Video viewing: do auditory salient events capture visual attention?

Antoine Coutrot, Nathalie Guyader, Gelu Ionescu, Alice Caplier

Annals of Telecommunications manuscript No.
(will be inserted by the editor)

Video viewing: do auditory salient events capture visual attention?

Antoine Coutrot · Nathalie Guyader · Gelu Ionescu · Alice Caplier

Received: 3 September 2012 / Accepted: 27 December 2012

Abstract We assess whether salient auditory events contained in soundtracks modify eye movements when exploring videos. In a previous study, we found that on average, non-spatial sound contained in video soundtracks impacts on eye movements. This result indicates that sound could play a leading part in visual attention models to predict eye movements. In this research, we go further and test whether the effect of sound on eye movements is stronger just after salient auditory events. To automatically spot salient auditory events, we used two auditory saliency models: the *Discrete Energy Separation Algorithm* and the Energy model. Both models provide a saliency time curve, based on the fusion of several elementary audio features. The most salient auditory events were extracted by thresholding these curves. We examined some eye movements parameters just after these events rather than on all the video frames. We showed that the effect of sound on eye movements (variability between eye positions, saccade amplitude and fixation duration) was not stronger after salient auditory events than on average over entire videos. Thus, we suggest that sound could impact on visual exploration not only after salient events but in a more global way.

Keywords saliency · eye movements · sound · videos · attention · multimodality · audio visual

1 Introduction

At any time, our brain perceives tremendous amount of information. Despite its substantial capacity, it cannot attach the same importance to each stimulus. To select the most pertinent ones, the brain uses a filter, called *attention*. When one visually explores its surroundings, the regions that are the most likely to attract attention are called *salient* regions. During the last decades, the modeling of saliency has been a very active field of research, from neurosciences to computer vision. Saliency models rely on the detection of spatial locations where the local properties of the visual scene such as color, motion, luminance, edge orientation... significantly differ from the surrounding image attributes [16] [22] [33]. Saliency models are evaluated by comparing the predicted salient regions with the areas actually looked at by participants during eye-tracking experiments.

Being able to predict the salient regions of an image or a video leads to a multitude of applications. For instance, saliency-based video compression algorithms are particularly efficient [15]. For each video frame, these algorithms encode the salient areas with a better resolution than the rest of the scene. Since one perceives only a small area around the center of gaze at high resolution (the fovea, around 3 $^\circ$ of visual angle), the distortion of non salient regions does not impact the perceived quality of the visual stimulus [3] [21]. Another application of saliency models is the automatic movie summarization [8]; video summary contains the most salient frames spotted by a visual attention model. The increasing availability of video bases implies a growing need for powerful indexation tools: automatically extracting the most salient frames of a video is an efficient way to evaluate its relevance. Saliency models also exist (although to a smaller extent) for audio signals. Auditory saliency models have been developed to detect the prominent syllable and word locations in speech [18] or to automatically annotate music with text tags (music style, mood, speech...) [32].

The existence of a strong interaction between vision and audition is well known, as reflected by the numerous audio-visual illusions [23] [35]. Previous studies showed that sound modified the way we explore visual scenes. For instance, a spatialized sound tends to attract gazes toward its loca-

A. Coutrot
Gipsa-lab - www.gipsa-lab.grenoble-inp.fr
E-mail: antoine.coutrot@gipsa-lab.grenoble-inp.fr

tion [4]. Onat and colleagues presented static natural images and spatially localized (left, right, up, down) simple sounds. They compared eye movements of observers when viewing visual only, auditory only or audio-visual stimuli. Results indicated that eye movements were spatially biased towards the regions of the scene corresponding to the sound sources [26]. Still, the combination of visual and auditory saliency models has rarely been investigated. Moreover, when used on videos, saliency models never take into account the information contained in the soundtrack. When running eye-tracking experiments with videos, authors do not mention soundtracks or explicitly remove them, making participants look at *silent movies* which is far from natural situations.

Our aim is to assess whether an auditory saliency model based on physical characteristics of the signal can be used to examine the impact of sound on observers' gaze while watching videos. Here, we do not focus on sound spatialization but simply on salient auditory events that might reinforce the saliency of visual events. In a previous study, we showed that soundtracks do have a global impact on visual exploration when watching videos [5]. In this study, we go further and examine whether this impact is stronger just after salient auditory events. For that purpose, we spotted salient auditory events in video soundtrack using two models. First, a popular auditory saliency model, the "Discrete Energy Separation Algorithm" (DESA). Second, a simple energy-based model. We analyzed the results of an eye - tracking experiment in which we recorded gazes of participants watching videos with and without their related soundtrack.

First, we present the results obtained in [5], where we tested the general impact of sound on eye movement parameters. We found that observers looking at videos with their soundtracks had different eye movements than observers looking at the same videos without sound. Second, we focus on sound impact on eye movements following auditory salient events spotted by a model. A founding rule of multisensory integration is the *temporal rule*: multisensory integration is more likely or stronger when the stimuli from different modalities are synchronous [29]. This rule has been established by comparing the electrical activity of some neurons when presenting simple visual stimuli (light flashes) with or without synchronous or delayed simple auditory stimuli (bursts). Studies showed that neurons activity was much stronger in multimodal than in unimodal condition, and that this reinforcement was maximal for synchronous stimuli [24] [25]. Here, we generalize this idea to more complex stimuli by identifying bursts to auditory saliency peaks and light flashes to the corresponding visual information. Thus, we compare the eye movements made over whole videos to those made over the few frames following auditory saliency peaks. It has been shown that audio and visual stimuli can be judged as synchronous across a broad range of physical offsets, typically in a 400 ms temporal window (see [27] for a review). This flexibility is probably due to the different propagation velocity between modalities in the environment (light: 300 000 km/s; sound: 0.34 km/s) and in the human body (conduction time from the retina to the brain: around 10 ms [11]; from the cochlea to the brain: around 50 ms [20]). Moreover, this window seems to be flexible with regard to input type. Complex stimuli are easier to integrate than simple ones, thanks to prior experience: one is more used to associate speech with moving lips or thunder with lightning than simple bursts with light flashes [12]. Thus, the temporal window during which a salient auditory event might significantly interact with visual information is around 400 ms but is not precisely determined. That is why in this research we chose to compare the eye movement parameters made over whole videos *vs.* the ones made over the 5 (200 ms), 10 (400 ms) and 25 (1s) frames following saliency peaks.

To summarize, the main goal of this study is to test whether the global effect of sound that was previously found on eye movements is reinforced just after salient audio events. The salient events are emphasized through two models: the DESA and the Energy models. We compared some eye movement parameters (the dispersion between eye positions, the mean saccade amplitude and the mean fixation duration) recorded on videos seen with and without their original soundtracks. The comparison was done over whole videos *vs.* over the few frames following salient audio events. To discuss our results, we ensured through an additional experiment that the salient audio events spotted by the models are effectively judged as more salient by listeners than random events.

2 Auditory saliency models

Attention, both in visual and auditory modalities, is mainly caught by features standing out from their background (*e.g.* motion, bright colors or high intensities). In a complex scene, the auditory system segregates sounds by extracting features such as spectral or temporal modulations [2]. In this section, we describe the two models used to spot auditory salient events in soundtracks. First, the Discrete Energy Separation Algorithm (DESA) is detailed. This algorithm has recently been brought forward in many fields of research involving the detection of auditory information, such as movie summarization or speech analysis [7] [8]. Second, we present a model merely based on the signal energy.

2.1 Discrete Energy Separation Algorithm

Even if our understanding of auditory saliency is still limited, previous studies had shown that extracting amplitude

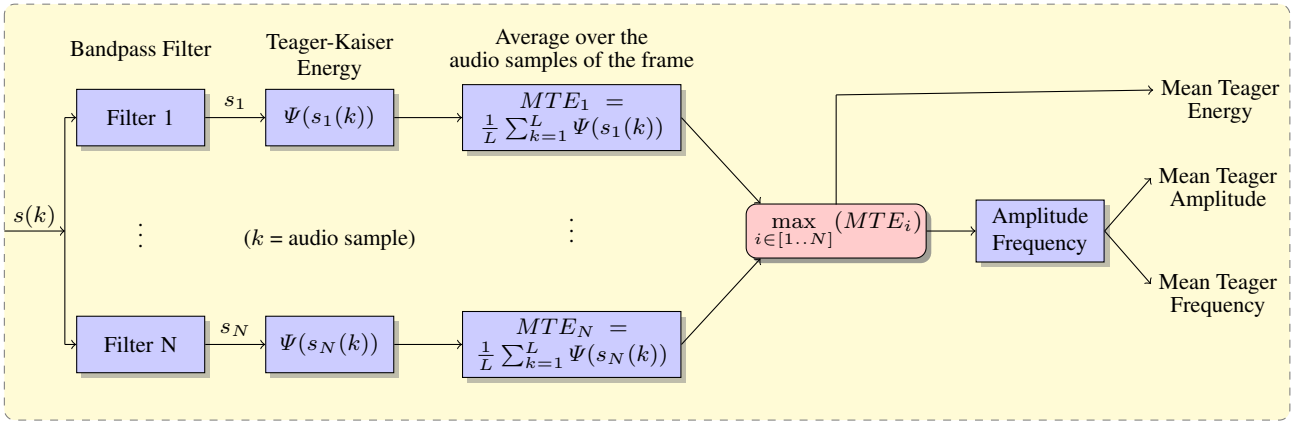

Fig. 1 Discrete Energy Separation Algorithm processing stages on an auditory signal s split into N frequency bands. Teager-Kaiser Energy, given by equation 1, is averaged over all audio samples k contained in a frame (there are L = 48000 x 0.04 = 1920 audio samples in a 40 ms audio frame sampled at 48 kHz). We chose the frequency band with the maximal Teager-Kaiser Energy (MTE). The Mean Teager Amplitude (MTA) and Frequency (MTF) are computed from the Teager-Kaiser Energy thank to equations 2 and 3. The three features are then combined to compute the auditory saliency value of the frame.

and frequency modulations is essential to predict the natural orienting behavior of humans to audio signals [9] [19]. The Discrete Energy Separation Algorithm (DESA) is an auditory saliency model based on the temporal modulation of amplitude and frequency in multiple frequency bands. The multiband demodulation analysis allows the captures of such modulations in the presence of noise, which is often a limiting factor when dealing with complex auditory scenes [1]. The DESA is simple and efficient. The process applied to each audio frame is described in Figure 1. The input signal s is separated in several frequency bands thank to Gabor filters. A Gabor filter is described in time as

$$h_i(t) = exp(-\alpha_i^2 t^2)cos(\omega_i t)$$

with (ω_i, α_i) respectively the central frequency and the filter bandwidth ($i \in [1..N]$ with N the total number of filters) [1]. Their placement and bandwidth have been chosen such that two neighboring filters intersect at half-peak.

$$\omega_i = \frac{3\Omega_c}{2^{i+1}}$$

$$\alpha_i = \frac{\omega_i}{2\sqrt{\ln 2}}$$

with Ω_c the highest frequency to be analyzed. Concretely, the video soundtracks were sampled at 48 kHz and separated in six frequency bands respectively centered on $\omega_i \in \{281, 562, 1125, 2250, 4500, 9000\}$ Hz. This spectrum covers a broad type of audible noises (*e.g.* speech: from 50 Hz to 8 kHz). Given an audio sample k, the Teager-Kaiser energy is computed for each frequency band:
Teager-Kaiser energy:

$$\Psi[s[k]] = s^2[k] - s[k+1]s[k-1] \quad (1)$$

The Teager-Kaiser energy is prized for its ease of implementation and its narrow temporal window, making it ideal for local (time) analysis of signals.The Teager-Kaiser energy is often used for detecting amplitude and frequency modulations in AM-FM signals [17] [31]. To separate the noise from the signal of interest, the frequency band in which the Teager-Kaiser energy is maximal is selected. In this frequency band, we separate the instantaneous energy into its amplitude and frequency components, according to the following equations.
Instant amplitude:

$$|a[s[k]]| = 2\frac{\Psi(s[k])}{\sqrt{\Psi(\dot{s}[k])}} \quad (2)$$

Instant frequency:

$$f[s[k]] = \frac{1}{2\pi}\arcsin\left(\sqrt{\frac{\Psi[\dot{s}[k]]}{4\Psi[s[k]]}}\right) \quad (3)$$

with $\dot{s}$ the derivative of the signal.
Each feature is averaged over a number of audio samples k corresponding to a frame duration (40 ms), to compute the mean Teager energy (MTE), the mean instant amplitude (MIA) and the mean instant frequency (MIF). The MTE, MIA and MIF are then normalized and combined to compute the auditory saliency value S of the current frame m. Here, we averaged the three features:

$$S(m) = w_1\text{MTE}(m) + w_2\text{MIA}(m) + w_3\text{MIF}(m)$$

with

$$w_1 = w_2 = w_3 = \frac{1}{3}$$

Since different weighing could lead to different results, one could adapt it according to the mean value of each feature. If the sound to be analyzed contains great energy variations (*e.g.* an argument with many raised voices), one is likely to

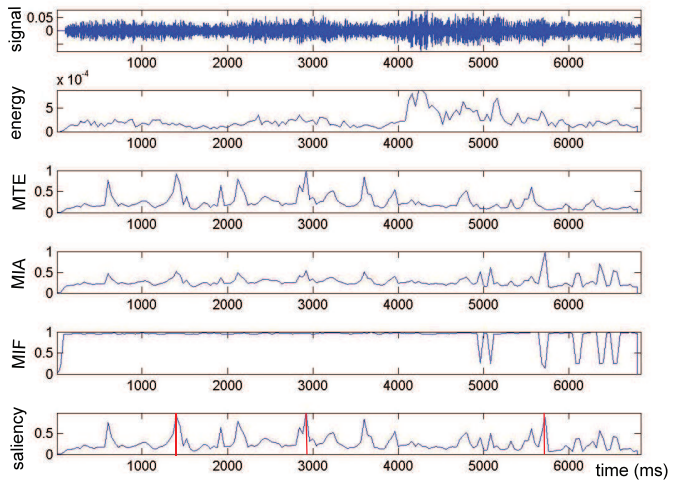

Fig. 2 Decomposition of a 6840 ms soundtrack (171 frames of 40 ms, upper plot) into energy (second plot), Mean Teager Energy (MTE, third plot), Mean Instant Amplitude (MIA, fourth plot) and Mean Instant Frequency (MIF, fifth plot). The combination of MTE, MIA and MIF gives the auditory saliency curve (lower plot) which is thresholded to spot the saliency peaks (vertical red bars).

give a preferential weighting to the MTE. On the contrary, if the sound contains great frequency variations (*e.g.* moving police siren presenting Doppler effect), the MIF will be preferred. Here, we chose an equally weighted combination to make the DESA as flexible as possible. Figure 2 illustrates the DESA algorithm applied to an audio signal (first plot). Three features were computed and averaged for each frame to provide the three curves MTE, MIA and MIF. Finally, the saliency curve was computed by averaging the three upper curves. Thresholding this auditory saliency curve gave the signal "saliency peaks" (vertical red bars). We normalized the number of saliency peaks over time. First, we chose a rate of one peak for two seconds: a N-second long signal had N/2 saliency peaks. Second, the time interval between two peaks had to be longer than one second: two neighboring peaks were distant enough so that the potential effect they might induce did not affect each other.

2.2 Energy model

We compared the peaks given by our auditory saliency model based on DESA algorithm, we also computed saliency auditory peaks using the energy curve (second plot of Figure 2), given by

$$E[s[k]] = s^2[k]$$

We extracted the "Energy peaks", i.e. the local maxima of the energy curve, at the same rate as saliency peaks (one peak for two seconds and at least one second between two peaks).

We used these two sets of peaks ("DESA peaks" and "Energy peaks") to evaluate the impact of sound on eye movements recorded during the eye-tracking experiment described below.

3 Methods

To observe the impact of salient auditory events on eye movements while freely watching videos, we set up an eye-tracking experiment. We built a base of 50 videos and asked 40 participants to watch it, half with and half without its related soundtracks. The experimental set-up and the data presented here were used in a previous paper [5].

3.1 Apparatus

Eye-movements were recorded using an Eyelink 1000 eye-tracker (SR Research). We used the eye-tracker in binocular "pupil - corneal reflect" tracking mode. Eye positions are sampled at 1 kHz with a nominal spatial resolution of 0.01 ° of visual angle. The device is controled by the software SoftEye [14] that allows to control stimuli presentation. Participants were sat 57 cm away from a 21 inch CRT monitor with a spatial resolution of 1024 x 768 pixels and a refresh rate of 75 Hz. Head was stabilized with a chin rest, forehead rest and headband. Soundtracks were listened using headphones (HD280 Pro, 64Ω, Sennheiser).

3.2 Participants and stimuli

Participants 40 persons participated at the experiment: 26 men and 14 women, aged from 20 to 29 years old. All participants had normal or corrected to normal vision and hearing and were French native speakers. They were naive about the aim of the experiment and were asked to watch videos freely. We discarded data from 4 participants due to recording problems.

Stimuli We used 50 video clips extracted from professional movies as varied as possible, to reflect the diversity of audiovisual scenes that one is likely to see and hear (dialogue, documentary film, drama, action movies). When the soundtrack contained speech, it was always in French. Each video sequence had a resolution of 720 x 576 pixels (30 ° x 24 ° of visual angle) and a frame rate of 25 frames per second. They lasted from 0.9 s to 35 s (*mean* = 8.7 s; *standard deviation* = 7.2 s). Video sequences lasted overall 23.1 min. We chose video shots with varied durations to avoid any habituation effect from the participants: one could not predict when each stimulus ended. As explained in the introduction, we chose

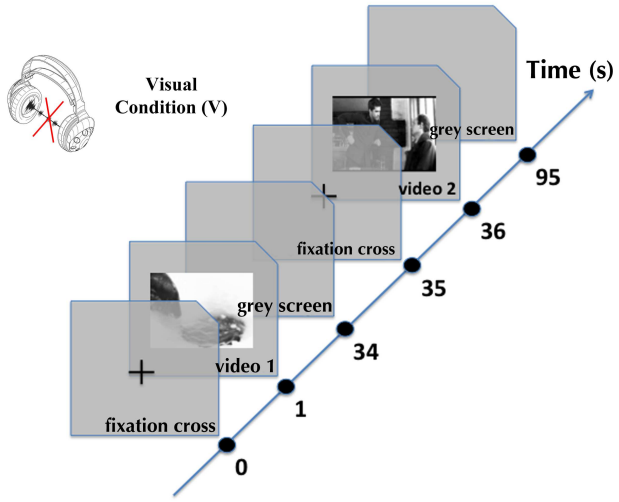

Fig. 3 Time course of two trials in the visual condition. A fixation cross is presented at the center of the screen, with gaze control. Then, a video sequence is presented in the center, followed by a grey screen. This sequence is repeated for the 50 videos, half without sound (Visual condition), the other half with their original soundtracks (Audio-Visual condition).

to focus on the influence of nonspatial sound on eye movements. Hence, we used monophonic soundtracks.
We analyzed the parameters of eye movements on average over each video shot rather than over entire videos. A shot cut is an abrupt transition from one scene to another that greatly impacts the way one explores videos [10] [28]. In a preliminary study [6], we elicited the effect of video editing (shots and cuts), studying the effect of sound over entire videos made up of several shots, and found no significant impact of sound on eye movements. However, this effect exists and has been observed when taking into account video editing, at least for shots longer than one second, which is the case of practically all the shots of our database [5]. Thus, in the present work, we did not study entire videos but we examined each shot. Shots were automatically detected using the mean pixel by pixel correlation value between two adjacent video frames. We ensured that the shot cuts detected were visually correct. Sequences contained different number of shots, with a total number of 163 shots.

3.3 Procedure

The experiment consisted in freely viewing 50 video sequences. The first 20 participants saw the first half of videos with their soundtracks and the other half without. This was counterbalanced for the last 20 participants. Each experiment was preceded by a calibration procedure, during which participants focused their gaze on 9 separate targets in a 3 x 3 grid that occupied the entire display. A drift correction was done between each video, and a new calibration was done at the middle of the experiment or if the drift error was above 0.5 $^\circ$. Before each video sequence, a fixation cross was displayed in the center of the screen for 1 second. After that time, and only if the participant looked at the center of the screen (gaze contingent display), the video sequence was played on a mean grey level background. Between two consecutive video sequences a grey screen was displayed for 1 second (see Figure 3). Participants wore headphones during the entire experiment, even when the stimuli were presented without soundtrack. To avoid presentation order effects, videos were run randomly. At the end, each video was seen by 20 persons with its related soundtrack and by other 20 persons without its related soundtrack.

3.4 Data extraction

The eye tracker system gives one eye position each millisecond, but since the frame rate is 25 frames per second, 40 eye positions per frame and per participant were recorded. Moreover, we only analyzed the guiding eye of each subject. In the following, an *eye position* is the median position that corresponds to the coordinates of the 40 raw eye positions: there is one eye position per frame and per subject. Eye positions corresponding to a frame during which participants made a saccade or a blink were discarded from analysis. For each frame and each stimulus condition, we discarded outliers, i.e. eye positions above $\pm$ 2 standard deviations from the mean. The eye-tracker software organizes the recorded movements in events: saccades, fixations and blinks.
Saccades are automatically detected by the Eyelink software using three thresholds: velocity (30 degrees/s), acceleration (8000 degrees/s^2) and saccadic motion (0.15 degree). The velocity threshold is the eye movement velocity that must be exceeded for a saccade to be detected. Acceleration threshold is used to detect small saccades. The saccadic motion threshold is used to delay the onset of a saccade until the eye has moved significantly.
Fixations are detected as long as the pupil is visible and as long as there is no saccade in progress.
For each stimulus condition, we discarded outliers, i.e. saccades (resp. fixations) whose amplitude (resp. duration) was above $\pm$ 2 standard deviations from the mean. Moreover, we discarded data from four subjects due to recording problems. We separated the recorded eye movements in two data sets.

- The data recorded in the Audio-Visual (AV) condition, i.e. when videos were seen with their original soundtrack.
- The data recorded in the Visual (V) condition, i.e. when videos were seen without any sound.

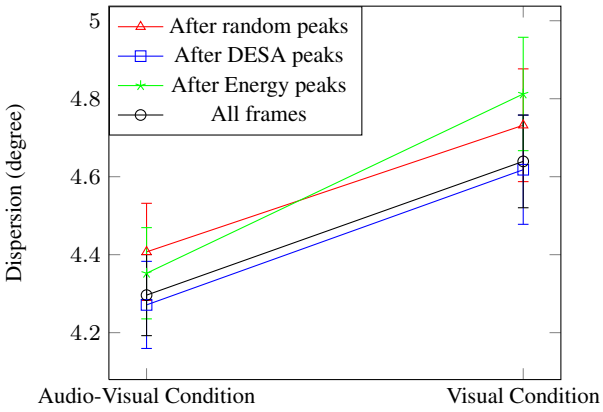

Fig. 4 Mean dispersion in Audio-Visual an Visual conditions. Dispersion values are averaged over all frames (blue) and over the 10 frames following each random (black), energy (green) and DESA saliency (red) peaks. Dispersions are given in visual angle (degrees) with error bars corresponding to the standard errors.

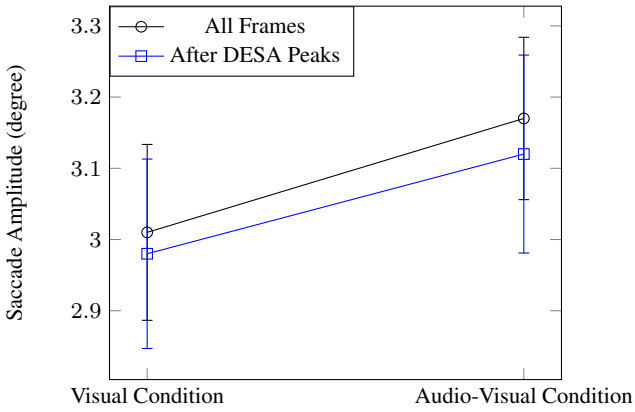

Fig. 5 Mean saccade amplitude in Audio-Visual an Visual conditions. Saccade amplitudes are averaged over all frames (black) and over the 10 frames following each DESA (blue) peak. Saccades amplitude are given in visual angle (degrees) with error bars corresponding to the standard errors.

4 Results

In this section, we examine the eye movements recorded in Visual (V) and Audio-Visual (AV) conditions. We compare eye movements averaged over all the frames of a same shot and over the few frames following each "Energy peak" or each "DESA peak". The presented results are based on a 10-frame time period after Energy and DESA peaks. The same analysis was carried out on a 5 and 25-frame period and led to the same results (see Section 4.1), that we chose not to plot. We discarded from analyzes shots without DESA (resp. Energy) peaks (36 shots). We examine several parameters in both V and AV conditions: the dispersion between the eye positions of different observers, which reflects the variability between them; the amplitude of the recorded saccades and the duration of the recorded fixations.

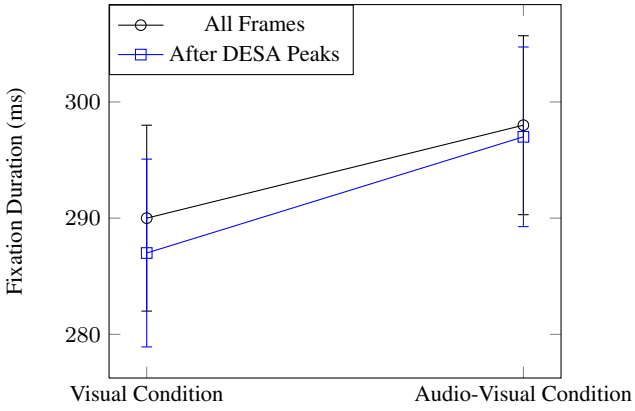

Fig. 6 Mean fixation duration in Audio-Visual an Visual conditions. Fixation durations are averaged over all frames (black) and over the 10 frames following each DESA (blue) peak. Fixation durations are given in milliseconds with error bars corresponding to the standard errors.

4.1 Eye position dispersion

To estimate the variability of eye positions between observers, we used a measure called *dispersion*. For a frame and for n participants (thus n eye positions $\mathbf{p} = (x_i, y_i)_{i\in[1..n]}$), the dispersion D is defined as follow:

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

In other words, the dispersion is the mean of the Euclidian distances between the eye positions of different observers for a given frame. If all participants look close to the same location, the dispersion value is small. On the contrary, if eye positions are scattered, the dispersion value increases.
In this analysis, we computed a dispersion value for each frame, in both V and AV conditions. First, we averaged dispersion over all frames of each shot. Then, we averaged dispersion over the 10 frames following each DESA and Energy peak. To control, we computed 1000 random sets containing "random peaks" at the same rate as Energy and DESA peaks. For each random set, we averaged dispersion over the 10 frames following each "random peaks" and took the mean of these 1000 "random" dispersion values. Results are shown in Figure 4. We first notice that in all cases, dispersion is smaller in Audio-Visual than in Visual Condition. To test the impact of sound on the dispersion values of the 127 video shots containing a DESA or Energy peak, we ran two analyzes of variance (ANOVAs). The first ANOVA was run with two factors: the stimulus condition (visual and audio-visual) and the window size (all frames and 10 frames) used to average dispersion after the Energy peaks. The second ANOVA was also run with two factors: the stimulus condition (visual and audio-visual) and the window size (all frames and 10 frames) used to average dispersion after the DESA peaks. For the first one, we found a main effect of

sound (F(1,126) = 28.2; $p < 0.001$) and no effect of Energy peaks (F(1,126) = 0.82; n.s). We found similar results for the second one: a main effect of sound (F(1,126) = 24.8; $p < 0.001$) and no effect of DESA saliency peaks (F(1,126) = 1.52; n.s).

To test the impact of the size of the averaging time window on the dispersion values of the 127 video shots containing a DESA or Energy peak, we ran two ANOVAs. The first ANOVA was run with two factors: the stimulus condition (visual and audio-visual) and the window size (averaged over 5, 10 and 25 frames after DESA peaks). We found a main effect of sound (F(1,126) = 31.8; $p < 0.001$) and no effect of the size of the averaging time window (F(2,124) = 0.9; n.s). The second ANOVA was also run with two factors: the stimulus condition (visual and audio-visual) and the window size (averaged over 5, 10 and 25 frames after Energy peaks). We also found a main effect of sound (F(1,126) = 39.6; $p < 0.001$) and no effect of the size of the averaging time window (F(2,124) = 0.2; n.s).

To sum up, we found that the presence of sound does impact on gaze dispersion but neither the proximity of DESA or Energy peaks nor the size of the averaging time window affects this parameter.

4.2 Saccade amplitude and fixation duration

Figure 5 and Figure 6 respectively compare the average saccade amplitude and fixation duration in Audio-Visual and Visual conditions over all frames and over the ten frames following each DESA peak. For the sake of clarity, we omitted the saccade amplitude and fixation duration made after Energy and random peaks. The results are similar as the plotted ones. Distributions follow a positively skewed, long-tailed distribution, which is classical when studying such parameters during scene exploration [13] [30]. We notice that participants tended to make smaller saccades and shorter fixations in V than in AV condition.

We performed two analyzes of variance (ANOVA) with two factors (visual and audio-visual ; all frames and 10 frames after DESA peaks) on the 36 participant's median saccade amplitude and median fixation duration. For saccade amplitude, it revealed a main effect of sound (F(1,35) = 4.9; p = 0.033) and no effect of DESA peaks (F(1,35) = 0.27; n.s). For fixation duration, there was still no effect of DESA peaks (F(1,35) = 0.01; n.s) and the effect of sound was not significant (F(1,35) = 2.1; p = 0.15).

5 General discussion

We compared eye positions and eye movements of participants freely looking at videos with their original soundtracks (AV condition) and without sound (V condition). In a previous study, we showed that soundtrack globally impacts on eye movements during video viewing [5]. We showed that in AV condition, the eye positions of participants were less dispersed and tended to shift more from the screen center, with larger saccades. We also showed that observers did not look at the same locations, according to the viewing condition. An interpretation of these results is that sound might strengthen visual saliency. Indeed, with sound, observers explored videos in a more uniform way, leading to a decrease of the dispersion between eye positions. This interpretation is supported by the results on saccade amplitude. Participants made shorter saccades in V condition, fluttering from one position to another. On the opposite, in AV condition, sound could have helped guiding participants' gaze, leading to larger, goal-directed saccade amplitudes.

In this study, we compared this impact of sound averaged over entire videos with the one averaged over the 5, 10 and 25 frames following the salient events of video soundtracks. To spot these salient events, we used two auditory saliency models: the Discrete Energy Separation Algorithm and the Energy model.

We found that the impact of sound on saccade amplitudes and on the dispersion between eye positions was very similar after DESA peaks, Energy peaks or in general over entire videos. This indicates that the temporal proximity of auditory events spotted by neither the DESA, nor the energy model, increases the effect of sound on eye movements. Moreover, the size of the temporal window over which we averaged eye movement parameters following salient events (5, 10 or 25 frames) did not affect the results. A reason for these results can be that the signal features extracted by the DESA (mean instant Teager-Kaiser energy, amplitude and frequency) might not satisfactorily reflect the way our brain processes auditory information to generate attention. Future studies should investigate more complete auditory saliency models, which use more sophisticated auditory features. For instance, the auditory features used in musical classification (mel-frequency cepstral coefficients (MFCC), delta-MFCC, chromagrams or zero-crossing rates [34]) could be used to successfully integrate sound to visual saliency models. Moreover, one may question the way these features are combined (here, linearly). For a given input, we could estimate the energy of each feature and adjust their weight accordingly. For instance, we could give the MIF a bigger weight when this feature is dominant, like in a whistling sound.

Nevertheless, this explanation is not entirely satisfactory. Indeed, although simple, the models we used in this study spotted auditory events that had a perceptual relevance, *i.e.* that actually were judged as salient by human listeners. To ensure this, we ran a control experiment.

Stimuli We chose 14 soundtracks (from 9 to 57 seconds)

from the dataset of soundtracks associated to the videos used in the main experiment. These soundtracks were randomly chosen, trying to have exemplar of sounds with music, moving objects, voices, etc. Audio stimuli had the same characteristics as described Section 3.2. For each soundtrack, we computed three sets of peaks: one set using the DESA model, one set using the Energy model and a last set using random peaks. For a given soundtrack, the three sets of peaks had the same number of peaks at least separated by the same minimum interval (one second), like in the main experiment.

Participants and set-up We asked 5 persons to listen to each soundtrack (without video) and to judge whether or not each peak corresponded to a salient audio event. Participants were seated in front of a computer screen with the headphones used for the main experiment. For each soundtrack, three curves with the three sets of peaks (DESA, Energy and Random) were displayed and a blue vertical line marked the progression of the sound. Participant had to tag as salient or not salient 104 peaks x 3 models = 312 peaks. They could replay the audio signals as many time as needed. In average, each participant spent forty-five minutes to fulfill the task.

Results For each soundtrack, we obtained a percentage of the peaks identified as salient by the participants for the DESA, Energy and Random models. We ran paired t-test to test which model had the highest percentage of peaks judged as salient. Both DESA and Energy models had a higher percentage than the Random model, with 43% *vs.* 20% for the DESA model (t(13)=5,46; $p < .0001$), and 60% *vs.* 20% for the Energy model (t(13)=5,28; $p < .0001$). The Energy model had a higher percentage than the DESA (t(13)=-2,8; $p < .01$). As it could be expected, the peaks spotted by both models were more relevant than the random ones. Moreover, we could notice that when the audio signal contained events that clearly stood out from the background (like speech or noise from a moving object), both DESA and Energy models computed relevant saliency peaks (*i.e.* events that were judged as salient by a majority of observers). On the contrary, when the input did not contain any particular event (*e.g.* smooth music, wind blowing), the performance dropped sharply: the peaks emphasized by the models were judged as salient only in few cases. While the relevance of the DESA and Energy models compared to the Random model was expected, we did not expect the better results of the Energy model. However, one should not draw hasty conclusions. The Energy model emphasizes the most evident changes in the audio signal, which are not necessarily the ones that actually draw attention. For instance, a little voice in a noisy environment will not be tagged as salient by the Energy model, although it obviously attracts the auditory attention. Conversely, the DESA model is likely to spot this voice, thanks to its MIF feature. To tag the peaks as salient or not, participants had to listen to each soundtrack several times (at least one time per set of peaks). After two or three listenings, only the most evident changes remain salient, which can explain the preference of the listeners for the Energy peaks.

Altogether, both DESA and Energy peaks were globally much more relevant than random ones, which legitimizes the presented analyses.

6 Conclusions

We have shown that while sound has a global impact on eye movements, this effect is not reinforced just after salient auditory events. This result can be explained if we consider that auditory saliency may entail much more complex and observer-dependent information. The emotions aroused by the soundtrack (*e.g.* music) or the information contained in it (*e.g.* speech) can drastically affect our attention on much larger time scale than 5, 10 or even 25 frames. In that case, the temporal proximity of salient events would not be a relevant parameter: the sound would impact on visual exploration in a global way. Altogether, these results indicate that if non-spatial auditory information does impact on eye movements, the exact auditory features capturing observers' attention remain unclear. To successfully integrate sound into visual saliency models, one should investigate the influence of a specific sound on a specific visual feature, and take into account context-sensitive information.

References

1. Bovik, A.C., Maragos, P., Quatieri, T.F.: AM-FM Energy Detection and Separation in Noise Using Multiband Energy Operators. IEEE Transactions on Signal Processing **41**(12), 3245–3265 (1993)
2. Bregman, A.S.: Auditory Scene Analysis, the perceptual organization of sound. MIT Press, Cambridge, MA, US (1990)
3. Cater, K., Chalmers, A., Ward, G.: Detail to attention: exploiting visual tasks for selective rendering. In: Eurographics Symposium on Rendering, pp. 270–280 (2003)
4. Corneil, B.D., Munoz, D.P.: The influence of auditory and visual distractors on human orienting gaze shifts. The Journal of neuroscience **16**(24), 8193–8207 (1996)
5. Coutrot, A., Guyader, N., Ionescu, G., Caplier, A.: Influence of soundtrack on eye movements during video exploration. Journal of Eye Movement Research **5**(4), 1–10 (2012)
6. Coutrot, A., Ionescu, G., Guyader, N., Rivet, B.: Audio Tracks Do Not Influence Eye Movements When Watching Videos. In: 34th European Conference on Visual Perception (ECVP 2011), p. 137. Toulouse, France (2011)
7. Evangelopoulos, G., Maragos, P.: Multiband Modulation Energy Tracking for Noisy Speech Detection. IEEE Transactions on Audio, Speech and Language Processing **14**(6), 2024–2038 (2006)
8. Evangelopoulos, G., Zlatintsi, A., Skoumas, G., Rapantzikos, K., Potamianos, A., Maragos, P., Avrithis, Y.: Video event detection and summarization using audio, visual and text saliency. In: Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP-09), pp. 3553–3556. Taipei (Taiwan) (2009)

9. Fritz, J.B., Elhilali, M., David, S.V.: Auditory attention — focusing the searchlight on sound. Current Opinion in Neurobiology **17**, 1–19 (2007)
10. Garsoffky, B., Huff, M., Schwan, S.: Changing viewpoints during dynamic events. Perception **36**(3), 366–374 (2007)
11. Gouras, P.: The effects of light-adaptation on rod and cone receptive field organization of monkey ganglion cells. The Journal of Physiology **192**(3), 747–760 (1967)
12. Guski, R., Troje, N.F.: Audiovisual phenomenal causality. Perception & Psychophysics **65**(5), 789–800 (2003)
13. Ho-Phuoc, T., Guyader, N., Landragin, F., Guérin-Dugué, A.: When viewing natural scenes, do abnormal colours impact on spatial or temporal parameters of eye movements ? Journal of Vision **12**(2), 1–13 (2012)
14. Ionescu, G., Guyader, N., Guérin-Dugué, A.: SoftEye software (IDDN.FR.001.200017.000.S.P.2010.003.31235) (2009)
15. Itti, L.: Automatic Foveation for Video Compression Using a Neurobiological Model of Visual Attention. IEEE Transactions on Image Processing **13**(10), 1304–1318 (2004)
16. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11), 1254–1259 (1998)
17. Kaiser, J.: On a simple algorithm to calculate the "energy" of a signal. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP-90, vol. 1, pp. 381–384. Albuquerque, NM, USA (1990)
18. Kalinli, O., Narayanan, S.: A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. In: Eighth Annual Conference of the International Speech Communication Association, pp. 1941–1944. Antwerp, Belgium (2007)
19. Kayser, C., Petkov, C.I., Lippert, M., Logothetis, N.K.: Mechanisms for allocating auditory attention: an auditory saliency map. Current Biology **15**, 1943–1947 (2005)
20. Kraus, N., McGee, T.: Electrophysiology of the human auditory system. In: A. Popper, R. Fay (eds.) The Mammalian Auditory Pathway: Neurophysiology, pp. 335–403. Springer, New York (1992)
21. Li, Z., Qin, S., Itti, L.: Visual attention guided bit allocation in video compression. Image and Vision Computing **29**, 1–14 (2011)
22. Marat, S., Ho-Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., Guérin-Dugué, A.: Modelling Spatio-Temporal Saliency to Predict Gaze Direction for Short Videos. International Journal of Computer Vision **82**(3), 231–243 (2009)
23. McGurk, H., MacDonald, J.: Hearing lips and seing voices. Nature **264**, 746–748 (1976)
24. Meredith, M.A., Nemitz, J.W., Stein, B.E.: Determinants of Multisensory Integration in Superior Colliculus Neurons. I. Temporal Factors. The Journal of neuroscience **7**(10), 3215–3229 (1987)
25. Meredith, M.A., Stein, B.E.: Spatial factors determine the activity of muitisensory neurons in cat superior colliculus. Brain Research **365**, 350–354 (1986)
26. Onat, S., Libertus, K., König, P.: Integrating audiovisual information for the control of overt attention. Journal of Vision **7**(10), 1–16 (2007)
27. Recanzone, G.H.: Interactions of auditory and visual stimuli in space and time. Hearing Research **258**(1-2), 89–99 (2009)
28. Smith, T.J., Levin, D., Cutting, J.E.: A Window on Reality: Perceiving Edited Moving Images. Current Directions in Psychological Science **21**(2), 107–113 (2012)
29. Stein, B., Meredith, M.: The merging of the senses, mit press edn. Cambridge,MA (1993)
30. Tatler, B.W., Baddeley, R.J., Vincent, B.T.: The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. Vision Research **46**, 1857–1862 (2006)
31. Teager, H.M.: Some observations on oral air flow during phonation. IEEE Transactions on Acoustics, Speech and Signal Processing **28**(5), 599–601 (1980)
32. Tingle, D., Kim, Y.E., Turnbull, D.: Exploring automatic music annotation with "acoustically-objective" tags. In: Proceedings of the international conference on Multimedia information retrieval, MIR '10, pp. 55–62. ACM, New York, NY, USA (2010)
33. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive psychology **12**, 97–136 (1980)
34. Tzanetakis, G., Cook, P.: Musical Genre Classification of Audio Signals. IEEE Transactions on Speech and Audio Processing **10**(5), 293–302 (2002)
35. Vroomen, J., de Gelder, B.: Sound Enhances Visual Perception: Cross-modal Effects of Auditory Organization on Vision. Journal of Experimental Psychology **26**(5), 1583–1590 (2000)