic context. To address this issue, Võ, Smith, Mital, and Henderson (2012) eye-tracked participants watching videos of a pedestrian engaged in an interview. They showed that observers' gazes were dynamically directed to the eyes, the nose, or the mouth of the interviewee, according to events depicted (speech onsets, eye contact with the camera, quick movement of the head). The authors also found that removing the speech signal decreased the number of fixations on the pedestrian's face in favor of the scene background.

Nevertheless, in daily life, conversations are often made of several speakers embedded in a complex scene (objects, background), not only listening to what is being said but interacting dynamically. Thus, Foulsham and colleagues eye-tracked observers viewing video clips of people taking part in a decision-making task. (Foulsham, Cheng, Tracy, Henrich, & Kingstone, 2010). These authors showed that gazes followed the speech turn-taking, especially when the speaker had high social status. These results indicate that during dynamic face viewing, our visual system operates in a functional, information-seeking fashion. A few very recent papers quantified how the turn-taking affects the gaze of a noninvolved viewer of natural conversations (Foulsham & Sanderson, 2013; Hirvenkari et al., 2013). These studies presented conversations to participants with the related speech soundtracks or without any sound. They both showed that sound changed the timing of looks. With the related speech soundtracks, speakers were fixated on more often and more quickly after they took the floor, leading to a greater attentional synchrony.

All the previously reviewed studies reported behavioral and eye movement analyses, but did not quantify the relative contributions of faces (mute or talking) and of classical visual features to guide eye movements. Birmingham and Kingstone (2009) showed static social scenes to observers and compared their eye positions to the corresponding low-level saliency maps (within the meaning of Itti & Koch, 2000). The authors showed that saliency did not predict fixations better than chance. They noticed that classical low-level saliency models do not account for the bias of observers to look at the eyes within static social scenes. But what about dynamic scenes, where motion is known to be highly predictive of fixations, much more than static visual features (Mital, Smith, Hill, & Henderson, 2010)? What are the relative powers of classical visual features to attract gaze? How is their attractiveness modulated by auditory information? In this study, we first quantified temporally how different visual features explain the gaze behavior of noninvolved viewers looking at natural conversations embedded in complex natural scenes. Five classical visual features were compared: the face of the conversation partners, the low-level static saliency, the low-level dynamic saliency, the center area, and chance (a uniform spatial distribution). We chose these features because they are often pointed out by the visual exploration literature. The center area reflects the center bias, i.e., the tendency one has to gaze more often at the center of the image than at the edges (Tseng, Carmi, Cameron, Munoz, & Itti, 2009). Then, we measured the influence of auditory information on these features. Previous studies showed that different types of sounds

interact differently with visual information when viewing videos (Vroomen & Stekelenburg, 2011; Song, Pellerin, & Granjon, 2013). Other studies dealing with static images and lateralized natural sounds showed that eye positions are biased toward sound sources, depending on saliencies of both auditory and visual stimuli (Onat, Libertus, & König, 2007). Like visual saliency, auditory saliency is measured by how much an auditory event stands out from the surrounding scene (Kayser, Petkov, Lippert, & Logothetis, 2005). Thus, one can legitimately hypothesize that different auditory scenes (with different auditory saliency profiles) would have different impacts in the way one listens in on a conversation. For instance, an abrupt auditory event, with local saliency peaks, may not influence gaze in the same way as a continuous auditory stream. We extracted conversation scenes from Hollywood-like movies. We recorded the eye movements of participants watching the movies either with the original speech soundtrack, with an unrelated speech soundtrack, with the noise of moving objects (abrupt onsets, e.g., falling cutlery), or with landscape continuous sound (slowly changing components, e.g., wind blowing). We modeled the different recorded gaze patterns with the expectation-maximization (EM) algorithm, a statistical method widely used in statistics and machine learning, and recently successfully applied to visual attention modeling (Gautier & Le Meur, 2012; Ho-Phuoc, Guyader, & Guerin-Dugue, 2010; Vincent, Baddeley, Correani, Troscianko, & Leonards, 2009). This method is a mixture model approach that uses participants' eye positions to estimate the relative contribution of different potential gaze-guiding features. In the following, we first study the impact of sound on classical (saccade amplitudes, fixation durations, dispersion between eye positions) and less classical (distance between scanpaths) eye movement parameters. Then, thanks to the EM algorithm, we analyze how auditory information modulates the relative predictive power of different visual items (faces, low-level static and dynamic visual saliencies, center bias).

Methods

The experiment described in the following is part of a broader study (Coutrot & Guyader, 2013). The stimuli and the eye-tracking data described below are available at http://www.gipsa-lab.fr/~antoine.coutrot/.

Participants

Seventy-two participants took part in the experiment: 30 women and 42 men, from 20 to 35 years old ($M = 23.5$; $SD = 2.1$). Participants were not aware of the purpose of the experiment and gave their informed consent to participate. This study was approved by the local ethics committee. All were French native speakers, had a normal or corrected-to-normal vision, and reported normal hearing.

Stimuli

The visual material consisted of 15 one-shot conversation scenes extracted from French Hollywood-like movies. Videos featured two to four conversation partners embedded in a natural environment. Videos lasted from 12 to 30 s ($M = 19.6$; $SD = 4.9$), had a resolution of 720×576 pixels^2 (28×22.5 squared degrees of visual angle), and a frame rate of 25 frames per second. We chose stimuli featuring conversation partners embedded in complex scenes (cafe, streets, corridor, office, etc.) involving different moving objects (glasses, spoons, cigarettes, papers, etc.). Faces occupied an area of $3.3 \pm 0.4 \times 5.2 \pm 0.9$ deg^2 and were separated from each other by $10° \pm 2°$. Thus, on average, each face only occupied $(3.3 \times 5.2) / (28 \times 22.5) = 2.7\%$ of the frame area. The auditory material consisted of 45 monophonic soundtracks: a first set of 15 soundtracks extracted from the conversation scenes (dialogues), a second set of 15 soundtracks made up of noises from moving objects (short abrupt onsets, e.g., falling cutlery), and a third set of 15 soundtracks extracted from landscape scenes (continuous auditory stream, e.g., wind blowing).

To investigate the effect of auditory information on gaze allocation during a conversation, we created four auditory versions of the same visual scene, each one of them corresponding to an auditory condition. The Original version in which visual scenes were accompanied by their original soundtracks, the Unrelated Speech version in which the original soundtrack was replaced by another speech soundtrack from the first set, the Abrupt Sounds version in which the original soundtrack was replaced by a soundtrack from the second set, and the Continuous Sound version in which the original soundtrack was replaced by a soundtrack from the third set. In the following, Unrelated Speech, Abrupt Sounds, and Continuous Sound conditions will be referred to as the Nonoriginal conditions. A soundtrack was associated to a particular visual scene only once. The soundtracks were monophonic and sampled at 48,000 Hz. All dialogues were in French.

Apparatus

Participants were seated 57 cm away from a 21-in. CRT monitor with a spatial resolution of 1024×768

	Auditory conditions			
	Original	Unrelated speech	Abrupt sounds	Continuous sound
Saccade amplitudes (degree)	4.5 ± 0.2	4.9 ± 0.2	5.0 ± 0.2	4.9 ± 0.2
Fixation durations (ms)	430 ± 23	423 ± 21	412 ± 21	419 ± 22
Dispersions (degree)	4.8 ± 0.5	5.3 ± 0.5	5.6 ± 0.6	5.5 ± 0.6

Table 1. General eye movement parameters in each auditory condition. *Notes*: Saccade amplitudes and fixation durations are averaged over participants, whereas dispersions are averaged over stimuli. ($M \pm SE$).

pixels and a refresh rate of 75 Hz. The head was stabilized with a chin rest, forehead rest, and headband. The audio signal was presented via headphones (HD280 Pro, 64Ω, Sennheiser, Wedemark, Germany). Eye movements were recorded using an eye-tracker (Eyelink 1000, SR Research, Eyelink, Ottawa, Canada) with a sampling rate of 1000 Hz and a nominal spatial resolution of 0.01 degree of visual angle. We recorded the eye movements of the dominant eye in monocular pupil–corneal reflection tracking mode.

Procedure

Each participant viewed the 15 different conversation scenes. The different auditory versions were balanced (e.g., four scenes in Original condition, four in Unrelated Speech condition, four in Abrupt Sounds condition, and three in Continuous Sound condition). Participants were told to carefully look at each video. Each experiment was preceded by a calibration procedure, during which participants focused their gaze on nine separate targets in a 3 × 3 grid that occupied the entire display. A drift correction was carried out between each video, and a new calibration procedure was performed if the drift error was above 0.5°. Before each video, a fixation cross was displayed in the center of the screen for 1 s. After that time, and only if the participant looked at the center of the screen (gaze contingent display), the video was played on a mean gray level background. Between two consecutive videos, a gray screen was displayed for 1 s. To avoid any order effect, videos were randomly displayed. Each visual scene was seen in each auditory condition by 18 different participants.

Data extraction

Eye positions

The eye-tracker system sampled eye positions at 1000 Hz. Since videos had a frame rate of 25 frames per second, 40 eye positions were recorded per frame and per participant. In the following, an eye position is the median of the 40 raw eye positions: There is one eye position per frame and per subject. We discarded from analysis the eye positions landing outside the video area.

Saccades

Saccades were automatically detected by the Eyelink software using three thresholds: velocity (30°/s), acceleration (8000°/s^2), and saccadic motion (0.15°).

Fixations

Fixations were detected as long as the pupil was visible and as long as there was no saccade in progress.

Face labeling

The face of each conversation partner was marked by an oval mask. Since faces were moving, the coordinates of each mask were defined dynamically for each frame of each video. We used Sensarea, an in-house authoring tool allowing spatio-temporal segmentation of video objects to be performed automatically or semi-automatically (Bertolino, 2012).

Eye-tracking results

How does sound influence eye movements when viewing other people having a conversation? In this section, we characterize how some general eye movement parameters such as saccade amplitudes and fixation durations are affected by the auditory content. We also analyze the variability of eye movements between participants. Then, we perform a temporal analysis to describe how a given soundtrack influences observers' sequence of fixations across the exploration (scanpaths).

Global analysis

Saccade amplitudes

For each participant, we computed the mean saccade amplitude in each auditory condition (see Table 1). One-way repeated measures ANOVA with mean saccade

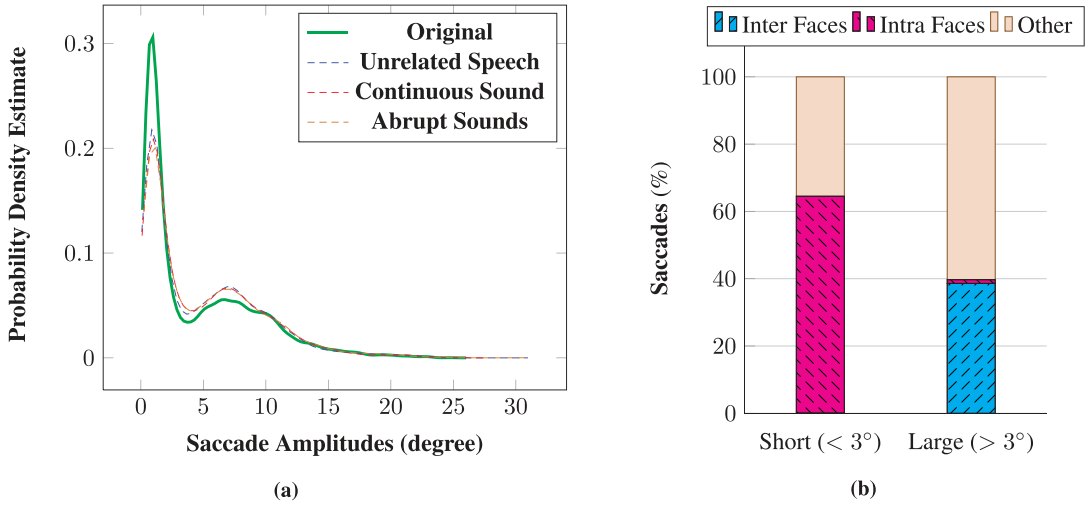

Figure 1. (a) Probability density estimate of saccade amplitudes in each auditory condition. The density is evaluated at 100 equally spaced points covering the range of data (ksdensity Matlab function). (b) Proportion of saccades starting from one face and landing on another one (Interfaces); starting from one face and landing on the same face (Intrafaces); and starting from or landing on the background (Other). Saccades are separated in two groups: <3° saccades (corresponding to the first mode of Figure 1a) and >3° saccades (corresponding to the second mode of Figure 1a).

amplitude per subject as a dependent variable and auditory condition (Original, Unrelated Speech, Abrupt Sounds, and Continuous Sound) as within-subject factor was performed. A principal effect of the auditory condition was found, $F(3, 213) = 7.72$; $p < 0.001$, and Bonferroni posthoc pairwise comparisons revealed that saccade amplitudes are higher for the three Nonoriginal conditions compared to the Original condition (all ps < 0.01). No difference was found between Non-original auditory conditions (all ps $= 1$).

Saccade amplitudes follow a bimodal distribution, with modes around 1° and 7°, as shown Figure 1a. We can notice that the first mode of Original distribution is significantly higher than the first mode of Unrelated Speech, Abrupt Sounds, and Continuous Sound distributions. (Three two-sample Kolmogorov-Smirnov tests between the Original condition and the three other conditions, all ps < 0.001). To further understand this bimodal distribution, we split the saccades into two groups: short (<3°) saccades, corresponding to the first mode, and large (>3°) saccades, corresponding to the second mode. In each group, we compared the proportion of saccades (a) starting from one face and landing on another one (Inter); (b) starting from one face and landing on the same one (Intra); and (c) starting from or landing on the background (Other; see Figure 1b). There are no Inter saccades in the first mode and almost no Intra saccades in the second mode. Thus it is reasonable to assume that the first mode represents the saccades made within a given face (from eyes to mouth, to nose, etc.) and that the second mode represents the saccade made between faces.

Fixation durations

We conducted one-way repeated measures ANOVA with mean fixation duration per subject as a dependent variable and auditory condition as within subject factor. We did not find any effect of the auditory condition, $F(3, 213) = 0.39$; $p = 0.76$. Fixation durations follow a classical positively skewed, long-tailed distribution.

Dispersion

To estimate the variability of eye positions between observers, we used a dispersion metric. For a frame and n observers ($\mathbf{p} = (x_i, y_i)_{i\in[1..n]}$ the eye position coordinates), the dispersion D is defined as follows:

$$D(\mathbf{p}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \qquad (1)$$

The dispersion is the mean Euclidian distance between the eye positions of different observers for a given frame. Small dispersion values reflect clustered eye positions.

We averaged dispersion values over all frames and compared the results obtained for the 15 videos in each auditory condition (Table 1). We conducted one-way repeated measures ANOVA with mean dispersion per video as a dependent variable and auditory condition as within subject factor. A principal effect of the auditory condition was found, $F(3, 42) = 17.97$; $p < 0.001$, and Bonferroni posthoc pairwise comparisons revealed that dispersion is higher in the three Non-original conditions compared to the Original condition

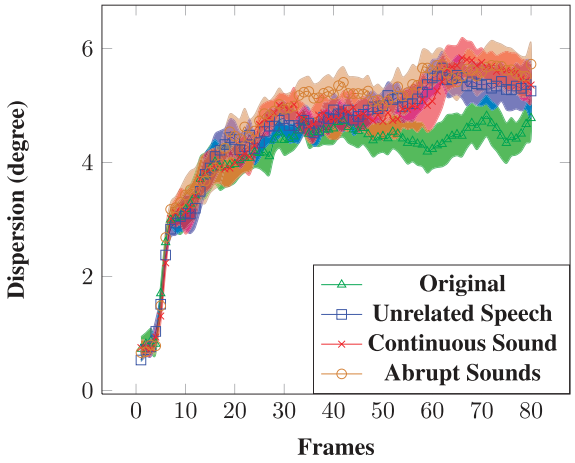

Figure 2. Temporal evolution of the dispersion between observers' eye positions. Dispersions are computed frame-by-frame and averaged over the 15 videos of each auditory condition. Values are given in degree of visual angle with error bars corresponding to the standard errors.

(all $ps < .001$). We found no difference between Nonoriginal conditions (Abrupt Sounds vs. Unrelated Speech: $p = 0.07$; Continuous Sound vs. Abrupt Sounds: $p = 1$; Continuous Sound vs. Unrelated Speech: $p = 0.79$).

We showed that in Nonoriginal auditory conditions, the dispersion between the eye positions of different subjects is higher and saccade amplitudes are larger. These results reflect a greater attentional synchrony in the Original condition: Eye positions are more clustered in a few regions of interest. To better understand these global results, we looked at the temporal evolution of gaze behavior and compared subjects' scanpaths in each auditory condition.

Temporal analysis

In this section, we first look at the temporal evolution of the variability between observers' eye positions (dispersion) and of their distance from the screen center (distance to center [DtC]). For the sake of clarity, only the evolution along the 80 first frames was plotted but analyses were carried out over whole videos. Then, for each auditory condition, we compare the number of fixations and the fixation sequences (scanpaths) landing on talking and mute faces. In the following, by talking face, we mean a face that talks in the Original auditory condition.

Dispersion

We represented the frame-by-frame evolution of dispersion (Figure 2). During the five first frames, dispersion remains small (around 0.5°), regardless of the auditory condition. Then, it increases sharply and reaches a plateau after the first second (around 25 frames) of visual exploration. During the first second, all dispersion curves are superimposed. But once the plateau has been reached, the dispersion curve in the Original condition stays below the others, as we found in the global analysis.

Distance to Center

DtC is defined, for a given frame, as the mean distance between observers' eye positions and the screen center. A small DtC value corresponds to a strong center bias, and can be seen as an indicator of the type of exploration strategy (active or passive). The center bias reflects the tendency one has to gaze more often at the center of the image than at the edges (see the Modeling section below). The DtC (not represented) follows the same pattern as dispersion. It stays small (around 0.5°) during the five first frames, then it increases sharply and reaches a plateau after the twentieth frame (around 6.5°). Contrary to dispersion, DtC curves do not differ significantly between auditory conditions during the whole experiment.

Fixation ratio

We matched the eye positions to the frame-by-frame labeled faces previously defined. We also manually spotted the time periods during which each face was speaking. Speaking and mute time periods were defined in the Original auditory condition, i.e., when the face was actually articulating. Thus, we were able to spatio-temporally distinguish talking faces from mute faces. For each of the 33 faces present in our stimuli and for each frame, we computed a fixation ratio, i.e., the number of fixations landing on the faces divided by the total number of fixations. We then averaged these ratios over the speaking and the mute periods of time (28 faces talked at least once and 27 faces were silent at least once; see Table 2). We found that talking faces attracted gaze around twice as much as mute faces, regardless of the auditory condition. One-way repeated measures ANOVA with fixation ratio on talking faces as a dependent variable and auditory condition as within factor was performed. A principal effect of the auditory condition was found, $F(3, 81) = 8.9$; $p < 0.001$, and Bonferroni posthoc pairwise comparisons revealed that talking faces were more fixated in the Original than in the three Nonoriginal conditions (all $ps < 0.001$), but that there was no difference between Nonoriginal conditions (all $ps = 1$).

The same analysis was performed with mute faces. We did not find any effect of the auditory condition, $F(3, 78) = 1.5$; $p = 0.21$. These ratios might seem low compared with the literature. This is understandable

	Auditory conditions			
	Original	Unrelated speech	Abrupt sounds	Continuous sound
Fixation in talking faces (%)	48 ± 5	40 ± 4	38 ± 4	38 ± 5
Fixation in mute faces (%)	20 ± 4	23 ± 3	22 ± 3	22 ± 3

Table 2. Fixation ratios (number of fixations landing on faces divided by the total number of fixations). *Notes*: Fixation ratios are computed for each face in each video. By averaging these ratios over speaking and silent time periods, we obtain fixation ratios for talking and mute faces. ($M \pm SE$).

since we used stimuli featuring conversation partners embedded in complex natural environments, and many objects that could also attract observers' gaze. To further understand how soundtracks impact on the timing of looks in talking and mute faces, we used a string edit distance to directly compare observers' scanpaths.

Scanpath comparison

To compare scanpaths, a classical method is to use the Levenshtein distance, a string edit distance measuring the number of differences between two sequences (Levenshtein, 1966). This distance gives the minimum number of operations needed to transform one sequence into the other (insertion, deletion, or substitution of a single character), and has been widely used to compare scanpaths. In this case, the compared sequence is the sequence of successive fixations made by an observer across visual exploration (see Le Meur & Baccino, 2013, for a review). Here, we used quite a simple approach, since we only intended to compare the observer fixation patterns in regions of interest (faces), without considering the distance between them. For a given video, we sampled the eye movement sequence of each subject frame by frame. To each frame, we assigned a character corresponding to the area of the scene currently looked at (face *a*, face *b*, …, background; see Figure 3). We also defined the ground truth sequence, or **GT**, of each video. If a video lasts m frames, **GT** is an array of length m, such as if face a speaks at frame i, then $\mathbf{GT}(i) = a$. If no face speaks at frame j, then $\mathbf{GT}(j) =$ background. This choice is quite conservative since even when no one is speaking, observers usually continue looking at faces. For each subject, we compared the Levenshtein distance between the fixation sequence recorded on each video and **GT**, normalized by the length m of the video. We conducted one-way repeated measures ANOVA with mean-normalized Levenshtein distance per subject as a dependent variable and auditory condition as within subject factor. A principal effect of the auditory condition was found, $F(3, 213) = 17.6$; $p < 0.001$, and Bonferroni posthoc pairwise comparisons revealed that the Levenshtein distance was smaller between **GT** and the eye movement sequences recorded in the Original

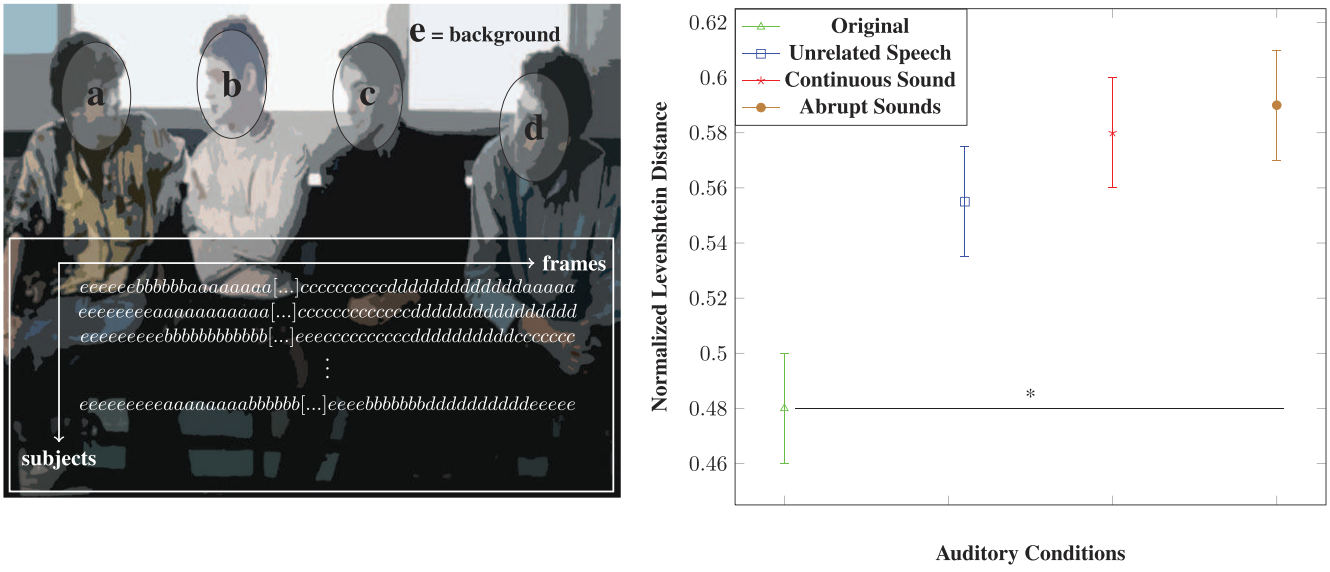

Figure 3. Left: Frames are split into five regions of interest (face *a*, face *b*, face *c*, face *d*, and background *e*). At the bottom, each line represents the scanpath of a subject: Each letter stands for the region the subject was looking at during each frame. Right: Mean normalized Levenshtein distance between the scanpaths and the ground truth sequence, in each auditory condition. Error bars correspond to the standard errors.

than in the three Nonoriginal conditions (all ps $<$ 0.001). No difference between Nonoriginal conditions was found (Abrupt Sounds vs. Unrelated Speech: $p = 0.12$; Continuous Sound vs. Abrupt Sounds: $p = 1$; Unrelated Speech vs. Continuous Sound: $p = 0.64$). Thus, we found a greater similarity between scanpaths and the ground truth sequences in Original than in Nonoriginal conditions.

Interim summary

We show that the presence of faces deeply impact visual exploration by attracting most fixations toward them. In particular, talking faces attract gaze around twice as many as mute faces, regardless of the auditory condition. In the Original auditory condition, eye positions are more clustered within face areas, leading to smaller saccade amplitudes. Temporal analysis reveals that, in contrast to mute faces, talking faces attract more observers' gazes in the Original condition. We find no significant difference between Nonoriginal conditions. These results are confirmed by the comparison between scanpaths and speech turn-taking, pointing out that in the Original condition, participants' gaze follows the speech turn-taking (**GT**) more closely than in Nonoriginal conditions.

To better characterize the differences between exploration strategies in each auditory condition, we quantify the importance of different visual features likely to drive gaze when viewing conversations. To do so, we model the probability distribution of eye positions by a mixture of different causes and separate their contributions with a statistical method, the EM algorithm.

Modeling

In this section, we quantify how soundtracks modulate the strength of potential gaze guiding features such as static and dynamic low-level visual saliencies, faces, and center bias (see below). To separate and quantify the contribution of the different gaze guiding features, we used the EM algorithm, a statistical method using observations (the recorded eye positions) to estimate the relative importance of each feature in order to maximize the global likelihood of the mixture model (Dempster, Laird, & Rubin, 1977). The EM algorithm is widely used in statistics and machine learning, and some recent studies have successfully applied it to visual attention modeling in static scenes (Gautier & Le Meur, 2012; Ho-Phuoc et al., 2010; Vincent et al., 2009). To our knowledge, EM has never been used on dynamic scenes. In order to represent the dynamic turn-taking of conversations, we computed the weights of the different features for each frame of each video.

Let $P(\mathbf{w}|f, v)$ be the probability distribution of n eye positions with coordinates $(\mathbf{w} = (x_i, y_i)_{i\in[1..n]}$, made by n different observers on frame f of video v. To break this probability distribution down into m different gaze guiding features, a classical method is to express P as a mixture of different causes Φ, each associated to a weight α:

$$P(\mathbf{w}|f, v) = \sum_{k=1}^{m} \alpha_k(f, v)\Phi_k(x, y, f, v), \text{with} \sum_{k=1}^{m} \alpha_k(f, v) = 1$$

P and Φ have the same dimensions as frames (720×576). The EM algorithm converges toward the most likely combination of weights, i.e., the one that optimizes the maximum likelihood of the data, given the eye position probability distribution P and the features Φ. The first step (expectation) takes all the visual features modeling the data (low-level static and dynamic saliencies, center bias, uniform distribution, and face masks) and converts them into two-dimensional (2-D) spatial probability distributions. Assuming that the current model (i.e., the weight combination) is correct, the algorithm labels each eye position with the corresponding probability of each 2-D spatial distribution. The second step (maximization) assumes that these probabilities are correct and sets the weights of the different features to their maximum likelihood values. These two steps are iterated, until a convergence threshold is reached. Finally, the best weight combination is found for each frame of each video in each auditory condition. This allows the frame-by-frame evolution of the relative importance of each feature to be followed. Below, we describe the features we chose for the mixture model: static and dynamic low-level saliencies, center bias, and faces (Figure 4).

Low-level saliency

To compute the saliency of video frames we used the spatio-temporal saliency model proposed in (Marat et al., 2009). This biologically inspired model, only based on luminance information, is divided into two main steps: a retina-like and a cortical-like stage. Before the retina stage, camera motion compensation is performed to extract only the moving areas relative to the background. The retina-like stage does not model the photoreceptor distribution. It extracts, on one hand, low spatial frequencies further processed in the dynamic pathway to extract moving areas in the video frame, and on the other hand, high spatial frequencies further processed in the static pathway to extract luminance orientation and frequency contrast. Then,

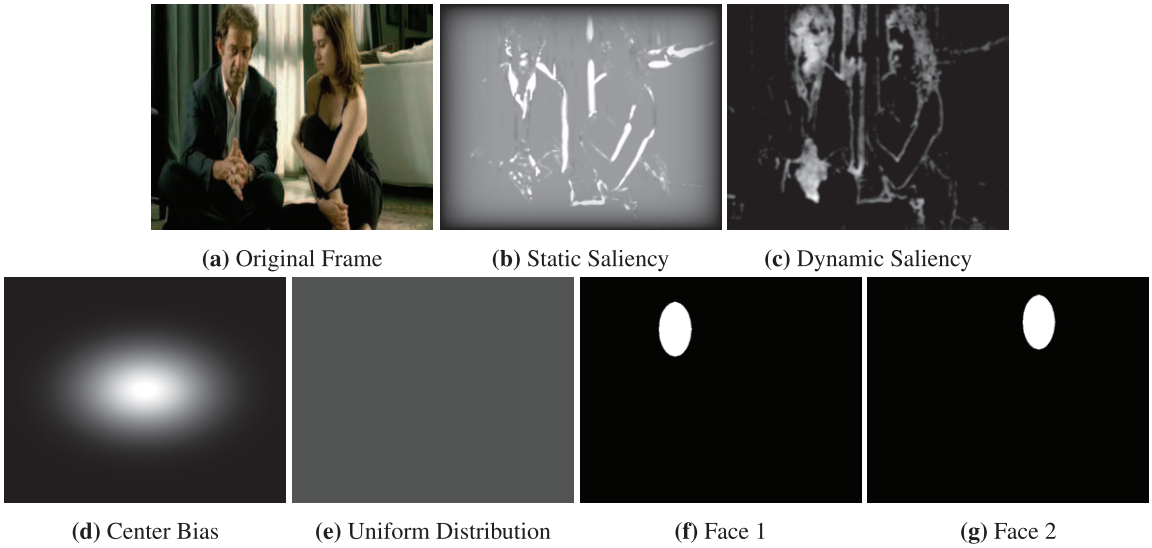

Figure 4. Features chosen to model the probability distribution of eye positions on each frame.

the cortical-like stage processes these two pathways with a bank of Gabor filters.

Static saliency: The Gabor filter outputs are normalized to strengthen the filtered frames having spatially distributed maxima. Then, they are added up, yielding a static saliency map (Figure 4b). This map emphasizes the high luminance contrast.

Dynamic saliency: Through the assumption of luminance constancy between two successive frames, motion estimation is performed for each spatial frequency of the bank of Gabor filters. Finally, a temporal median filter is applied over five successive frames to remove potential noise from the dynamic saliency map (Figure 4c). This map emphasizes the moving areas, returning the amplitude of the motion.

Center bias

Most eye-tracking studies reported that subjects tend to gaze more often at the center of the image than at the edges. Several hypotheses have been proposed to explain this bias. Some are stimuli-related, like the photographer bias (one often places regions of interest at the center of the picture); others are inherent to the oculomotor system (motor bias) or to the observers' viewing strategy (Marat et al., 2013; Tatler, 2007; Tseng et al., 2009). As in Gautier and Le Meur (2012), the center bias is modeled by a time-independent bidimensional Gaussian function, centered at the screen center as $N(0, \Sigma)$, with $\sum = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$ the covariance matrix and σ_x^2, σ_y^2 the variance. We chose σ_x and σ_y proportional to the frame size ($28° \times 22.5°$), and ran the algorithm with several values ranging from $\sigma_x = 2°$ to $\sigma_x = 3.5°$ and $\sigma_y = 1.6°$ to $\sigma_y = 2.8°$. Changing these values did not significantly change the outputs. The results presented in this study were obtained with $\sigma x = 2.3°$ and $\sigma_y = 1.9°$ (Figure 4d).

Uniform distribution

Fixations occur at all positions with equal probability (Figure 4e). This feature is a catch-all hypothesis that stands for any fixations that are not explained by other features. The lower the weight of this feature is, the better the other features will explain the data.

Faces

For a given frame, we created as many face maps as faces present in the frame. Face maps were made up of the corresponding face binary masks described in the Method section (Figure 4f, g). In Figure 5a, the All Faces weight corresponds to the sum of the weights of the different face maps in the frame.

For each video, the weight of each feature was averaged over time. We compared the weights of the different features for each video, regardless of the auditory condition, as well as the weights of each feature in the different auditory conditions (Figure 5a). Faces were by far the most important features explaining gaze behavior, regardless of the auditory condition (weights ≥ 0.6). This result matches the fixation ratios reported in Table 2: The fixation ratio in all faces (i.e., mute + talking) is around 60%.

We performed repeated measures ANOVA with Feature Weights (Static Saliency, Dynamic Saliency, Center Bias, Uniform, and Faces) and Auditory Conditions (Original, Unrelated Speech, Continuous

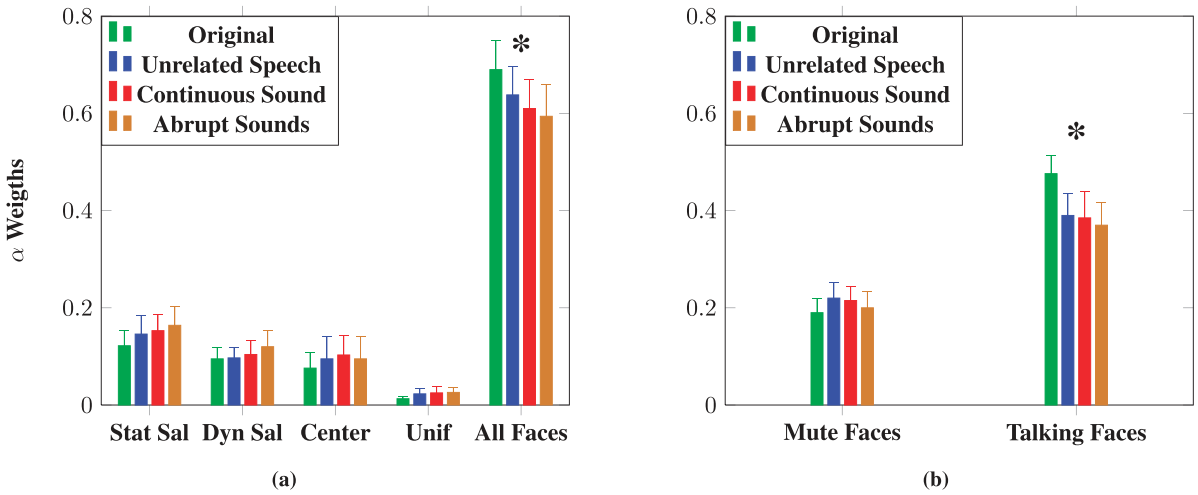

Figure 5. (a) Weights of the features chosen to model the probability distribution of eye positions (the sum of the five features equals one). (b) Contributions of talking and mute faces to the all faces weight (the sum of the two equals the all faces weight). For each video, weights are averaged over all frames. Results are then averaged over all videos and error bars correspond to the standard errors. *Marks a significant difference between auditory conditions for the corresponding feature (Bonferroni pairwise posthoc comparisons, see below for further details).

Sound, and Abrupt Sounds) as within-subject factors. The main effect of Feature Weights yielded an F ratio of $F(4, 56) = 145.95$, $p < 0.001$. Bonferroni pairwise comparisons revealed that the weight of Faces was significantly higher than the others (all ps < 0.001). There was no significant difference between Static Saliency, Dynamic Saliency, and Center Bias (Static vs. Dynamic: $p = 1$; Static vs. Center Bias: $p = 0.67$; Dynamic vs. Center Bias: $p = 1$). Uniform distribution was lower than Static and Dynamic Saliencies (Uniform vs. Static: $p < 0.001$; Uniform vs. Dynamic: $p = 0.06$), but was not significantly different from Center Bias (Uniform vs. Center Bias: $p = 0.18$). The main effect of Auditory conditions yielded an F ratio of $F(3, 42) = 74.39$, $p < 0.001$. The interaction effect was also significant, with an F ratio of $F(12, 168) = 6.43$, $p < 0.001$. Bonferroni pairwise comparisons between each auditory condition for each feature were calculated.

Static Saliency, Dynamic Saliency, Center Bias and Uniform features: No significant difference between auditory conditions (all ps $= 1$).

All Faces: We found that All Faces weight is higher in the Original auditory condition than in other conditions (Original vs. Unrelated Speech: $p = 0.019$; Original vs. Abrupt Sounds: $p < 0.001$; Original vs. Continuous Sound: $p < 0.001$). We found no significant difference between Unrelated Speech, Abrupt Sounds, and Continuous Sound (Abrupt Sounds vs. Unrelated Speech: $p = 0.32$; Continuous Sound vs. Abrupt Sounds: $p = 1$; Unrelated Speech vs. Continuous Sound: $p = 1$).

Talking and mute faces

We tagged manually the periods of time during which each face was speaking or mute (in the Original auditory condition), as it was done to calculate the fixation ratios. By averaging the weights of face maps over these periods of time, we were able to separate the contribution of talking faces from mute faces. The weights shown in Figure 5b nicely match the fixation ratios reported in Table 2 around 20% for mute faces regardless of the auditory condition, around 50% for talking faces in the Original condition, and 40% in the Nonoriginal condition.

We conducted repeated measures ANOVA with the face category (mute and talking) and the auditory condition (Original, Unrelated Speech, Continuous Sound, and Abrupt Sounds) as within-subject factors. The main effect of the face category yielded an F ratio of $F(1, 14) = 106.75$, $p < 0.001$. The maps containing the talking faces had a mean weight of 0.45 and the maps containing the mute faces had a mean weight of 0.2. The main effect of auditory conditions yielded an F ratio of $F(3, 42) = 5.16$, $p = 0.004$. The interaction effect was also significant, with an F ratio of $F(3, 42) = 20.14$, $p < 0.001$.

Bonferroni pairwise posthoc comparisons revealed that talking face weights were higher in the Original auditory condition than in the other conditions (all ps < 0.001). For the weights of the mute face map, we found no difference between the auditory conditions (Original vs. Unrelated Speech: $p = 0.70$; Original vs. Abrupt Sounds: $p = 1$; Original vs. Continuous Sound: $p = 1$; Unrelated Speech vs. Continuous Sound: $p = 1$;

Abrupt Sounds vs. Unrelated Speech: $p = 1$; Continuous Sound vs. Abrupt Sounds: $p = 1$).

Interim summary

We show that in dynamic conversation scenes, low-level saliencies (both static and dynamic) and center bias are poor gaze-guiding features compared to faces, and especially to talking faces. Even if the related speech enhances talking face weight by 10%, gaze is mostly driven toward talking faces by visual information. Indeed, even with unrelated auditory information, the weight of talking faces is still twice the weight of mute faces. We found no difference between unrelated auditory conditions.

Discussion

Gaze attraction toward faces is widely agreed upon. However, when trying to model visual attention, authors rarely take faces into account and never consider the auditory information that is usually part of dynamic scenes. In this paper, we quantify how auditory information influences gaze when viewing a conversation. For this purpose we eye-tracked participants viewing conversation scenes in different auditory conditions (original speech, unrelated speech, noises of moving objects, and continuous landscape sound), and we compared their gaze behaviors. First, we comment on our results with reference to previous studies on faces and visual attention. Then we discuss how speech and other sounds modulate gaze behavior when viewing conversations. Finally, we propose groundwork for an audiovisual saliency model.

Faces: Strong gaze attractors

We found that in every auditory condition, faces attract the most fixations (>60%). This central role of faces in visual exploration is reflected by saccade amplitude distribution. Classically, saccades made during the free exploration of natural scenes follow a positively skewed, long-tailed distribution (Bahill, Adler, & Stark, 1975; Coutrot et al., 2012; Tatler, Baddeley, & Vincent, 2006). In contrast, here we found a bimodal distribution, with modes around 1° and 7°. An interpretation is that when viewing scenes including faces, participants make at least two kinds of saccades: intraface (from eyes to mouth to nose, etc.) and interface (from one conversation partner to another) saccades. We tested this hypothesis by comparing the proportion of intraface and interface saccades and their amplitudes. We found that almost all intraface saccades were concentrated within the first mode, while all interface saccades were concentrated within the second one. This result is confirmed by the mean face area (around 3° × 5°, matching the first mode) and the mean distance between conversation partners (around 10°, matching the second mode) present in our stimuli. Moreover, fixation durations were longer (around 420 ms) than usually reported in the literature (250–350 ms), which supports the idea of long explorations of a few regions of interest, like faces (Pannasch, Helmert, Herbold, Roth, & Henrik, 2008; Smith & Mital, 2013).

Studies have long established the specificity of faces in visual perception (Yarbus, 1967), but the use of static images made the generalization of their results to the real world problematic. Recently, some social gaze studies used dynamic stimuli to get as close as possible to ecological situations and confirmed that observers spend most of the time looking at faces (Foulsham et al., 2010; Frank, Vul, & Johnson, 2009; Hirvenkari et al., 2013; Võ et al., 2012). This exploration strategy leads eye positions to cluster on faces (Mital et al., 2010), and more generally induces a decrease in eye position dispersion, as compared to natural scenes without semantically rich regions (e.g., landscapes; Coutrot & Guyader, 2013). Our results are consistent with a very strong impact of faces on gaze behavior when exploring natural dynamic scenes. They extend previous findings by highlighting that the presence of faces in natural scenes leads to a bimodal saccade amplitude distribution corresponding to the saccades made within a same face and between two different faces. This strong impact of faces occurred even though the stimuli we chose featured conversation partners who were embedded in complex natural environments (cafe, office, street, corridor) and many objects that could also attract observers' gaze.

We also quantified and compared the strength of different gaze guiding features such as static and dynamic low-level visual saliencies, faces, and center bias. Our results show that after a short predominance of the center bias (during the first five frames), faces are by far the most pertinent features to explain gaze allocation. This five-frame delay is classically reported for reflexive saccades toward peripheral target (latency around 150–250 ms; Carpenter, 1988; Yang, Bucci, & Kapoula, 2002). Then, we found that although the weights are globally high for every face, they are even higher for talking faces, regardless of the auditory condition. This indicates that visual cues are sufficient to efficiently drive gaze toward speakers. Yet, the quite low weights we found for both static and dynamic low-level dynamic saliencies suggest that their contribution to gaze guiding is slight. This result reinforces previous studies claiming that classical visual attention models do not account for human eye fixations when looking

at static images involving complex social scenes (Birmingham & Kingstone, 2009). Thus, to explain the attractiveness of speakers even without their related speech, higher level visual cues might be invoked, such as expressions or body language (Richardson, Dale, & Shockley, 2008). However, these are more difficult to model.

Influence of related speech

We found that if the fixation ratio is globally high for every face, it is even higher for talking faces, regardless of the auditory condition. As stated in the previous paragraph, this result suggests that since observers are able to follow speech turn-taking without the related speech soundtrack, visual and auditory information are in part redundant in guiding the viewers' gaze (as was also reported in Hirvenkari et al., 2013). So, what is the added value of sound? A body of consistent evidence shows that with the related speech, observers follow the speech turn-taking even more closely. First, the dispersion between eye positions made with the related speech was found to be smaller than without it (as was also reported in Foulsham & Sanderson, 2013). Second, when we modeled the gaze-attracting power of different visual features, the weights of talking faces were found to be significantly higher with than without the related speech. Third, the first mode of saccade amplitude distribution (corresponding to the intraface type of saccade) was found to be much greater with than without the related speech. These results show that without the related speech soundtrack, observers were less clustered on talking faces, making fewer little saccades (from eyes to mouth to nose), usually made to better understand speakers momentary emotional state, or to support speech perception by sampling mouth movements and other facial nonverbal cues (Buchan, Paré, & Munhall, 2007; Vatikiotis-Bateson et al., 1998; Võ et al., 2012). Finally, we compared the scanpaths between subjects in each auditory condition to a ground truth sequence representing speech turn-taking. We found a greater similarity between subjects' scanpaths and the ground truth sequence in the original auditory condition. This result is coherent with the recent studies of Hirvenkari et al. (2013) and Foulsham and Sanderson (2013), which noted the temporal relationship between speech onsets and the deployment of visual attention. Both studies reported that with the related speech soundtrack, fixations on the speaker increased right after speech onset, peaking about 800 ms to 1 s later. Removing the sound did not affect the general gaze pattern, but it did change the speed at which fixations moved to the speaker. It may be consistent with considering speech as an alerting signal telling that another conversation partner has taken the floor. Without related speech, observers have to realize that the speakership has shifted and seek the new speaker, which could explain the lower similarity between their scanpaths and the speech turn-taking. In this section, we discussed gaze behavior between Original and Nonoriginal conditions, but what about the differences between Nonoriginal conditions?

Influence of other soundtracks

Our results show an effect of the related speech on eye movements while watching conversations. But what about unrelated sounds? Studies showed that presenting natural images and lateralized natural sounds biased observers' gazes towards the part of the image corresponding to the sound source (Onat et al., 2007; Quigley, Onat, Harding, Cooke, & König, 2008). Moreover, this spatial bias is dependent on the image saliency presented without any sound, meaning that gaze behavior is the result of an audiovisual integration process. Yet, our study is quite different from these, since we used unspatialized soundtracks and dynamic stimuli. Other studies investigated the perception of audiovisual synchrony for complex events by presenting speech versus object-action video clips at a range of stimulus onset asynchronies (Vatakis & Spence, 2006). Participants were significantly better at judging the temporal order of streams (auditory or visual) for the object actions than for the speech video clips, meaning that cross-modal temporal discrimination performance is better for audiovisual stimuli of lower complexity, compared to stimuli having continuously varying properties. Indeed, authors argued that since speech presents a fine temporal correlation between sound and vision (phoneme and viseme), judging temporal order in audiovisual speech may be more difficult than for abrupt noises like moving object sounds (Vroomen & Stekelenburg, 2011). Thus, audiovisual integration seems to be linked to the abrupt or slowly changing nature of audiovisual component signals, and to their correlation. That is why we chose to investigate how visual exploration is influenced by unrelated speech soundtracks (is speech special?), sound of moving objects (abrupt sound onsets), and landscape sounds (slowly varying components).

Surprisingly, we found no difference between the three Nonoriginal auditory conditions, whether in terms of dispersion between eye positions, saccade amplitudes, fixation durations, scanpath comparisons, fixation ratios in faces (mute or talking), or weights of any features computed by the EM algorithm. A reason for this absence of effect might be found in the notion of audiovisual binding.

A classical view of audiovisual integration is that audio and visual streams are separately processed before interaction automatically occurs, leading to an integrated percept (Calvert, Spence, & Stein, 2004). Other studies suggested that audiovisual fusion could be conceived as a two-stage process, beginning by binding together pieces of audio and video that present a certain amount of spatio-temporal correlation, before the actual integration (Berthommier, 2004). A recent study reinforced this idea by showing that it is possible to unbind visual and auditory streams (Nahorna, Berthommier, & Schwartz, 2012). To do so, the authors used the McGurk effect as a marker of audiovisual integration: The more it occurs, the more visual and auditory information the participants integrate. Results showed that if a given McGurk stimulus (visual /ga/ dubbed onto an acoustic /ba/) is preceded by an incoherent audiovisual context, the amount of McGurk effect (perception of /da/; McGurk & MacDonald, 1976), and thus the audiovisual integration, is largely reduced. The authors showed that even a very short incoherent audiovisual context (one syllable) is enough to cause unbinding.

In our study, there might be no difference in gaze behavior between Nonoriginal auditory conditions simply because unrelated speech, object noise, and landscape sound soundtracks are not temporally correlated enough with the visual information to pass through the binding stage, preventing any further integration. In the three Nonoriginal auditory conditions, observers might just filter out the unbound audio information and focus on the sole visual stream. Thus, any unrelated soundtracks or no soundtrack at all might result to the same gaze behavior, only driven by visual information. This interpretation is confirmed by the results of two recent papers that compared the gaze behavior of participants watching videos with or without their original soundtrack (Coutrot et al., 2012; Foulsham & Sanderson, 2013). Foulsham et al. (2010) used dynamic conversations as stimuli and found higher dispersion between eye positions and larger saccade amplitudes in the visual condition than in the audiovisual condition, which is coherent with our previous results (Coutrot et al., 2012). In fact, we also found higher dispersion in the visual condition than in audio-visual conditions, but without larger saccade amplitudes. Since we used various videos as stimuli (not specifically involving faces), these results corroborate the idea developed at the beginning of this Discussion: that the presence of faces induces an intraface and interface type of saccade. As explained, removing the original soundtrack increases the inter/intraface saccade ratio, resulting in an increase in saccade amplitude. On the contrary, when the visual scenes do not involve faces, removing the original soundtrack yields smaller saccades: Observers might become less active and make less goal-directed saccades.

It is interesting to notice that this binding phenomenon has been understood and used by filmmakers for a long time. For instance, the French composer and film theorist Michel Chion (1994, p. 40) denies the very notion of soundtrack as a coherent unity:

> By stating that there is no soundtrack I mean first of all that the sounds of a film, taken separately from the image, do not form an internally coherent entity on equal footing with the image track. Second, I mean that each audio element enters into simultaneous vertical relationship with narrative elements contained in the image (characters, actions) and visual elements of texture and setting. These relationships are much more direct and salient than any relations the audio element could have with other sounds. It's like a recipe: Even if you mix the audio ingredients separately before pouring them into the image, a chemical reaction will occur to separate out the sounds and make each react on its own with the field of vision.

Chion (1994), Nahorna et al. (2012), and this study agree on the necessity for sound to "enter into simultaneous vertical relationship" (i.e., to be correlated) with visual information so as to be bound and integrated with it, or using Chion's words, to "react" with it.

Toward an audiovisual saliency model

In many situations, low-level visual saliency models fail to predict fixation locations (Tatler et al., 2011). For scenes involving semantically interesting regions (Nyström & Holmqvist, 2008; Rudoy, Goldman, Shechtman, & Zelnik-Manor, 2013) and faces (Birmingham & Kingstone, 2009), it has been shown that high-level factors override low-level factors to guide gaze. In this paper, we modeled the probability distribution of eye positions across each video with the EM algorithm, a statistical method allowing the contribution of different gaze guiding features such as faces, low-level visual saliency and center bias to be separated and quantified. Regardless of the auditory condition, the weight associated to faces exceeded by far the weight associated to any other features. We found that the weight of low-level saliency is at the same level as center bias or chance. This supports the idea that low-level factors are not pertinent to explain gaze behavior when faces are present and extends it to dynamic scenes. We also found that even if the related speech enhances talking faces' weight by 10%, gaze is mostly driven toward talking faces by visual information. Indeed, even

with unrelated auditory information, the talking face weight is still twice the mute faces' weight.

Thus, in addition to already existing face detectors (Cerf et al., 2008; Marat et al., 2013), future audiovisual saliency models should include visual or audiovisual speaker diarization algorithms. Distinguishing silence from speech situations, and identifying the location of the active speaker in the latter case, remains a challenge, particularly in ecological—and thus noisy—environments. Yet many recent studies try to address this issue, for instance by exploiting the coherence between the speech acoustic signal and the speaker's lip movements (Blauth, Minotto, Jung, Lee, & Kalker, 2012; Noulas, Englebienne, & Kröse, 2012; see Anguera et al., 2012, for a review).

To sum up, to predict eye positions made while viewing conversation scenes, we think that future saliency models should detect talking and silent faces. If the scene comes with its related speech soundtrack, 50% of the total saliency should be attributed to talking faces, 20% to mute faces. The remaining should be shared between center bias (mainly during the five first frames) and low-level saliency. If the scene comes with unrelated soundtrack, the weight of talking faces should be slightly lowered to the benefit of the other features. Nevertheless, talking faces should remain the most attractive feature.

Conclusion

We find that when viewing ecological conversations in complex natural environment, participants look more at faces in general and at talking faces in particular, regardless of the auditory information. This result suggests that if auditory information does influence viewers' gaze, visual information is still the leading factor. We do not find any difference between the different types of unrelated soundtracks (unrelated speech, moving object abrupt noises, and continuous landscape sound). We hypothesize that unrelated soundtracks are not correlated enough with the visual information to be bound to it, preventing any further integration. However, hearing the original speech soundtrack makes participants follow the speech turn-taking more closely. This behavior increases the number of small intraface saccades and reduces the variability between eye positions. Using a statistical method, we quantify the propensity of several classical visual features to drive gazes (faces, center bias, static and dynamic low-level saliencies). We find that classical low-level saliency globally fails to predict eye positions, whereas faces (and especially talking faces) are good predictors. Therefore, we suggest the joint use of face detector and speaker diarization algorithms to distinguish talking from mute faces and label them with appropriate weights.

Keywords: faces, speech, gaze, scanpath, saliency, audiovisual integration, expectation-maximization, database

Acknowledgments

The authors would like to thank Jean-Luc Schwartz, Jonas Chatel-Goldman, and two anonymous referees for their enlightening comments on the manuscript.
Commercial relationships: none.
Corresponding author: Antoine Coutrot.
Email: antoine.coutrot@gipsa-lab.fr.
Address: Gipsa-lab, CNRS & Grenoble-Alpes University.

References

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transaction on Audio, Speech, & Language Processing*, *20*(2), 356–370.

Bahill, T., Adler, D., & Stark, L. (1975, June). Most naturally occurring human saccades have magnitudes of 15 degrees or less. *Investigative Ophthalmology & Visual Science*, *14*(6), 468–469, http://www.iovs.org/content/14/6/468. [PubMed] [Article]

Bailly, G., Perrier, P., & Vatikiotis-Bateson, E. (2012). *Audiovisual speech processing*. Cambridge, UK: Cambridge University Press.

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

Berthommier, F. (2004). A phonetically neutral model of the low-level audiovisual interaction. *Speech Communication*, *44*(1–4), 31–41.

Bertolino, P. (2012). Sensarea: An authoring tool to create accurate clickable videos. In *10th workshop on content-based multimedia indexing* (pp. 1–4). Annecy, France.

Bindemann, M., Burton, A. M., Hooge, I. T. C., Jenkins, R., & de Haan, E. H. F. (2005). Faces retain attention. *Psychonomic Bulletin & Review*, *12*(6), 1048–1053.

Bindemann, M., Burton, A. M., Langton, S. R. H., Schweinberger, S. R., & Doherty, M. J. (2007). The control of attention to faces. *Journal of Vision*, *7*(10):15, 1–8, http://www.journalofvision.org/

content/7/10/15, doi:10.1167/7.10.15. [PubMed] [Article]

Birmingham, E., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision Research*, *49*, 2992–3000.

Blauth, D. A., Minotto, V. P., Jung, C. R., Lee, B., & Kalker, T. (2012). Voice activity detection and speaker localization using audiovisual cues. *Pattern Recognition Letters*, *33*(4), 373–380.

Boltz, M. G. (2004). The cognitive processing of film and musical soundtracks. *Memory & Cognition*, *32*(7), 1194–1205.

Boremanse, A., Norcia, A., & Rossion, B. (2013). An objective signature for visual binding of face parts in the human brain. *Journal of Vision*, *13*(11):6, 1–18, http://www.journalofvision.org/content/13/11/6, doi:10.1167/13.11.6. [PubMed] [Article]

Borji, A., & Itti, L. (2012). State-of-the-art in visual attention modeling. *IEEE Transactions on Patterns Analysis and Machine Intelligence*, *35*(1), 185–207.

Branigan, E. (2010). Soundtrack in mind. *Projections*, *4*(1), 41–67.

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

Buswell, G. T. (1935). *How people look at pictures: A study of the psychology of perception in art*. Chicago: University of Chicago Press.

Calvert, G., Spence, C., & Stein, B. E. (2004). *Handbook of multisensory processes*. Cambridge, MA: MIT Press.

Carpenter, R. H. S. (1988). *Movements of the eyes* (2nd rev. & enlarged ed.). London, England: Pion Limited.

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. In *Advances in neural information processing systems* (pp. 241–248).

Chion, M. (1994). *Audio-vision: Sound on screen*. New York: Columbia University Press.

Cohen, A. J. (2005). How music influences the interpretation of film and video: Approaches from experimental psychology. In R. A. Kendall & R. W. H. Savage (Eds.), *Selected reports in ethnomusicology: Perspectives in systematic musicology* (pp. 15–36). Los Angeles: Department of Ethnomusicology, University of California.

Coutrot, A., & Guyader, N. (2013). Toward the introduction of auditory information in dynamic visual attention models. In *IEEE international workshop on image analysis for multimedia interactive services (WIAMIS)* (pp. 1–4). Paris, France.

Coutrot, A., Guyader, N., Ionescu, G., & Caplier, A. (2012). Influence of soundtrack on eye movements during video exploration. *Journal of Eye Movement Research*, *5*(4), 1–10.

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision*, *10*(4):16, 1–17, http://www.journalofvision.org/content/10/4/16, doi:10.1167/10.4.16. [PubMed] [Article]

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–38.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review*, *105*(3), 482–498.

Foulsham, T., Cheng, J. T., Tracy, J. L., Henrich, J., & Kingstone, A. (2010). Gaze allocation in a dynamic situation: Effects of social status and speaking. *Cognition*, *117*(3), 319–331.

Foulsham, T., & Sanderson, L. A. (2013). Look who's talking? Sound changes gaze behaviour in a dynamic social scene. *Visual Cognition*, *21*(7), 922–944.

Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, *110*, 160–170.

Gautier, J., & Le Meur, O. (2012). A time-dependent saliency model combining center and depth biases for 2D and 3D viewing conditions. *Cognitive Computation*, *4*, 1–16.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233.

Hershler, O., & Hochstein, S. (2005). At first sight: A high-level pop out effect for faces. *Vision Research*, *45*, 1707–1724.

Hirvenkari, L., Ruusuvori, J., Saarinen, V.-M., Kivioja, M., Peräkylä, A., & Hari, R. (2013). Influence of turn-taking in a two-person conversation on the gaze of a viewer. *PLoS ONE*, *8*(8), 1–6.

Ho-Phuoc, T., Guyader, N., & Guerin-Dugue, A. (2010). A functional and statistical bottom-up saliency model to reveal the relative contributions of low-level visual guiding factors. *Cognitive Computation*, *2*(4), 344–359.

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

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene

analysis. *IEEE Transactions on Patterns Analysis and Machine Intelligence*, *20*(11), 1254–1259.

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*(11), 4302–4311.

Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology*, *15*, 1943–1947.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227.

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.

Le Meur, O., & Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior Research Methods*, *45*(1), 251–266.

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.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Marat, S., Ho-Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., & Guérin-Dugué, A. (2009). Modelling spatio-temporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision*, *82*(3), 231–243.

Marat, S., Rahman, A., Pellerin, D., Guyader, N., & Houzet, D. (2013). Improving visual saliency by adding 'face feature map' and 'center bias'. *Cognitive Computation*, *5*(1), 63–75.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

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*(1), 5–24.

Nahorna, O., Berthommier, F., & Schwartz, J.-L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *Journal of the Acoustical Society of America*, *132*(2), 1061–1077.

Noulas, A. K., Englebienne, G., & Kröse, B. J. A. (2012). Multimodal speaker diarization. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *34*(1), 79–93.

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

Onat, S., Libertus, K., & König, P. (2007). Integrating audiovisual information for the control of overt attention. *Journal of Vision*, *7*(10):11, 1–16, http://www.journalofvision.org/content/7/10/11, doi:10.1167/7.10.11. [PubMed] [Article]

Pannasch, S., Helmert, J. R., Herbold, A.-K., Roth, K., & Henrik, W. (2008). Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions. *Journal of Eye Movement Research*, *2*(4), 1–19.

Quigley, C., Onat, S., Harding, S., Cooke, M., & König, P. (2008). Audio-visual integration during overt visual attention. *Journal of Eye Movement Research*, *1*(2), 1–17.

Richardson, D., Dale, R., & Shockley, K. (2008). Synchrony and swing in conversation: Coordination, temporal dynamics, and communication. In I. Wachsmuth, M. Lenzen, & G. Knoblich (Eds.), *Embodied communication* (pp. 75–94). New York: Oxford University Press.

Rudoy, D., Goldman, D. B., Shechtman, E., & Zelnik-Manor, L. (2013). Learning video saliency from human gaze using candidate selection. In *Conference on computer vision and pattern recognition* (pp. 4321–4328). Portland, OR.

Schwartz, J.-L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: A taxonomy of models of audiovisual fusion in speech perception. In R. Campbell, B. Dodd, & D. K. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 85–108). Hove, UK: Psychology Press.

Smith, T. J., & Mital, P. K. (2013). Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of Vision*, *13*(8):16, 1–24, http://www.journalofvision.org/content/13/8/16, doi:10.1167/13.8.16. [PubMed] [Article]

Song, G., Pellerin, D., & Granjon, L. (2013). Different types of sounds influence gaze differently in videos. *Journal of Eye Movement Research*, *6*(4), 1–13.

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). New York: Lawrence Erlbaum.

Tatler, B. W. (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*(14):4, 1–17, http://www.journalofvision.org/content/7/14/4, doi:10.1167/7.14.4. [PubMed] [Article]

Tatler, B. W., Baddeley, R. J., & Vincent, B. T. (2006). The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, *46*, 1857–1862.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5):5, 1–23, http://www.journalofvision.org/content/11/5/5, doi:10.1167/11.5.5. [PubMed] [Article]

Theeuwes, J., & Van der Stigchel, S. (2006). Faces capture attention: Evidence from inhibition of return. *Visual Cognition*, *13*(6), 657–665.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

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*(7):4, 1–16, http://www.journalofvision.org/content/9/7/4, doi:10.1167/9.7.4. [PubMed] [Article]

Vatakis, A., & Spence, C. (2006). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research*, *1111*, 134–142.

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisualspeech perception. *Perception & Psychophysics*, *60*(6), 926–940.

Vincent, B. T., Baddeley, R. J., Correani, A., Troscianko, T., & Leonards, U. (2009). Do we look at lights? Using mixture modelling to distinguish between low- and high-level factors in natural image viewing. *Visual Cognition*, *17*(6–7), 856–879.

Võ, M. L. H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision*, *12*(13):3, 1–14, http://www.journalofvision.org/content/12/13/3, doi:10.1167/12.13.3. [PubMed] [Article]

Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition*, *118*(1), 75–83.

Yang, Q., Bucci, M. P., & Kapoula, Z. (2002). The latency of saccades, vergence, and combined eye movements in children and in adults. *Investigative Ophthalmology & Visual Science*, *43*(9), 2939–2949, http://www.iovs.org/content/43/9/2939. [PubMed] [Article]

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum.

Zeppelzauer, M., Mitrovic, D., & Breiteneder, C. (2011). Cross-modal analysis of audio-visual film montage. In *International conference on computer communications and networks* (pp. 1–6). Maui, Hawaii.