How different kinds of sound in videos can influence gaze

Guanghan Song, Denis Pellerin, Lionel Granjon

HOW DIFFERENT KINDS OF SOUND IN VIDEOS CAN INFLUENCE GAZE

Guanghan Song, Denis Pellerin, Lionel Granjon

Department Images and Signal, GIPSA-Lab
BP 46, 38402 Grenoble Cedex, France

ABSTRACT

This paper presents an analysis of the effect of thirteen different kinds of sound on visual gaze when looking freely at videos to help to predict eye positions. First, an audio-visual experiment was designed with two groups of participants, with audio-visual (AV) and visual (V) conditions, to test the sound effect. Then, an audio experiment was designed to validate the classification of sound we proposed. We observed that the sound effect is different depending on the kind of sound, and that the classes with human voice (speech, singer, human noise and singers) have the greatest effect. Finally, a comparison of eye positions with a visual saliency model was carried out, which proves that adding sound to video decreases the accuracy of prediction of the visual saliency model.

1. INTRODUCTION

Traditional research on attention often considers a single sensory modality (aural or visual) at a time. However, in the real world, normally our attention has to be coordinated cross modally, and we select information from a common external source across several modalities. Early application of audio-visual (AV) saliency model is focused on low-level fusion (at the extracted saliency curves) [1]. This low-level fusion could not contain information from the saliency region. C. Quigley et al. [2] proposed AV integration research during overt attention, including spatial information. Recently, cross-modal interaction of auditory and visual modalities has played an important role in human spatial saliency and for video coding [3]. To extend, some other researchers only consider the common content of audio and video speech signals to detect the active speaker among different candidates [4].

Our previous research [5] showed that sound affects human gaze differently depending on the sound type, and the effect is bigger for the on-screen speech class (the speakers appear on screen) rather than non-speech class (any kind of audio signal other than speech) and non-sound class (intensity below 40 dB). Compared to a visual attention model developed in our laboratory, the accuracy of prediction decreases in the group of participants with AV condition compared to the group with V condition under the on-screen speech class.

In our previous research, we only considered three sound classes. To extend, in this paper we study the effect of thirteen sound classes on visual gaze. Hence, we first describe in Section 2 an audio-visual experiment of two groups of participants with audio-visual (AV) and visual (V) conditions. In Section 3, an audio experiment to validate the classification we have proposed, is presented. We analyse the effect of sound on eye positions in Section 4. In Section 5, a comparison of the eye positions with a visual saliency model is presented.

2. AUDIO-VISUAL EXPERIMENT

Based on previous research [5], sound classes seem to affect human gaze differently. The purpose of this new experiment is to compare the behavior of human gaze of more refined sound classes, and to analyse the effect on eye movement of the advent of a new sound during a clip snippet (video excerpt lasting 6 to 10 seconds was called a clip snippet).

In this experiment, the video database was chosen from films which were relevant and interesting for both visual and audio content. In the visual domain, the database contains various content, including objects, events, characters, sports and so on. In the audio domain, during one clip snippet, the first sound lasts to about the middle of the clip snippet. Then, the sound switches to the second sound, which is unrelated to the first sound. Here, we only observe the behavior of human gaze after the second sound so as not to be disturbed by the simultaneous audio and visual changes when a new clip snippet starts.

The participants viewed the videos without any task. We put small parts of videos from different sources together with unrelated semantic contents. In order to prevent the participants from understanding the language in the video, we chose foreign languages for the participants, like Chinese, Irish, Japanese etc.

2.1. Stimuli

Eighty clip snippets were selected from heterogeneous sources for a total of 16402 frames (around 11 minutes). Each clip

This research is supported in part by the Rhône-Alpes region (France) with LIMA project. The authors would like to thank their colleagues G. Ionescu, A. Rahman and R. Drouilhet.

snippet was converted to the same video format (25 fps, 842× 474 pixels/frame). The 80 clip snippets were then recombined into 10 clips, each clip being the concatenation of 8 clip snippets from different sources and different sound classes of second sound. We used gray level stimuli. Two sets of stimuli were built from these clips, one with AV condition (frames + audio signal), and one with V condition (frames only).

2.2. Participants

Thirty-six human participants (18 women and 18 men, aged from 20 to 34 years old) viewed half clips with V condition, and the other half clips with AV condition. 18 participants first viewed 5 clips with V condition, then viewed another 5 clips with AV condition. The other 18 participants, first viewed 5 clips with AV condition, then viewed other 5 clips with V condition. Each clip appeared with AV and V condition in the same frequency. All participants had normal or corrected to normal vision, and reported normal hearing. They were ignorant to the purpose of the experiment.

2.3. Apparatus and experiment design

Human eye position was tracked by an Eyetracker Eyelink II (SR Research). During the experiment, the participants were sitting in front of a 19-inch color monitor (60 Hz refresh rate) with their chin supported. The viewing distance between the participant and the monitor was 57 cm. The usable field of view was $35° \times 20°$. A headphone carried the stereo sound. A 9-point calibration was carried out every five clips. 10 clips are presented to each participant in random order. Before each clip, we presented a drift correction, then a fixation in the center of the screen. Participants were asked to look at the 10 clips without any particular task.

2.4. Human eye position density maps

The eye-tracker records eye positions at 250 Hz. We recorded ten monocular eye positions per frame and per participant. A 2-D Gaussian was added to each position. The standard deviation of the Gaussian was chosen to match the fovea field of view ($0.5°$). For each frame k, we obtained a human eye position density map noted $M_h(x, y, k)$, where (x, y) are spatial coordinates of eye position.

3. SOUND CLASSIFICATION EXPERIMENT

In this paper, because we focus on the effect of the second sound, we only consider the sound classification of second sound.

3.1. Sound classification

Referred to M. E. Niessen et al. [6], we classify the second sound to thirteen classes (Fig.1). The difference between clusters of classes "on-screen with one sound source" and "on-screen with more than one sound source" is the number of sound sources on the screen. Here, we call one sound source a visual location in the scene associated to an event in the sound track. In this instance the sound can be associated with a spatial location. The "out-screen sound source" group is different from the other two in that there is no sound source on the screen when the second sound appears.

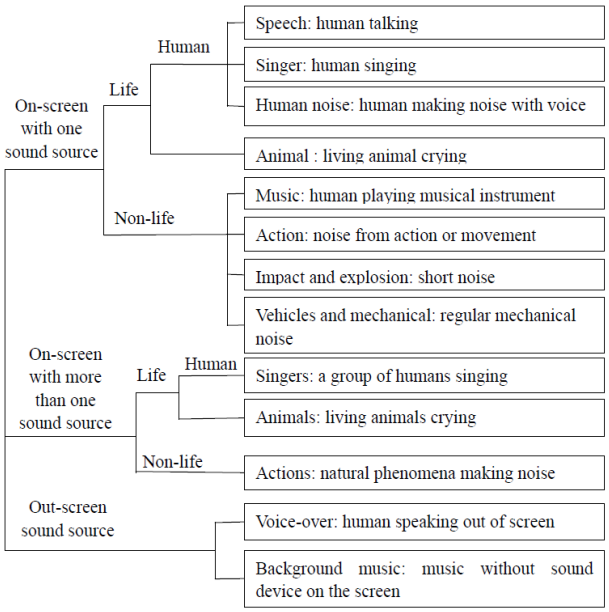

Fig. 1. Classification of the second sound

3.2. Validation of sound base

Second sound of all the clip snippets are classified to thirteen classes in Fig. 1. For each class, there are 5 to 13 samples. In order to validate this classification, we carried out an audio experiment. We presented audio excerpts of the second sound to participants through headphones for a classification task. Only the second sound of each clip snippet is presented, in random order. Because of sound fade, the audio excerpts begin 200ms before the second sound appears, the total duration of each stimulus is more than 800ms. 5 participants carried out this experiment. Because it is hard to tell the difference between "speech" and "voice-over" without visual information, we gathered them together. The same process has done for "music" and "background music" classes. After each cutting second sound presented, the participants were asked to classify the sound presented among the eleven classes. .

If the participant chose the same class as we proposed, we consider it is a correct recognition. The minimal correct classification rate is 80%, and the mean correct classification rate is 90%. We conclude the classification is convenient for audio-visual experiment.

4. EYE POSITION ANALYSIS

In order to investigate the effect of sound on visual gaze, we considered the eye positions of participants with AV condition compared with participants with V condition. A comparison of the different eye positions between the two groups is presented.

Fig. 2. An example of the eye positions of two groups of participants (with AV condition –red points, with V condition –green points) of speech class.

4.1. Criterion

In order to measure the distance of eye positions between the two groups for each frame, we adopted the criterion named median distance (md). It is defined as:

$$md = median(d_{i,j}), i \in \mathcal{N}, j \in \mathcal{N}^{'} \quad (1)$$

where $\mathcal{N}$ is the group with AV condition, $\mathcal{N}^{'}$ is the group with V condition, $d_{i,j}$ is the Euclidean distance between eye positions of participants i and j.

4.2. Comparison among three clusters of sound classes

First, in order to investigate the influence of sound source, we analysed the median distance (md) between groups with AV and V conditions, among three clusters of classes: "on-screen with one sound source", "on-screen with more than one sound source" and "out-screen sound source", with Kruskal-Wallis test. In this section, for each clip snippet, we took 25 frames (from frame 6 to 30, to eliminate the reaction time of about 5 frames) after the beginning of the second sound. Because we consider continuous measurement along time, the fixation position for most participants does not change between two adjacent frames. Hence, we assume a set (8 frames, which is bigger than the average value of one fixation duration) of continuous frames as one independent sample. We compute an md for each frame, and subsample by computing the mean of 8 adjacent frames as an independent sample in Kruskal-Wallis test. In Fig. 3, "out-screen sound source" presents the lowest md among the three clusters of classes. The difference is significant, between "on-screen with one sound source" and "out-screen sound source" ($\chi^2(1) = 14.44$, $p < 10^{-3}$), and also significantly between "on-screen with more than one sound source" and "out-screen sound source"($\chi^2(1) = 16.92$, $p < 10^{-4}$). The md of "out-screen sound source" gets the lowest value among these three clusters of classes with median distance criterion, suggesting the lowest difference between the groups with AV and V conditions.

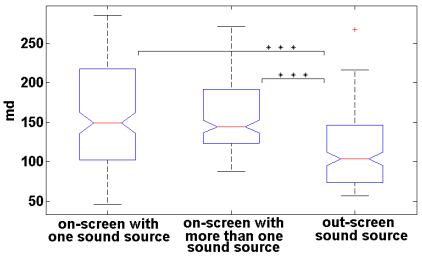

Fig. 3. Median distance (md) between participants with AV and V conditions in three clusters of classes ("on-screen with one sound source", "on-screen with more than one sound source" and "out-screen sound source").

4.3. Analysis of thirteen sound classes separately

We did not analyse sound effect directly through audio information, but through the eye positions of participants which are also based on visual information. In order to reduce the influence of visual information, we created a baseline for the statistical comparison by performing a randomization [7]: extract 18 participants randomly from groups with AV and V conditions to create a new group called G1. The rest of the participants make up another new group, called G2. After that, we calculated the md between G1 and G2 for each frame. We repeated this procedure 5000 times, obtaining for each frame a distribution of 5000 random md values. Then, we took the mean of the 5000 md values as the reference (md_R). Finally, we calculated the difference ($md_{AVV} - md_R$) where md_{AVV} represents the median distance between participants with AV and V conditions.

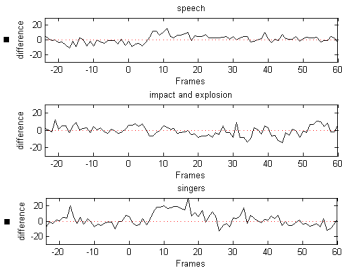

Fig. 4. Example of difference ($md_{AVV} - md_R$) along time for sound classes of speech, impact and explosion, and singers. Frame zero corresponds to the beginning of the second sound. Classes marked with ■ are significantly different between $\overline{md}_{AVV}$ and $\overline{md}_{R_i}$

In Fig. 4, different sound classes behave differently. To validate whether the difference between md_{AVV} and md_R is significant, we took a duration of 1s (25 frames) from frame 6 to 30 after the beginning of the second sound. We compared

$\overline{md}_{AVV}$ (the mean of md_{AVV} over the 25 frames) to the distribution of $\overline{md}_{R_i}$ ($i = 1, 2, ...5000$), where $\overline{md}_{R_i}$ is the mean of md between G1 and G2 over the 25 frames for the random trial i. To estimate the probability of $\overline{md}_{R_i}$ being bigger than $\overline{md}_{AVV}$, we calculated $p = n/5000$ where n is the number of $\overline{md}_{R_i}$ which is bigger than $\overline{md}_{AVV}$. In Table.1, $\overline{md}_{AVV}$

Table 1. Probability of $\overline{md}_{R_i}$ being bigger than $\overline{md}_{AVV}$ from frame 6 to 30 after the beginning of the second sound

sound class	p	sound class	p
speech ■	0.011	human noise ■	0.015
singer ■	0.013	impact and explosion	0.993
animal	0.698	vehicles and mechanicals	0.127
music	0.063	actions	0.827
action	0.744	voice-over	0.982
singers ■	0.006	background music	0.309
animals	0.682		

is significantly bigger than $\overline{md}_{R_i}$ ($p < 0.02$), from frame 6 to 30 after the beginning of the second sound, for speech, singer, human noise, and singers classes (marked with ■) suggesting that human voice has the greatest effect.

5. COMPARISON WITH A VISUAL SALIENCY MODEL

In order to test the accuracy of prediction of a visual saliency model for videos with sound, we compare eye positions of both groups of participants with saliency maps given by a well-known model proposed by L. Itti and C. Koch [8].

5.1. Criteria

For the evaluation, we chose the metric Normalized Scanpath Saliency (NSS), which was especially designed to compare eye fixations with the salient areas emphasized by the saliency model [5]. It is calculated by averaging the pixels that correspond to the eye positions. The NSS metric, which corresponds to Z-score, is computed as follows:

$$NSS(k) = \frac{\overline{M_h(x,y,k) \times M_m(x,y,k)} - \overline{M_m(x,y,k)}}{\sigma_{M_m(x,y,k)}} \quad (2)$$

where, $M_h(x, y, k)$ is the human eye position density map standardized to mean 0 and variance 1, and $M_m(x, y, k)$ is the model saliency map.

5.2. Results

With the purpose of testing the prediction accuracy of dynamic pathway of the model, we calculated the NSS for each clip snippet from the onset of the second sound for both the V (NSS_V) and AV (NSS_{AV}) conditions. We then calculated $\overline{NSS}_V$ (respectively $\overline{NSS}_{AV}$), which is the mean of NSS_V (respectively NSS_{AV}) for all the clip snippets. Finally, we considered the NSS difference ($\overline{NSS}_V - \overline{NSS}_{AV}$) between groups of participants with AV and V conditions.

We observed that the NSS difference after the onset of second sound first raises above 0, then decreases after a while. In order to validate whether the difference is significant, we choose the temporal window from frame 6 to 56. We took the mean of NSS difference of 8 continuous frames as one independent sample. The mean of NSS difference is significantly different from 0 (with t-test $p = 0.016$).

From the results, we conclude that the accuracy of prediction from dynamic pathway of the model decreases in a group with AV condition compared to group with V condition during frame 6 to 56 after the second sound appears.

6. CONCLUSION AND PERSPECTIVES

This study presents the analysis of sound effect on human gaze when looking freely at videos. From our comparison between $\overline{md}_{AVV}$ (mean of median distance between the two groups of participants) and randomization distribution $\overline{md}_{R_i}$, we can conclude that sound affects human gaze differently depending on the sound kind, and the effect is greater for human voice (speech, singer, human noise, and singers classes). Compared to a visual attention model, the accuracy of prediction decreases in the group with AV condition compared to the group with V condition. Hence, in future work, it will be interesting to create an audio-visual attention model by adding a sound pathway, for example, locating the sound source of human voice automatically. This research can be applied on video compression and video summarization by enhancing the selection of salient regions in image.

7. REFERENCES

[1] Y. Ma, X. Hua, L. Lu, et al., "A generic framework of user attention model and its application in video summarization," *IEEE Trans. on Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.

[2] C. Quigley, S. Onat, and S. Harding, "Audio-visual integration during overt visual attention," *J. Eye Movement Research*, vol. 1, pp. 1–17, 2008.

[3] J. S. Lee, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation of foveated video coding using audio-visual focus of attention," *IEEE J. Select. Topics in Signal Processing*, vol. 5, no. 7, pp. 1322–1331, Nov. 2011.

[4] P. Besson, V. Popovici, J. M. Vesin, et al., "Extraction of audio features specific to speech production for multimodal speaker detection," *IEEE Trans. on Multimedia*, vol. 10, no. 1, pp. 63–73, Jan. 2008.

[5] G. Song, D. Pellerin, and L. Granjon, "Sound effect on visual gaze when looking at videos," *19th European Signal Processing Conf.*, pp. 2034–2038, Aug. 2011.

[6] M. E. Niessen, L. van Maanen, and T. C. Andringa, "Disambiguating sounds through context," *IEEE Int. Conf. on Semantic Computing*, pp. 88 – 95, Aug. 2008.

[7] E. S. Edgington and P. Onghena, *Randomization Tests*, Chapman Hall/CRC, 4 edition, 2007.

[8] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Reviews Neuroscience*, vol. 2, pp. 194–203, March 2001.