Exploration libre de vidéos : influence du son sur les mouvements oculaires consécutifs à un événement sonore saillant

Antoine Coutrot, Gelu Ionescu, Nathalie Guyader, Alice Caplier

Exploration libre de vidéos : influence du son sur les mouvements oculaires consécutifs à un événement sonore saillant

Antoine Coutrot[1] Gelu Ionescu[1] Nathalie Guyader[1] Alice Caplier[1]

[1] Laboratoire Gipsa-lab

Université de Grenoble - CNRS UMR 5216 - www.gipsa-lab.fr

{prénom.nom}@gipsa-lab.grenoble-inp.fr

Résumé

Dans cette étude, nous utilisons un modèle de saillance sonore (le Discrete Energy Separation Algorithm), capable de détecter automatiquement les événements sonores saillants des bande-son accompagnant des vidéos, pour analyser l'effet du son sur les mouvements oculaires de participants explorant des vidéos. Ce modèle est basé sur une analyse multifréquentielle d'attributs élémentaires, tels l'amplitude et la fréquence moyennes instantanées ainsi que l'énergie de Teager-Kaiser. Les modulations de ces attributs sont combinées linéairement afin de donner une courbe de saillance sonore en fonction du temps. Son seuillage permet de repérer les événements sonores les plus susceptibles d'attirer l'attention. Nous comparons les paramètres des mouvements oculaires (dispersion, amplitude, durée, fréquence) de participants ayant vu les vidéos avec et sans leur bande sonore associée. sur les vidéos étudiées la présence du son modifie certains paramètres étudiés. Cependant cette différence n'est pas plus marquée sur les quelques trames suivant les événements sonores saillants repérés par le modèle qu'en moyenne sur l'ensemble des trames de la vidéo. Ceci laisse à penser que le regard n'est que peu influencé par les attributs élémentaires sonores considérés, du moins sur la base de vidéos étudiées et lorsque l'on s'intéresse à des extraits de courte durée.

Mots clefs

saillance audio-visuelle, attention, vidéos, mouvements oculaires.

1 Introduction

A chaque instant, le cerveau est soumis à un important flot d'informations sensorielles. Malgré ses considérables capacités, il ne peut accorder la même importance à chacune. Afin de sélectionner les plus pertinentes, il utilise un filtre appelé l'attention. Nous nous intéressons ici au lien entre l'attention "visuelle" et "auditive" et plus particulièrement à l'influence du son sur les mouvements oculaires qui précèdent l'attention visuelle. Il est connu que ces deux sens interagissent fortement comme en témoignent les nombreuses illusions audio-visuelles [1, 2].

Depuis les années 1980, de nombreux modèles dits d'attention visuelle ont été développés pour prédire les mouvements oculaires des personnes visionnant des images ou des vidéos [3, 4, 5]. Certains de ces modèles s'inspirent du fonctionnement du système visuel et permettent de prédire la saillance visuelle de stimuli dynamiques. Ces modèles sont ensuite comparés à des données issues d'expériences oculomètriques et montrent une très bonne fiabilité. Malgré leur efficacité, ces modèles n'ont pas été testés dans des situations écologiques de visualisation de vidéos avec leur bande-son. Bien que certains modèles de saillance sonore existent [6, 7], ils n'ont pour le moment pas été intégrés à un modèle de saillance visuelle. Une recherche [8] utilise un modèle combinant à la fois la saillance visuelle et sonore pour générer automatiquement des résumés de vidéos en extrayant les trames les plus saillantes. Une autre étude [9] a également testé un modèle de fusion des informations visuelles et auditives mais en utilisant uniquement des sons latéralisés très simples et des images statiques.

Nous avons voulu dans cette étude tester si un modèle de saillance sonore basé sur des caractéristiques physiques du signal pouvait être utilisé pour analyser l'effet du son sur les mouvemements oculaires de participants explorant des vidéos. En effet, nous savons que le son influence notre perception visuelle. Par exemple, un son orienté provoque plus de mouvements oculaires dans sa direction [9, 10]. Nous nous intéressons ici non pas à l'orientation spatiale du son mais simplement aux événements sonores pouvant renforcer ou non la saillance d'événements visuels. Pour cela, nous présentons un modèle de détection automatique d'événements sonores saillants au sein de vidéos. Il se base sur une analyse multifréquentielle d'attributs élémentaires. Ce modèle permet de repérer au court du temps les événements sonores les plus susceptibles d'attirer l'attention.

Ce modèle est appliqué à une expérience d'oculométrie où ont été enregistrés les mouvements oculaires de participants visionnant des vidéos avec et sans leur bande sonore associée. Cette expérience, au travers des mouvements oculaires des participants, a pour but de tester si des événements sonores saillants peuvent influencer la saillance d'événements visuels. Dans ce but, nous étudions

d'une part l'influence générale du son (présence versus absence) sur les paramètres des mouvements oculaires. D'autre part, nous analysons les mêmes paramètres mais cette fois-ci uniquement à la suite des pics de saillances sonores. En effet, nous pouvons formuler l'hypothèse que l'effet du son est plus marqué au juste après les pics de saillance sonore qu'en moyenne sur l'ensemble des vidéos.

2 Un modèle de saillance sonore : le DESA (Discrete Energy Separation Algorithm)

Le Discrete Energy Separation Algorithm (DESA) a été récemment mis en avant dans des applications nécessitant la détection d'information sonore, tels les résumés automatiques de vidéos [8, 11, 12].
Cet algorithme est fondé sur l'extraction des composantes AM-FM (modulations d'amplitude et de fréquence) dominantes d'un signal sonore. Après la séparation du signal en plusieurs bandes de fréquences et, en partant de l'hypothèse que pour un court laps de temps (de l'ordre de la dizaine de millisecondes), le signal sonore montre une certaine stationnarité dans le domaine de l'énergie et de la fréquence, des nouveaux attributs (énergie, amplitude et fréquence instantanées), sont extraits pour chaque intervalle de temps (correspondant ici à la durée d'une trame, 40 ms). Seule est retenue la bande fréquentielle où l'énergie de Teager-Kaiser est dominante. Depuis son introduction en 1980, sa simplicité d'implémentation et son étroite fenêtre temporelle ont fait de l'énergie de Teager-Kaiser un opérateur privilégié pour détecter les modulations d'amplitude et de fréquence dans les signaux AM-FM [13, 14].
Pour séparer le signal sonore en bandes de fréquences, nous utilisons des filtres de Gabor. Le placement et la bande passante des filtres de Gabor est choisie de manière à ce que l'enveloppe d'un filtre coupe celles de ses premiers voisins à mi-hauteur [15] :

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

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

avec (ω_i, α_i) respectivement la fréquence centrale et la largeur de bande du filtre ($i = 1..N$ avec N le nombre de filtres choisis) et Ω_c la fréquence la plus grande que l'on souhaite analyser. Concrètement, les bandes sonores des vidéos que nous avons analysées ont été échantillonnées à un taux de 48 kHz et séparées en six bandes sonores centrées respectivement sur {281, 562 , 1125 , 2250 , 4500 , 9000} Hz.
Le DESA est simple et efficace. Le traitement appliqué à chaque trame est décrit dans la Figure 1 où l'indice k représente l'échantillon audio (il y a $48000 \times 0.04 = 1920$ échantillons audio dans une trame de 0.04 s dont la bande-son est échantillonnée à 48kHz).

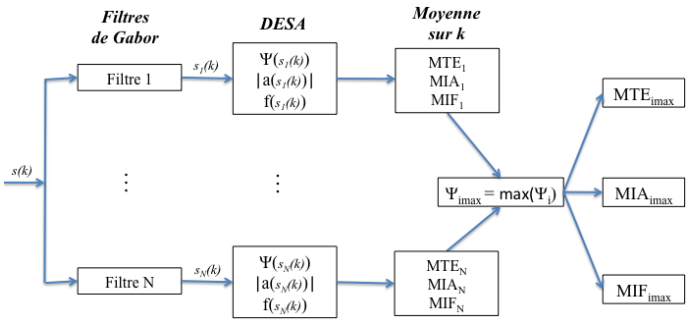


Figure 1 – *Extraction des attributs élémentaires du signal audio $s(k)$. Décomposition selon un banc de filtres de Gabor, calcul de l'énergie, de l'amplitude et de la fréquence instantanées selon les équations 1, 2 et 3. Moyenne sur les k échantillons audio et choix du filtre présentant l'énergie de Teager-Kaiser Ψ moyenne maximale. Les attributs (MTE, MIA et MIF) du filtre sélectionné servent au calcul de la saillance sonore de la trame traitée.*

Le signal sonore original $s(k)$ est séparé en bandes sonores de différentes fréquences par les filtres de Gabor. Ensuite, chaque signal $s_i(k)$ est analysé par le bloc DESA d'où sont extraites l'énergie instantanée fondée sur l'opérateur de Teager-Kaiser (MTE), l'amplitude moyenne instantanée (MIA) et la fréquence moyenne instantanée (MIF), respectivement définies par les formules suivantes :

$$\Psi[x[n]] = x^2[n] - x[n+1]x[n-1] \tag{1}$$

$$|a[n]| = 2\frac{\Psi(x[n])}{\sqrt{\Psi(\dot{x}[n])}} \tag{2}$$

$$f[n] = \frac{1}{2\pi}\arcsin\left(\sqrt{\frac{\Psi[\dot{x}[n]]}{4\Psi[x[n]]}}\right) \tag{3}$$

Le résultat est alors moyenné sur une période représentant la durée d'une trame vidéo (en l'occurrence 40 ms). On sélectionne ensuite le filtre considéré comme étant dominant au sens de l'énergie de Teager-Kaiser. On combine les paramètres MTE, MIA et MIF de la bande fréquentielle sélectionnée pour évaluer la saillance audio S correspondant à la trame traitée m. Une possibilité est la combinaison linéaire des trois paramètres :

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

Ici nous avons choisi une pondération équilibrée :

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

La Figure 2 montre les différentes étapes de traitement appliquées à la bande sonore (1 canal échantillonné à 48 kHz) d'une vidéo de 6840 ms (171 trames de 40 ms) avant d'extraire le profil de sa saillance sonore. Le seuillage de ce dernier donne les "pics de saillance sonore" dont nous nous

servons pour évaluer l'effet du son sur les mouvements oculaires dans l'expérience décrite à la section suivante.

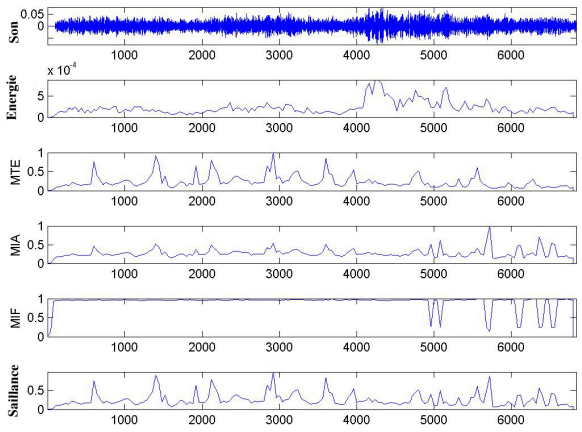


Figure 2 – *Les différents attributs extraits d'une bande fréquentielle du signal sonore (en haut) permettant de calculer la courbe de saillance sonore trame par trame (en bas). La deuxième ligne représente l'énergie "traditionnelle" du signal, la troisième l'énergie de Teager-Kaiser, la quatrième et la cinquième l'amplitude et la fréquence moyennes instantanées. L'axe des abscisses est gradué en millisecondes.*

3 Expérience

Afin d'observer les effets induits par les événements sonores saillants sur les mouvements oculaires lors de l'exploration libre de vidéos, nous avons élaboré une expérience d'oculométrie. Nous avons constitué une base de 50 vidéos que nous avons projetée avec et sans son à 40 participants. Les mouvements oculaires enregistrés lors de cette expérience nous ont permis de tester si la saillance sonore modifiait certains de leurs paramètres caractéristiques, et donc si l'attention sonore pouvait ou non influencer l'attention visuelle.

3.1 Apparatus

Les positions oculaires ont été enregistrées à l'aide d'un oculomètre Eyelink 1000 (SR Research [1]). Ce système permet d'enregistrer les positions de deux yeux à une fréquence de 1000 Hz et une précision d'environ 0,5 degrés. L'appareil est contrôlé par l'interface SoftEye [16] qui permet de contrôler l'ordre et la modalité d'apparition des stimuli. A la fin de chaque expérience on obtient le décours temporel des stimuli projetés et des positions des yeux du participant. Nous pouvons également avoir accès aux événements de type saccade et fixation. C'est à partir de ces fichiers que nous avons extrait les résultats présentés dans la section suivante.

1. www.sr-research.com

3.2 Participants et stimuli

Quarante personnes ont passé l'expérience : 26 hommes et 14 femmes, âgés entre 20 et 29 ans. Tous les participants avaient une vue normale ou corrigée. Les participants étaient naïfs quant au but de l'expérience et avaient pour consigne de regarder les vidéos librement et sans contrainte.
Comme nous l'avons expliqué à la Section 1, nous faisons l'hypothèse que la saillance visuelle sera davantage modifiée à la suite des pics de saillance sonore qu'en moyenne sur les vidéos. Comme nous nous sommes intéressés ici à la nature de l'information sonore et à son apport à l'information visuelle et non à la localisation des sources sonores, les bande-son de chaque vidéo étaient toutes monophoniques.
Nous avons utilisé des vidéos issues de films commerciaux aussi variées que possible (visages, paysages et objets en mouvements pour la vidéo, parole, musique et bruits de l'environnement pour le son) afin de rendre compte de la diversité des scènes audio-visuelles auxquelles nous sommes chaque jour confrontés.
Les vidéos avaient une durée comprise entre 8 et 60 secondes (moyenne 27.7 s, écart-type 13.2 s), une résolution de 720×576 pixels (30 degrés horizontal) et un débit de 25 images (trames) par seconde. Chaque séquence était composée d'un nombre variable de plans. Il y avait au total 160 plans, d'une durée allant de 0.9 s à 35 s (moyenne 8.7 s, écart-type 7.2 s). Le tout avait une durée de 23.1 min.

3.3 Procédure

Comme le montre la Figure 3, l'expérience consistait à visionner librement et sans contrainte l'enchainement des 50 vidéos décrites plus haut, avec entre chaque vidéo un écran gris et une croix de fixation centrale.
Pendant l'expérience les participants étaient assis face à l'oculomètre, leur menton posé sur une mentonnière, en face d'un écran de 21". Ceci permettait de garder une distance constante de 57 cm entre leur visage et l'écran. La résolution de l'écran était 1024x768 pixels et sa fréquence de rafraîchissement de 75 Hz. En début d'expérience, les participants devaient fixer 9 points apparaissant successivement à l'écran afin de calibrer l'oculomètre. Un "drift" (correction) de contrôle entre chaque vidéo permettait de s'assurer que le participant n'avait pas bougé. Si tel était le cas (drift central supérieur à 0.5 degré), une calibration était effectuée à nouveau. Durant toute la durée de l'expérience, les participants portait un casque audio Sennheiser HD280 Pro. Afin d'éviter les effets d'ordre ou de fatigue des participants, les vidéos étaient tirées aléatoirement et sans remise dans la base de 50 vidéos. Pour les 20 premiers participants, les 25 premières vidéos furent jouées avec le son et les 25 dernières sans, et pour les 20 derniers participants, ce fut le contraire. Au final, chaque vidéo fut visionnée par 20 personnes avec le son et par 20 autres sans le son.

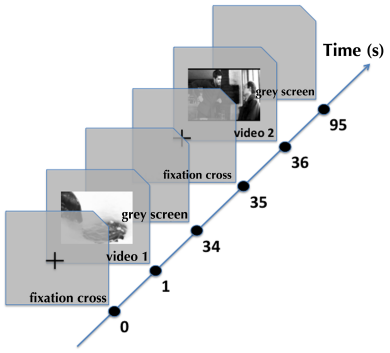


Figure 3 – *Fragment du déroulement d'une expérience présentant l'enchainement de deux vidéos (33 s et 59 s) encadrées par une croix de fixation (1 s) et un écran gris (1 s). Une expérience complète est un enchainement de 50 vidéos.*

4 Résultats

Dans cette section, nous exploitons les informations recueillies à partir des mouvements oculaires des participants ayant participé à l'expérience décrite ci-dessus. Nous comparons les paramètres des mouvement oculaires des participants sur chaque vidéo vues avec et sans le son : d'abord en moyennant sur l'ensemble des trames de la vidéo, ensuite en moyennant uniquement sur les 10 trames suivant les événements sonores saillants détectés par le modèle. Nous analysons plusieurs grandeurs :

- la dispersion des positions oculaires qui reflète la variabilité entre les positions des différents participants ;
- la fréquence des saccades (il s'agit du nombre de saccades effectuées dans une vidéo divisé par le nombre de trames de cette vidéo) ;
- l'amplitude des saccades effectuées ;
- la durée des fixations effectuées.

Pour chacune de ces grandeurs, on réalise une analyse statistique par item (160 plans). Nous travaillons sur les plans et non sur les vidéos entières car les participants ont des mouvements oculaires comparables (diminution brutale de la dispersion, biais central) à chaque changement de plan, indépendamment du stimulus visuel [17]. Pour quantifier la significativité des différences mesurées entre les paramètres étudiés, nous utilisons le test de Kolmogorov-Smirnov. Il s'agit d'un test non paramétrique basé sur les propriétés des fonctions de distributions empiriques. Il permet de tester si deux ensembles de données appartiennent à la même distribution (par exemple, dans notre cas, les mouvements oculaires effectués avec ou sans son). L'avantage de ce test est qu'il ne requiert aucune connaissance *a priori* de la distribution. Le Tableau 1 recense le nombre de données à partir desquelles ont été faites les analyses.

Tableau 1 – *Nombre de saccades et de fixations à partir desquelles ont été menées les analyses.*

	Nombre de mouvements oculaires	
Conditions	Saccades	Fixations
Avec son	46030	48108
Sans son	47261	48702

4.1 Variabilité des positions oculaires

Afin de quantifier les variations des positions des mouvements oculaires parmi les différents participants, nous avons défini un outil que nous appellerons la dispersion D. Pour n participants (donc n points de fixation $(x_i, y_i)_{i\in[1..n]}$),

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

En d'autres termes, il s'agit de la moyenne des distances euclidiennes entre les différentes positions oculaires : pour une image donnée, plus les différents participants regardent au même endroit, plus la dispersion est petite. A l'inverse, si les séquences de positions varient beaucoup d'un participant à l'autre, la dispersion sera importante.
La deuxième colonne du Tableau 2 montre qu'il existe une différence significative entre la dispersion des participants ayant vu un plan donné avec et sans sa bande sonore associée (test de Kolmogorov-Smirnov (ks-test) : $p = 0.018$). Le son apporte donc bien une information supplémentaire à l'information visuelle en ce sens que sa présence provoque une diminution de la variance des positions des regards des participants.
A l'inverse, il n'y a pas de différence significative entre la dispersion moyennée sur chaque plan et celle consécutive aux pics de saillance sonore (deuxième ligne du Tableau 2, ks-test : $p = 0.26$). Les événements sonores saillants tels que définis par le modèle ne semblent pas avoir d'effet sur la position des mouvements oculaires. Cependant il serait nécessaire de détailler davantage ces données en les classant par type de son et d'image : un son de parole n'est certainement pas saillant de la même manière que pourrait l'être un vrombissement de moteur.
La Figure 4 permet d'illustrer cette grandeur : l'ellipse de dispersion (dont les axes ont été calculés par analyse en composantes principales) des fixations des participants

Tableau 2 – *Dispersion moyennée sur chaque plan et sur les 10 trames suivant les pics de saillance sonore dans les deux modalités : avec et sans son (moyenne $\pm$ écart type).*

	Dispersion (pixels)	
Conditions	plans entiers	après sons saillants
Avec son	109 ± 2.4	110.9 ± 2.7
Sans son	116.7 ± 3	121.6 ± 3.3

ayant vu la vidéo avec le son et celles des participants l'ayant vu sans le son se recouvrent largement au cours du temps. Ici nous pouvons constater que l'œil du poisson marteau est le centre de l'attention des participants dans les deux modalités.

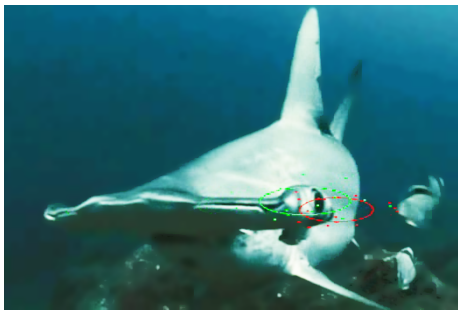
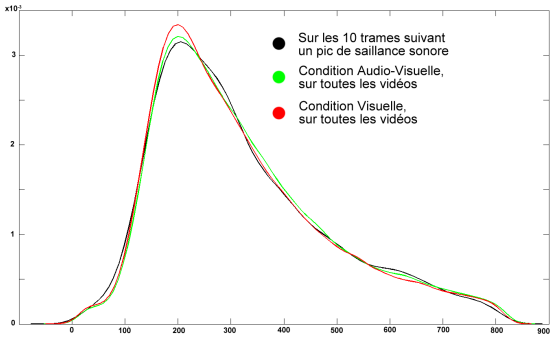
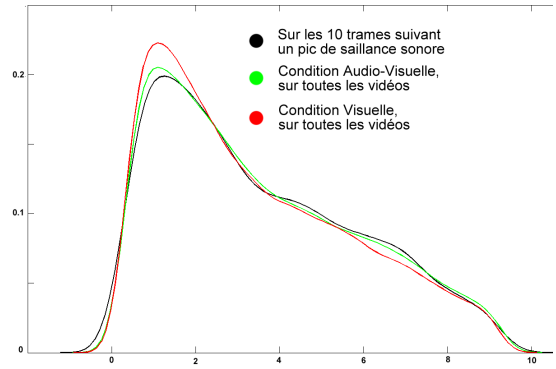


Figure 4 – *trame extraite d'une vidéo projetée lors de l'expérience. L'ellipse verte (respectivement rouge) représente la dispersion des fixations des participants ayant vu la vidéo avec (respectivement sans) sa bande sonore associée.*

4.2 Durées des fixations, amplitudes des saccades



(a) Distribution de la durée des fixations (ms)



(b) Distribution de l'amplitude des saccades (°)

Figure 5 – *Distribution la durée des fixations (en haut) et de l'amplitude des saccades (en bas) effectuées par les participants sur tous les plans avec son (en vert), sans son (en rouge) et durant les 10 trames suivants un pic de saillance sonore (en noir).*

Tableau 3 – *Nombre de saccades par seconde et par participant effectuées sur chaque plan et sur les 10 trames suivant les pics de saillance sonore dans les deux modalités : avec et sans son (moyenne ± écart type).*

	fréquence (s^{-1})	
Conditions	plans entiers	après sons saillants
Avec son	2.28 ± 0.05	2.06 ± 0.05
Sans son	2.37 ± 0.05	2.18 ± 0.05

La Figure 5a compare les distributions des durées de fixation, la Figure 5b compare les distributions des amplitudes des saccades, en rouge sur chaque plan sans le son, en vert sur chaque plan avec le son et en noir sur les 10 trames suivant les pics de saillance sonore (avec le son). On retrouve l'allure classique des distributions de durée de fixations et d'amplitude de saccades pour les stimuli dynamiques [18] (moyenne de durée de fixations autour de 320 ms et de 4 degrés pour l'amplitude des saccades). Là aussi, si la présence de son apporte une différence significative (ks-test, $p < 0.001$), par contre, la proximité d'événements sonores saillants n'induit pas plus de différence entre les distributions de ces deux paramètres (ks-test : $p > 0.04$).
Qualitativement, l'absence de son semble provoquer davantage de courtes saccades et de courtes fixations, ce qui est cohérent avec ce que nous venons de voir avec la dispersion : sans le son, le regard des participants serait moins concentré sur une zone saillante et aurait davantage tendance à explorer la scène.

Comme le montre le Tableau 3, dans la modalité avec son, la proximité d'un son saillant induit une baisse significative (ks-test : $p = 0.002$) du nombre de saccade par seconde. Ceci va dans le sens de notre hypothèse selon laquelle l'attention des participants est mieux captée juste après un pic de saillance sonore : leur regard est davantage fixé sur une région d'intérêt, ils font donc moins de saccades.
Cependant cette tendance s'observe également pour la modalité sans son, bien que dans une moindre mesure (ks-test : $p = 0.032$). Ceci traduit le fait que même sans le son les participants sont davantage focalisés à la suite des pics de saillance sonore, car à ces instants-là, la saillance visuelle est souvent plus forte qu'en moyenne sur la vidéo. En effet, lorsqu'une voiture explose, les participants n'ont pas besoin de la saillance sonore correspondant au bruit de la déflagration pour se concentrer sur la zone d'intérêt.

5 Conclusion

Nous avons utilisé un modèle de saillance sonore capable de détecter dans la bande-son de vidéos les événements sonores les plus susceptibles de capter l'attention des observateurs pour évaluer la propension de la saillance sonore à moduler la saillance visuelle. Les mouvements oculaires de 40 participants visionnant 50 vidéos avec et sans leur bande-son associée ont été enregistrés. La comparaison de ces deux jeux de données a mis en évidence des différences

significatives liées à la présence ou à l'absence de son. Cependant, les paramètres des mouvements oculaires n'ont pas été fondamentalement modifiés juste après les pics de saillance sonore détectés par le modèle. Ceci laisse à penser que si le signal sonore associé à une scène naturelle est une information importante à prendre en compte lors de l'élaboration de modèle de saillance, cette dernière n'est que peu sensible aux attributs élémentaires de la modalité auditive que nous avons considérés, du moins sur la base de vidéos étudiées et lorsque l'on s'intéresse à des extraits de courte durée. Il serait intéressant de refaire ces analyses avec des pics de saillance sonore détectés à partir d'attributs élémentaires différents, ou différemment combinés. D'autre part, l'effet significatif du son constaté pourrait venir de liens plus haut niveau entre l'information sonore et l'information visuelle. Afin d'explorer cette piste, de futures études pourront analyser plus en détail la relation entre ces deux modalités en contrôlant par exemple les attributs visuels et sonores présents dans les stimuli (un son de parole ne serait pas saillant de la même manière avec un visage qu'il ne le serait avec un paysage).

Références

[1] H McGurk et J MacDonald. Hearing lips and seing voices. *Nature*, 264 :746–748, dec 1976.

[2] Jean Vroomen et Beatrice de Gelder. Sound Enhances Visual Perception : Cross-modal Effects of Auditory Organization on Vision. *Journal of Experimental Psychology*, 26(5) :1583–1590, oct 2000.

[3] Anne M. Treisman et Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12 :97–136, 1980.

[4] Laurent Itti, Christof Koch, et Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11) :1254–1259, nov 1998.

[5] Sophie Marat, Tien Ho-Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin, et Anne Guérin-Dugué. Modelling Spatio-Temporal Saliency to Predict Gaze Direction for Short Videos. *International Journal of Computer Vision*, 82(3) :231–243, feb 2009.

[6] S Shamma. On the role of space and time in auditory processing. *Trends in cognitive sciences*, 5(8) :340–348, 2001.

[7] Ozlem Kalinli et Shrikanth Narayanan. A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. Dans *Eighth Annual Conference of the International Speech Communication Association*, pages 1941–1944, Antwerp, Belgium, aug 2007.

[8] K Rapantzikos, G Evangelopoulos, P Maragos, et Y Avrithis. An audio-visual saliency model for movie summarization. Dans *IEEE Int' l Workshop on Multimedia Signal Processing (MMSP-07)*, pages 320–323, Chania, Greece, oct 2007. Springer-Verlag New York Inc.

[9] Selim Onat, Klaus Libertus, et Peter König. Integrating audiovisual information for the control of overt attention. *Journal of Vision*, 7(10) :1–16, jul 2007.

[10] Brian D. Corneil et Douglas P. Munoz. The influence of auditory and visual distractors on human orienting gaze shifts. *The Journal of neuroscience*, 16(24) :8193–8207, dec 1996.

[11] G. Evangelopoulos et P. Maragos. Speech Event Detection using Multiband Modulation Energy. Dans *InterSpeech-2005*, pages 685–688, Lisbon, jun 2005.

[12] G Evangelopoulos, A Zlatintsi, G Skoumas, K Rapantzikos, A Potamianos, P Maragos, et Y Avrithis. Video event detection and summarization using audio, visual and text saliency. Dans *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP-09)*, pages 3553–3556, Taipei (Taiwan), apr 2009.

[13] H. M. Teager. Some observations on oral air flow during phonation. *IEEE Transactions On Acoustics, Speech and Signal Processing*, 28(5) :599–601, oct 1980.

[14] Jim Kaiser. On a simple algorithm to calculate the "energy" of a signal. Dans *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-90*, volume 1, pages 381–384, Albuquerque, NM, USA, apr 1990.

[15] Alan C. Bovik, Petros Maragos, et Thomas F. Quatieri. AM-FM Energy Detection and Separation in Noise Using Multiband Energy Operators. *IEEE Transactions On Signal Processing*, 41(12) :3245–3265, dec 1993.

[16] Gelu Ionescu, Nathalie Guyader, et Anne Guérin-Dugué. SoftEye software. IDDN.FR.001.200017.000.S.P.2010.003.31235, 2009.

[17] Uri Hasson, Ohad Landesman, Barbara Knappmeyer, Ignacio Vallines, Nava Rubin, et David J. Heeger. Neurocinematics : The Neuroscience of Film. *Projections*, 2(1) :1–26, may 2008.

[18] Michael Dorr, Thomas Martinetz, Karl R. Gegenfurtner, et Erhardt Barth. Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, 10(28) :1–17, aug 2010.