UNIVERSITÉ DE GRENOBLE

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité: Signal, Image, Parole, Télécoms

Arrêté ministériel : 7 août 2006

Présentée par

Guanghan SONG

Thèse dirigée par **Denis PELLERIN**

préparée au sein du

GIPSA-lab

dans l'**École Doctorale: Electronique, Electrotechnique, Automatique & Traitement du Signal**

Effect of sound in videos on gaze: Contribution to audio-visual saliency modeling

Thèse soutenue publiquement le **14 juin 2013**,

devant le jury composé de:

M. Hervé GLOTIN

Professeur, Université de Toulon, Rapporteur

M. Patrick LE CALLET

Professeur, Université de Nantes, Rapporteur

M. Didier COQUIN

Professeur, Université de Savoie, Annecy, Examinateur

M. Christophe GARCIA

Professeur, INSA de Lyon, Examinateur

M. Denis PELLERIN

Professeur, Université Joseph Fourier, Grenoble, Directeur de thèse

UNIVERSITÉ DE GRENOBLE

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité: Signal, Image, Parole, Télécoms

Arrêté ministériel : 7 août 2006

Présentée par

Guanghan SONG

Thèse dirigée par **Denis PELLERIN**

préparée au sein du

GIPSA-lab

dans l'**École Doctorale: Electronique, Electrotechnique, Automatique & Traitement du Signal**

Effet du son dans les vidéos sur la direction du regard : Contribution à la modélisation de la saillance audiovisuelle

Thèse soutenue publiquement le **14 juin 2013**,

devant le jury composé de:

M. Hervé GLOTIN

Professeur, Université de Toulon, Rapporteur

M. Patrick LE CALLET

Professeur, Université de Nantes, Rapporteur

M. Didier COQUIN

Professeur, Université de Savoie, Annecy, Examinateur

M. Christophe GARCIA

Professeur, INSA de Lyon, Examinateur

M. Denis PELLERIN

Professeur, Université Joseph Fourier, Grenoble, Directeur de thèse

Acknowledgments

I would like to thank Prof. Denis PELLERIN for his serious and kind supports and helps. With his patient teaching, I began to progressively realize how to do research. Denis kindly provided me the possibility to gain research experiences. I never felt any hierarchy while I was working with Denis. He greatly supported me to start our collaboration with Mr. Lionel GRANJON.

I would like to thank Mr. Lionel GRANJON for his honest helps and supports in this thesis. He has spent lots of time for solving the statistical problems of the thesis, and proposed very helpful suggestions for our experiment design. Without helps and supports from Lionel, finishing this thesis was an impossible task.

I would like to thank Prof. Gang FENG. He was the representative of Gipsa-lab in a phD-recruitment forum, where I got the chance to do a phD at Gipsa-lab. He provided a very convenient way for me to communicate with my supervisor, Prof. Denis PELLERIN. Without his heartfelt help, I would not have this chance to study in this wonderful country.

I would like to sincerely appreciate the exceptional help of Prof. Michèle ROMBAUT and Dr. kai WANG. I always can get valuable comments from them.

I would like to thank the reviewers of this thesis, Prof. Hervé GLOTIN and Prof. Patrick LE CALLET for the evaluation of this work and their fruitful comments. I would like to thank the president of the jury members, Prof. Christophe GARCIA and Prof. Didier COQUIN.

I would like to thank Mr. Gelu IONESCU. He provided me a good condition for the experiment, which is important for my thesis.

I would like to thank Jérôme MARS, Lucia BOUFFARD-TOCAT and Isabelle RAFFIN for their exceptional helps during my thesis.

I would like to thank my boyfriend Zhongyang for his great helps. Our discussions were very leading in this thesis.

I would like to thank my friends Raluca, Xiyan, Hao, Anis, Weiyuan, Haiyang, Shahrbanoo, Ladan and Damien for their invaluable helps.

Abstract

Humans receive large quantity of information from the environment with sight and hearing. To help us to react rapidly and properly, there exist mechanisms in the brain to bias attention towards particular regions, namely the salient regions. This attentional bias is not only influenced by vision, but also influenced by audio-visual interaction. According to existing literature, the visual attention can be studied towards eye movements, however the sound effect on eye movement in videos is little known.

The aim of this thesis is to investigate the influence of sound in videos on eye movement and to propose an audio-visual saliency model to predict salient regions in videos more accurately. For this purpose, we designed a first audio-visual experiment of eye tracking. We created a database of short video excerpts selected from various films. These excerpts were viewed by participants either with their original soundtrack (AV condition), or without soundtrack (V condition). We analyzed the difference of eye positions between participants with AV and V conditions. The results show that there does exist an effect of sound on eye movement and the effect is greater for the on-screen speech class. Then, we designed a second audio-visual experiment with thirteen classes of sound. Through comparing the difference of eye positions between participants with AV and V conditions, we conclude that the effect of sound is different depending on the type of sound, and the classes with human voice (*i.e.* speech, singer, human noise and singers classes) have the greatest effect. More precisely, sound source significantly attracted eye position only when the sound was human voice. Moreover, participants with AV condition had a shorter average duration of fixation than with V condition. Finally, we proposed a preliminary audio-visual saliency model based on the findings of the above experiments. In this model, two fusion strategies of audio and visual information were described: one for speech sound class, and one for musical instrument sound class. The audio-visual fusion strategies defined in the model improves its predictability with AV condition.

Résumé

Les êtres humains reçoivent de grandes quantités d'informations provenant de l'environnement gràce à la vision et à l'audition. Pour nous aider à réagir rapidement et efficacement, il existe des mécanismes dans le cerveau qui porteut notre attention sur des régions particulières, à savoir les régions saillantes. Ce biais attentionnel n'est pas seulement influencé par la vision, mais aussi influencé par l'interaction audiovisuelle. Alors que l'attention visuelle a fait l'objet de nombreuses études, l'effet du son sur les mouvements oculaires a encore peu été exploré.

L'objectif de cette thèse est d'étudier l'influence du son dans les vidéos sur le mouvement des yeux et de proposer un modèle de saillance audiovisuelle pour prédire les régions saillantes dans les vidéos avec plus de précision. À cette fin, nous avons conçu une première expérience audiovisuelle d'oculométrie. Nous avons créé une base de données d'extraits de vidéos choisis dans divers films. Ces extraits ont été regardés par les participants de l'expérience, soit avec leur bande son originale (condition audiovisuelle AV), soit sans bande sonore (condition visuelle V). Nous avons analysé la différence des positions des yeux entre les participants dans les conditions AV et V. Les résultats montrent qu'il existe effectivement un effet du son sur le mouvement des yeux et que cet effet est plus important pour la classe de son « parole à l'écran ». Ensuite, nous avons conçu une seconde expérience audiovisuelle avec treize classes de son. En comparant la différence des positions des yeux entre les participants avec les conditions AV et V, nous avons observé que l'effet du son est différent selon le type de son, et que les classes contenant de la voix humaine (c'est-à-dire les classes de "parole", "chanteur", "bruit humain" et "chanteurs") ont le plus grand effet. De plus, la source sonore a fortement attiré le regard seulement lorsque le son contenait de la voix humaine. En outre, les participants avec la condition AV avaient une durée moyenne de fixation des yeux plus courte que les participants avec la condition V. Enfin, nous avons proposé un modèle préliminaire de saillance audiovisuelle basé sur les résultats des expériences. Dans ce modèle, deux stratégies de fusion d'informations audiovisuelles ont été proposées: l'une pour la classe de son "parole", et l'autre pour la classe de son "instrument de musique". Les stratégies de fusion audiovisuelles définies dans le modèle améliorent la précision de prédiction des régions saillantes pour la condition AV.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In daily life, human receives a large quantity of information from the environment using five senses: vision, hearing, taste, smell and touch. Among these five senses, we rely most on the sense of sight (*vision*). About 80% of the information we take from the environment is provided by our eyesight [Begbie 1996]. Through visual perception, we obtain a lot of information, which helps us to better analyze the environment. For example, during a navigation task, sight helps us to avoid the obstacles. Visual perception of the environment is a complex task, which requires a large number of mechanisms. In the brain, the visual cortex is responsible for processing this visual input. The primary visual cortex transmits information to two primary pathways: one called the dorsal stream, which is associated with motion, representation of object locations, and control of the eyes; the other is the ventral stream, which is associated with shape recognition and object representation.

Large amounts of visual information reach our eyes at all times, and our visual ability is not infinite. In order to react fast and properly after receiving information from the environment, there exist mechanisms in our brain to identify a subset of available sensory information from a scene before further processing it. These mechanisms of attention guide the bias toward particular regions. The eyes are going to focus on some particular regions called *salient regions* that attract attention. The sensors of vision are eyes, which work by allowing light to enter and converting it into electro-chemical impulses in neurons. In the eye, high-resolution images are provided by the center of retina called fovea, which is responsible for sharp central vision. Larger visual field with lower resolution is provided by the rest part of the retina.

The study of eye movement enables a better understanding of the visual system and the mechanisms in our brain to select salient regions. Modeling the visual attention system will help to predict salient regions. There are a lot of applications of this kind of models. The selection of salient region can be used for example to control the level of compression in videos, or to more efficiently guide a mobile robot.

Besides vision, hearing (*audition*) is also an important sense for human to gather information from the environment. For example, during a navigation task, alarm sound also helps us to avoid obstacles. In the brain, the auditory cortex is a region that processes sound and thereby contributes to our ability to hear. The neurons of the primary auditory cortex can be considered to have receptive fields covering a range of auditory frequencies, in a manner that, the neurons at one end of the auditory cortex respond to low frequencies, and those at the other end respond to high frequencies. The rest of auditory cortex areas handle further processing and make it possible to distinguish sounds as speech, music, or noise. In order to react fast after hearing the sound from the environment, there also exist mechanisms of attention in the brain to guide the bias toward the particular salient events in audios. Modeling the auditory attentional system will help to predict these salient events, and can also be applied on event detection, such as speech or music detection.

1.1 Problems

Our different senses receive correlated information from the same objects or events, and this information is combined in our brain. Hence, human behavior is not only influenced by each sense separately, but also influenced by the interaction of different senses. It appears thus important to study for example how vision interacts with hearing. Early researches considered one sense separately from other sensory modalities. The integration of features within single modality (such as visual or auditory) has been actively studied.

Recently, studies of cross-modal integration have been proposed. [Quigley 2008] investigated the influence of audio-visual interaction on eye movement. They focused on how different locations of sound source influence eye movement. The sound was played by loudspeakers in different locations (four conners of a screen), meanwhile the visual stimuli were static images. However, sound effect on gaze for videos was still unknown:

- Has the sound an influence on eye movement, when looking at videos (dynamic and complex stimuli) with its original soundtrack?
- Moreover, is this influence different, depending on the type of sound?

A few audio-visual saliency models, which simulate the behavior influenced by audio-visual interactions, have been utilized in some applications, for instance to select keyframes in videos [Lee 2011, Wang 2012]. In these examples, a visual saliency model is used to predict salient regions in frames, and an audio saliency model is used separately to predict salient audio events. Each model gives a one-dimensional saliency value for every

frame (therefore the spacial information provided by the visual saliency model is lost). Keyframes are selected based on the combination of visual and audio saliency curves. A first work has been performed during the internship of H. Buhr [Buhr 2009]. It allowed to implement and test the audio saliency model of Kayser *et al.* [Kayser 2005]. The problem was complex and the results were difficult to be interpreted, then it has been decided in GIPSA-lab to design and analyze audio-visual experiments to better understand the phenomena. The first experimental study concerns this PhD. A second study started in 2011 with the Master thesis of A. Coutrot, then currently pursuing with a PhD.

1.2 Objectives

The first objective of this thesis is to provide a better understanding of the influence of audio-visual interaction on human behavior. For that, we designed audio-visual experiments to investigate the influence of sound on human gaze when looking at videos. More precisely, we want to find answers to the two questions described in section A.1.

Secondly, based on the knowledge acquired from the above experiment, we want to complete an existing visual saliency model with an additional audio pathway. The objective is to propose an audio-visual saliency model that predicts more accurately salient regions for videos with soundtrack.

1.3 Contributions

In this thesis, two audio-visual experiments are designed to explore the influence of audio-visual interaction on eye movement. During the experiments, participants are divided into two groups: the first group of participants watch the video data with its original soundtrack (audio-visual (AV) condition); the second group of participants watch the same video data without any sound (visual (V) condition). The sound influence on gaze when looking at videos, is investigated, through the analysis of the difference of eye positions between the two groups of participants. Based on this study, a preliminary audio-visual saliency model is proposed to predict salient regions in videos with soundtrack. The contributions of this thesis are briefly summarized as follows:

- ***Sound influences eye movement in videos.***

  To answer the first question whether there is an influence of sound on eye movement in videos, the first audio-visual experiment (experiment I) was designed to investigate the influence of audio-visual interaction on eye movement. Design and analysis of this experiment I are described below:

- *Design of the experiment:* First, one data set consisting of short video excerpts selected from various films, had been created with two conditions: AV condition, the video data set with original soundtrack; V condition, the same data set without soundtrack. Then two groups of participants were asked to watch the same video data set, but with two different conditions (AV and V). The eye positions of the participants of each group were tracked and recorded by an eye tracker.
- *Analysis of the eye position data:* We defined three classes of sound: on-screen speech class (the speakers appear on screen), non-speech class (any kind of audio signal other than speech) and the non-sound class (intensity of sound signal below 40 dB). We observed that the difference of eye positions for the two groups of participants (respectively with AV and V conditions) was greatest for the on-screen speech class. Moreover, the prediction accuracy of a visual saliency model decreased when it was applied on videos with AV condition rather than those with V condition.

- ***Different types of sound influences eye movement in videos differently.***

  Our first experiment showed that sound influenced eye movements differently depending on the sound type. To enrich our study of the influence of audio-visual interaction on eye movement, the second audio-visual experiment (experiment II) was designed. Design and analysis of this experiment II are described below:

  - *Design of the experiment:* Another data set consisting of short video excerpts selected from various films, had been created with AV and V conditions. Compared to the data set in experiment I, we introduced thirteen more refined sound classes. Another difference was that the video excerpts were chosen with two successive sounds (for example, music then speech). The onset of second sound, which was relevant to visual scene, occurring in the middle of excerpts, to avoid the simultaneous change of visual and audio contents as for cuts.
  - *Analysis of the eye position data:* We investigated thirteen types of sound separately and observed that the effect of sound was different depending on the kind of sound, and the classes with human voice (i.e. speech, singer, human noise and singers) had the greatest effect. Furthermore, we assumed that the sound source in the frame attracted attention and therefore calculated the distance between sound source and eye positions of the group of participants with AV condition. The results suggested that only particular types of sound attract

human eye position to the sound source. Finally, in order to find whether the sound has influences on fixation duration, we analyzed the difference in fixation durations between AV and V conditions.

- **Proposal of a preliminary audio-visual saliency model.**

  Through the analysis of the distance between sound source and eye positions of the group with AV condition in experiment II, we found that not always the sound source represents the attractive (salient) regions, depending on the type of sound. Based on this observation, we proposed two kinds of fusion of the audio-visual information:

  - *For speech class:* The sound source (talking face) was the attractive (salient) region for participants with AV condition. Hence, if the sound is classified to speech class, we considered the speaker's face as the salient region.
  - *For music class:* The player's face was experimentally more attractive than the sound source (musical instrument). Hence, if the sound is classified to music class, we considered the player's face as the salient region.

1.4 Organization

This thesis presents the work described in section A.3 with the following structure:

Chapter 2 - Visual and audio attentions, interaction and saliency-based model

This chapter introduces the background of this thesis. It begins with a brief introduction to human visual and audio systems, and the attentional mechanisms. Then current research on the influence of audio-visual interaction on human behavior is presented. Finally, main visual or audio saliency-based models are described.

Chapter 3 - Experiment I: Is there an influence of sound on eye movement in videos?

This chapter presents the details of audio-visual experiment I, which purposes to investigate the sound influence on eye movement. This analysis is based on the comparison of the different eye positions between the two groups of participants respectively with AV condition and V condition.

Chapter 4 - Experiment II: Which type of sound influences eye movement in videos?

This chapter presents the details of audio-visual experiment II, following by a deeper investigation of the influence of audio-visual interaction on eye movement with a more refined sound classification. Besides the analysis of the difference of eye positions of the two group of participants with AV condition and V condition, the sound influence on the duration of eye fixation is also presented.

Chapter 5 - Preliminary audio-visual saliency model

This chapter proposes two different fusion strategy of motion and face for speech and music classes. For speech class, fusion strategy I is used to identify talking face as salient region. For music class, fusion strategy II is used to identify player's face as salient region.

Chapter 6 - Conclusions and perspectives

This chapter provides a general discussion on the results and models presented in the previous chapters. Based on this discussion, we draw conclusions of this dissertation work, and propose some future working directions.

Chapter 2

Visual and audio attentions, interaction and saliency-based model

From our sight and hearing, we receive a large quantity of information about the environment. The fast processing of this information helps to react rapidly and properly. Hence, there exists mechanism in our brain to bias attention towards particular regions or events, called salient regions or events. This attentional bias is not only influenced by visual and auditory information separately, but also influenced by audio-visual interaction.

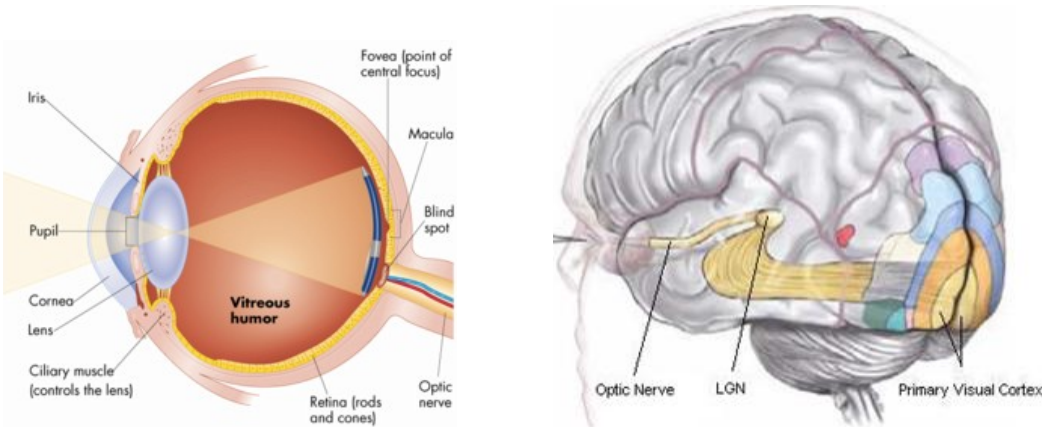
In this chapter, we briefly present the structures of human visual and auditory systems. Then, we introduce the main researches on auditory-visual integration in the brain and the influence of audio-visual interaction on human behavior. Finally, several computational visual or audio saliency models, which simulate attentional mechanism to predict regions with interest, are described.

2.1 Human visual and auditory system

2.1.1 Visual system

The human visual system consists of two functional parts: the eye and part of the brain. Eyes are organs of sight, detecting light and converting it into electro-chemical impulses in neurons. The brain processes the complex image analysis. In the following, we introduce the biological composition of the eye briefly.

Fig. 2.1 (a) shows a cross section of the human eye with the identification of its most important parts. Our perception of a visual scene is determined by the light rays (emitted or reflected) from that scene. When these light rays are strong enough and within the right range of the electromagnetic spectrum (about 300 to 700 nm), the healthy eye sends an electric signal to the brain through the optic nerve. When a light ray crosses the eye, it will pass through the cornea, the aqueous humor, the iris, the lens, and the vitreous



(a) (b)

Figure 2.1: Visual system (a) Structure of the eye [Astroweb] (b) The primary visual cortex, the lateral geniculate nucleus (LGN), and the optic nerve in the brain. [Stanford-Site].

humor, and finally the retina. The cornea is a transparent protective layer, which acts as a lens and refracts the light. The iris can determine the amount of light that can pass through by changing its size. The light rays are detected and converted to electrical signals by photoreceptors in the retina.

In the retina, there are two types of photoreceptors: *rods* and *cones*. The rods respond only to light and dark, spreading all over the retina, except the fovea, with an abundant quantity (about 100 millions in a human eye) [Osterberg 1935]. The cones locate in one small area of the retina (the fovea) with smaller number (6 to 7 millions), and are sensitive to color. The fovea is the area of the retina where our vision is sharpest, corresponding to an area of about 5° of the visual field. It is characterized by a density of cones much larger than in the peripheral area. This difference in distribution of photoreceptors can be explained by goal-saving system resources on the resolution of the visual information. Hence, in order to have the best visual acuity, we need to move the eye to the area, where we want to analyze in detail, with the center of the retina (fovea). About 50% of the information, which is extracted by the retina, came from the fovea. The rest 50% of the information is from other part of retina. That is why we concentrate on the detail of the center of the fovea to have best visual information.

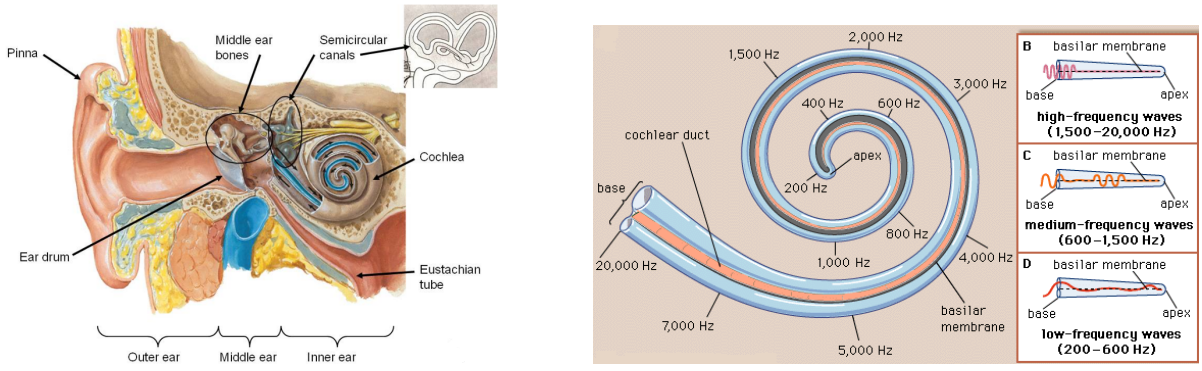
The visual cortex is the largest system in the human brain and is responsible for processing the visual image. Visual input to the brain goes from eye to lateral geniculate nucleus (LGN) and then to primary visual cortex (V1) (see Fig. 2.1 (b)). The LGN is a sensory relay nucleus in the thalamus of the brain. Early responses of V1 neurons consist of sets of selective spatio-temporal filters. In the spatial domain, the functioning of V1 can be considered as many spatially complex Fourier transforms, or more precisely,

Gabor transforms. Theoretically, these filters together can carry out neuronal processing of spatial frequency, orientation, motion, speed (temporal frequency).

2.1.2 Auditory system

In earlier processing of the human auditory system, the sound enters the ear, affecting the cochlea to initiate vibrations of the basilar membrane, then, transducer into spatio-temporal response on auditory nerve. Human ear consists of three stages, outer ear, middle ear and inner ear (see Fig. 2.2 (a)), each fulfilling an essential function.

- *Outer ear:* It locates on the external part of the ear, gathering sound energy and directing sound waves to the eardrum of the middle ear. The configuration of the outer ear boosts the frequencies around 3 kHz of sound pressure. Human speech sounds are distributed in this band around 3 kHz. This is an explanation why the outer ear amplification makes us most sensitive to frequencies of human speech.
- *Middle ear:* It locates between the eardrum and the oval window of the inner ear's cochlea. The main function of this structure is to translate the sound waves into mechanical vibrations as efficiently as possible and to send this information to the following inner ear.
- *Inner ear:* It is a bony labyrinth, comprising two main functions: the organ of hearing (cochlea) and the organ of balance (vestibular apparatus).



(a) (b)

Figure 2.2: Auditory system (a) Structure of the ear [StanfordWeb] (b) Overview of the cochlear functions [IfdWeb].

In the inner ear, the cochlea is the most important auditory portion. It is a spiraled, hollow and conical chamber of bone. The main function of the cochlea is to separate a complex sound into its constituent tonal components. It transduces the waves from the

middle ear to nerve impulses, then, transmitted to the brain through nerve fibers. In order to separate different frequencies in the sound wave, different regions of the basilar membrane through the oval window have particular frequency to initiate oscillations of the organ (see Fig. 2.2 (b)). The orifice of the cochlea is sensitive to the higher frequencies, and lower frequencies travel deeper down the cochlea. Thus, all the points along the cochlea response to a special frequency [Shamma 2001]. A more detailed explanation is described by Pickles [Pickles 2012].

The output of the cochlea carries the impulses to the brain. This nerve consists of thousands of fibers and arises extends to the brain stem. In the brain stem, these fibers contact with the cochlear nucleus, and making the next stage of neural processing in the auditory system. Hence, the cochlear nucleus is the first station in the human auditory system, and the binaural preprocessing begins at this stage.

In the brain, the region which processes the auditory information is called auditory cortex. The mechanisms of the auditory cortex are still unknown according to our knowledge. Primary study shows that the neurons in the auditory cortex are organized according to the different frequencies of sound. It is in a manner that, the neurons at one end of the auditory cortex respond to low frequencies, and those at the other end respond to high frequencies.

In analogy to the visual cortex, there exist multiple auditory areas to distinguish a complete frequency map. The purpose of this frequency map is not clear. However, it is likely to reflect the fact that the auditory system, such as the cochlea, is arranged by the sound frequency.

Auditory scene analysis (ASA)

The process of auditory system taking the mixture of sound from a complex natural environment and sorts it into packages of acoustic evidence, is called "auditory scene analysis (ASA)", in which each package of acoustic evidence has arisen from a single source of sound. This grouping helps to separate information from different sources for pattern recognition [Bregman 1990].

When we use our sense of hearing to understand the properties of sound events from the surroundings, often, we are interested in a single stream of events, such as a violin playing, a person talking, or a car approaching. However, in a natural listening environment, the acoustic energy produced by each sound event sequence is mixed together. All the energy raise from concurrent events at the listener's ears. To understand how the brain could build separate perceptual descriptions of sound-generating events despite this mixing of evidence, auditory scene analysis (ASA) is proposed.

The formation of ASA consists of processes of sequential and simultaneous grouping:

– *Sequential grouping:* It is determined by similarities in the spectrum from one moment to the next. Sequential integration is not only involved in the grouping of a sequence of discrete, but also in the sequential integration of frequency components, for example the integration of the speech of a single voice in a mixture of voices.

– *Simultaneous grouping:* When sounds are mixed, the auditory system must divide up the total set of acoustic components into subsets that come from different sources. To achieve this purpose, it uses properties of the incoming mixture that tend to be true whenever a subset of its components has come from a common source. The grouping of simultaneous components affects many aspects of perception, including the number of sounds, pitch, timbre, loudness.

2.2 Audio-visual interaction

2.2.1 Auditory-visual integration in the brain

Usually, human receive information of the surroundings from more than one sense. Hence, study of the mechanisms of cross-modal integration is necessary. Although the integration of features within single modality (such as visual or auditory) has been actively studied, research into similar cross-modal integration processes is less explored. The information processed by cross-modal integration comes from many different sensory systems (such as visual and auditory), meanwhile, in single modality, information is from single system. Hence, the process of cross-modal integration is more complex. In humans, most research work on multi-sensory integration is still at the stage of demonstrating the phenomenon and understanding the operative factors at perceptual and behavioral levels [King 2009, Mercier 2012]. In the following, we mainly discuss the current research on *when* and *where* auditory and visual information integrate.

When auditory and visual information integrate?

It has been recognized that integrating auditory and visual information implies a decision about whether or not two (or more) sensory cues originate from the same event [Stein 1993, Körding 2007]. It means that the auditory and visual information only occurs if there is sufficient evidence that they are due to a common event (see Fig. 2.3).

Recent studies of nervous system show that auditory and visual stimuli can be integrated by bimodal cells, exhibiting spatially overlapping both in auditory and visual receptive fields [Groh 2002, O'Brien 2010]. Some researchers suggest a spatio-temporal "window" for auditory-visual integration. When auditory and visual stimuli are within this window, they are always perceived as spatially coincident. If the time difference

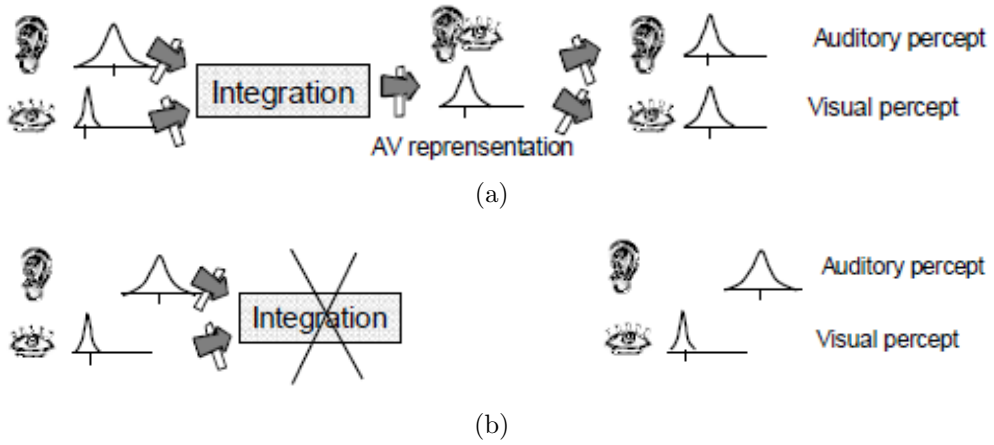


Figure 2.3: Auditory-visual integration (a) The auditory and visual perception of the single audio-visual event with overlapping. (b) The auditory and visual perception of different audio-visual event (the two signals without overlapping) [InriaWeb].

between the auditory and visual stimuli is larger than this "window", the integration between them will decrease a lot. The size of this "window" extends over approximately 100 ms [Lewald 2001]. A deeper study discussed the influence of duration of stimuli of auditory and visual to the size of this "window". The results indicate that when auditory and visual stimuli have unequal durations, there exists a shift in perceived synchrony. This shift in perceived synchrony was observed in the expected negative direction for longer auditory stimuli durations and in a positive direction for longer visual stimuli durations [Kuling 2012].

Where auditory and visual information integrate in the brain?

The perceptual coherence of auditory and visual information is achieved by integrative brain processes. The exact interaction area where the visual and auditory pathways meet to govern processing in the nervous system, still remains unknown, especially when it comes to attentional modulations [Ahveninen 2012]. There are a number of cortical area candidate for examining the neural substrates of multi-sensory processing [Ghazanfar 2006, Carriere 2008]. The cortical anatomy of multi-sensory areas in the primate brain is shown in Fig. 2.4.

Where auditory and visual information integrate in the brain? It is an important question for researchers on auditory-visual integration to find out that whether humans brought multi-sensory information together in the primary sensory processing. Moreover, which cortex areas are associated with this processing? For example, after audio stimuli, the auditory information activates the primary visual cortex directly, or it has to be first processed by the primary auditory cortex and then higher-order association areas?

To discover the neural mechanisms of auditory-visual integration, studying the infor-

mation flow of audio-visual process in the human brain is a crucial pathway [Liang 2008, Marchant 2012]. [Molholm 2002] proposed an investigation of the timing and topography of cortical auditory-visual integrations, using high-density event-related potentials (ERPs) during a reaction-time task. There are two ways to investigate the temporal aspects of brain processing during audio-visual integration: electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) [Burton 2006]. The limitation of EEG is that it cannot provide the actual propagation route across different brain regions in great detail. Meanwhile, to deal with high spatial resolution but relatively low temporal resolution, fMRI emphasized spatial localization of brain activity during audio-visual processing, but only a few fMRI studies have investigated the temporal sequence of brain activations. A rather recent fMRI study by [Alpert 2008] focused on the temporal characteristics of audio-visual processing, using mutual information to assess the relative timing of activations in different brain areas under simultaneous audio-visual (AV) stimulations as well as separate auditory and visual stimulations [Driver 2008]. A number of studies have demonstrated that the relative timing of audio-visual stimuli is especially important for speech signals in multi-sensory integration, although the neuronal mechanisms underlying this complex behavior are unknown. A recent research indicated that a disruption in the temporal synchrony of an audio-visual signal related prefrontal neurons could underlie the loss in intelligibility which occurs with asynchronous speech stimuli [Romanski 2012].

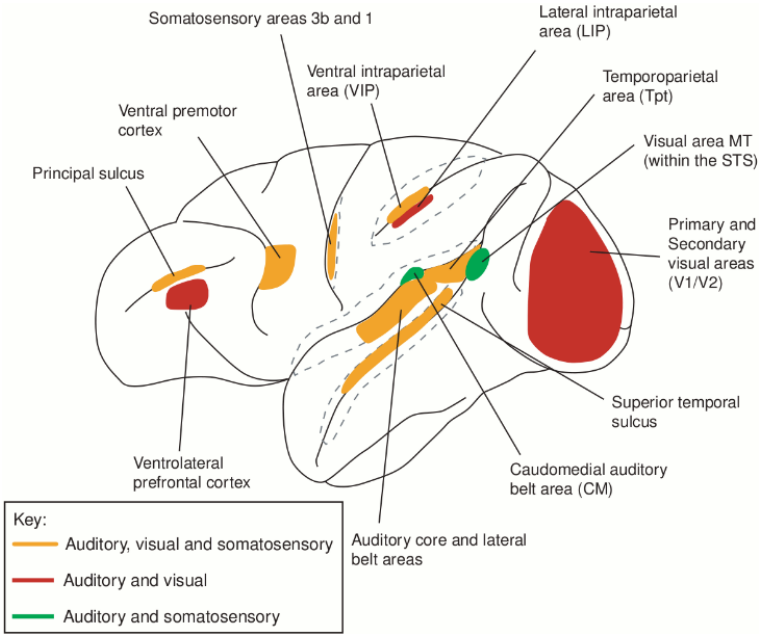


Figure 2.4: The cortical anatomy of multi-sensory areas in the primate brain. Colored areas represent regions where the information from multi-sensory information interact. In V1 and V2, the multi-sensory interactions seem to be restricted to the representation of the peripheral visual field [Ghazanfar 2006].

The mechanisms of auditory-visual integration in the brain are briefly introduced above. Based on these mechanisms, the influence of audio-visual interaction on human behavior is discussed in the following section.

2.2.2 Influence of audio-visual interaction on human behavior

From current psychophysical studies, we know that human response significantly faster for spatially and temporally overlapping bimodal audio-visual than for unimodal (audio or visual) [Sinnett 2008, Frens 1995, Corneil 2002]. The studies on audio-visual interaction concentrate on two areas: the influence of visual input on auditory perception and the influence of acoustic input on visual perception. *Speech* is a special audio stimuli: numerous studies are focused on audio-visual interaction of speech [Alho 2012]. Study from [Tuomainen 2005] provided an evidence of the existence of a specific mode of multi-sensory speech perception. Other types of sound are less investigated.

Visual cues influence audio perception

A lot of researches provide the evidences of the perceptual fusion between audio and visual information, especially for speech. An early evidence is the "McGurk Effect". The "McGurk Effect" is a phenomenon that demonstrates a perceptual fusion between auditory and visual (lip-reading) information in speech perception. In this experiment, a film of a young woman's talking head was shown to the participants, and repeated utterances of the syllable [ba] had been dubbed on to lip movements for [ga], normal adults reported hearing [da] [McGurk 1976]. More research about this "McGurk Effect" are continued. Research from [Cohen 1994] provided an evidence of this "McGurk Effect" perceivers with all language backgrounds, and it also works on young infants [Rosenblum 1997]. Another well-known audio-visual interaction is that visual "lip-reading" helps to understand speech, when speech is in poor acoustical conditions or in foreign language [Jeffers 1971, Summerfield 1987]. Another audio-visual interaction is that 'lip-reading' seeing the speaker's lips enables the listener to better extract useful acoustic information from noise [Schwartz 2004].

Auditory cues also influence visual perception

Previous studies showed that when the auditory and visual signals came from one and the same location, the sound can guide attention toward a visual target [Perrott 1990, Spence 1997]. Besides, other studies demonstrated that synchronous auditory and visual events can improve visual perception [Vroomen 2000, Dalton 2007]. Another study considered the situation that audio and visual information are not from the same spatial place.

The result showed that the synchronous sound "pip" makes the visual object pop out from its complex environment phenomenally [Van der Burg 2008].

More recently, some observations of the mechanisms of speech stimuli and visual interaction demonstrated that lip-read information was more strongly paired with speech information than non-speech information [Vroomen 2011].

[Quigley 2008] investigated how different locations of sound source influence eye movement. The sound was played by loudspeakers in different locations (left, right, up and down), meanwhile, the visual stimuli were static images. The results showed that eye movements were spatially biased towards the regions of the scene corresponding to the locations of the loudspeakers. Auditory influences on visual location also depend on the size of visual target [Heron 2004].

While the interaction of features within audio and visual modalities has been actively studied, the sound effect on human gaze when looking at videos with their original soundtrack is less explored. In the context of research at GIPSA-lab, Coutrot *et al.* [Coutrot 2012a] showed by using some metrics (dispersion of eye positions, Kullback-Leibler divergence, and fixation duration) that sound has an effect on human gaze. Our previous results in [Song 2011a, Song 2011b] with different experimental conditions also concluded that sound affects human gaze in videos.

2.3 Attention and eye movements

2.3.1 Attentional processes

Mechanisms exist in the human brain to identify a subset of available sensory information from a scene before further processing [Itti 1998, Kalinli 2007]. These mechanisms of attentional guidance play a key role in the allocation of resources. These attentional focused regions are influenced by two types of processes: one is "bottom-up" and the other is "top-down".

Bottom-up process

Bottom-up process is an early involuntary and task-independent attentional process. In visual signal, it helps to select and gate visual information based on saliency in the image itself, with various low-level features (orientation, color, motion, etc.) [Desimone 1995]. In audio signals, bottom-up saliency is focused on abrupt changes, transitions and abnormalities in the stream of soundtrack events, like sudden noises or change of sound type in movies [Kalinli 2007].

Top-down process

Top-down process is a voluntary and task-dependent attentional process. It is based on previously learned models (such as expectations, task demands, and emotions) to understand complex scenes, for example searching plates on table (in visual signal) or searching speech in soundtrack (in audio signal).

Trying to understand the relationship between bottom-up and top-down guided selection processes has captured the attention of researchers [Nordfang 2010]. The studies show that bottom-up mechanisms are faster and top-down mechanisms implement our longer-term cognitive strategies [Connor 2004, Parkhurst 2002, Tatler 2005].

2.3.2 Eye movements

Eye movements tightly linked to visual attention [Hoffman 1995, Awh 2006]. Furthermore, eye movements also represent the influence of audio-visual interaction on human behavior [Võ 2012]. The eyes can be moved voluntarily, however, most eye movements are through reflexes. There are three principal types of eye movements: vergence movements (or convergence), saccades, and pursuit movements (or smooth pursuit).

Vergence movements

Vergence movements (or convergence) are the movements to point the fovea of both eyes on a near object, to make sure that the image of the object being looked at falls on the corresponding spot on both retinas. When we change our binocular fixation between targets differing in distance but not in direction relative to the head, we perform this vergence movements of the eyes [Erkelens 2011].

Saccades

Saccades are the rapid eye movements that we make while scanning a visual scene between targets differing in direction. During each saccade, the eyes put the regions, which we are interested in, on the center of the fovea. This eyes movements are controlled by a local, non-visual feedback loop, and are extremely fast (30 and 80 ms) and can reach an angular speed of more than 900° per second.

Between two continuous saccades, the eyes stop moving on a region in the visual scene for a period of time. During this period of non-movement, the visual information is processed by the brain. This period is called a *fixation* and usually lasts between 250 and 500 ms.

The saccade amplitude and fixation duration are related to the quantity of information to be processed during this fixation. One of the main purpose for these saccades is to

scan a greater area with the high resolution of the fovea [Findlay 2009]. Saccades are the fastest eye movements compared to vergence and pursuit eye movements [Oyster 1999]. According to a recent research, near borders, saccades were large and mainly directed in parallel with the borders [Hooge 2012].

Pursuit movements

Pursuit movements (or smooth pursuit) are the movements that the eyes make during the tracking of moving objects. In order to gather more visual information, this eye movement tries that moving object remains stable on the fovea [Carlson 2009]. The pursuit movements are unlike saccades that they are "smooth", without any stop during the pursuit. These pursuit movements cannot be initiate without the moving object [Stoper 1973]. Compared to saccades, the speed of movements is slower, with a maximum speed of about 100° per second.

2.4 Computational attention models

In the previous sections, we introduced the studies of attentional mechanism of visual and auditory-visual integration in neuroscience aspect. Moreover, psychologists have studied the behavioral correlates of visual attention for human and the influence of audio-visual interaction on human behavior. Inspired by these studies, computer scientists tried to simulate this attentional mechanism to create computational attention model, which helps to select important objects from mass of information. This computational attention model provides another way to better understand attentional mechanism. Furthermore, this computational attention model is useful for applications. It can help to select regions of interest to enhance efficiency in applications such as video compression, image synthesis, and robot guidance.

2.4.1 Feature Integration Theory (FIT)

Bottom-up process is driven by low-level features. The saliency models depending on the intrinsic features of the visual stimuli, are called "bottom-up models" and most of them are inspired by the Feature Integration Theory (FIT) of Treisman and Gelade [Treisman 1980].

The first FIT posits that attention must be directed serially to each stimulus in a display whenever conjunctions of more than one separable feature are needed to characterize or distinguish the possible objects presented [Treisman 1980]. Later, other experiments provided more support for this FIT model, that when conjunctive targets are grouped, participants serially scan between groups of targets rather than between individual targets

in visual search [Treisman 1982]. Finally, FIT was completed by that parallel processing occurs when the target has a unique distinguishing feature [Treisman 1985]. Based on previous researches, Wolfe proposed a guided search model. The heart of the guided search model is that attentional deployment of limited resources is guided by the output of the earlier parallel processes [Wolfe 1994].

Based on this FIT theory, a visual stimulus is represented by several elementary feature maps such as luminance, texture, color, edges, orientations and motion. Then, each feature map is normalized to emphasize the regions that are different from their context. All the feature maps are fused together to compose a saliency map. This saliency map emphasizes the salient regions of the input visual scene. The fusion of the different feature maps is usually carried out by a simple sum.

2.4.2 Visual saliency models

Inspired by the Feature Integration Theory (FIT), Koch and Ullman proposed the first visual saliency model based on bottom-up process in 1985 [Koch 1985]. They concluded that early processing stages are able to predict salient regions of a visual scene fairly quickly. Based on this model, a series of visual saliency models had been proposed in recent 20 years. The most popular one is proposed by Itti and Koch [Itti 1998]. In our lab, it exists a spatio-temporal bottom-up saliency model, proposed by Marat et al. [Marat 2010]. This model performs among the upper third of thirty-five computational models of visual attention on video stimuli [Borji 2012].

The Koch-Ullman model

In 1985, Koch and Ullman proposed the first biologically plausible visual attention model [Koch 1985]. It modeled visual *bottom-up* process and it is based on the properties of the visual stimuli to predict the salient areas that attract attention.

This visual attention model has three different stages (see Fig. 2.5) to predict the salient regions:

– From the input image, different visual features are extracted separately, and each feature has its own feature map. These feature maps represent the topographic map in the brain. They decompose the visual stimuli like the visual cortex decompose the elementary visual stimuli. These feature maps extract different elementary features, such as orientation of line segments, colors, motion disparity, etc.. Lateral inhibition within these feature maps enhances the local conspicuity. These regions are the evidences that correspond to the attribute. The output of these maps is combined in the saliency map, which encodes salient features in the visual scene.

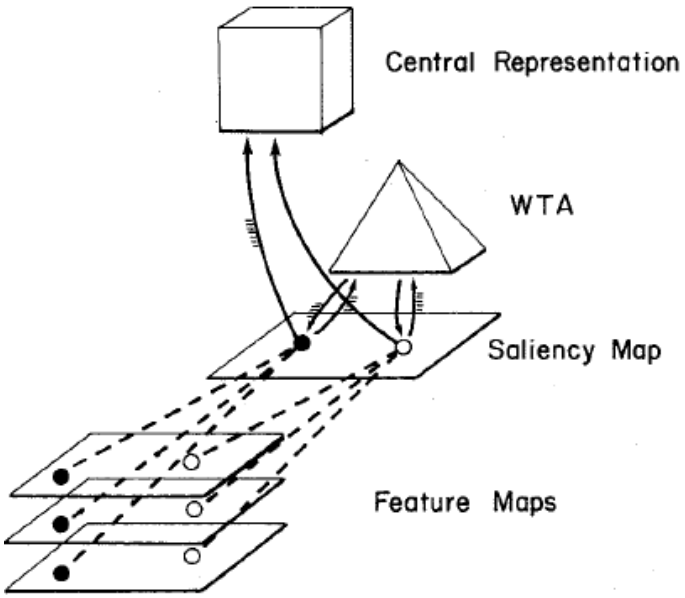


Figure 2.5: The framework of the visual attention model proposed by Koch and Ullman [Koch 1985].

– The results from *Winner Take All (WTA)* can determine the more salient regions in the saliency map. These regions are called *Focus of Attention (FOA)*. The WTA and the details of how it is implemented are important contributions of Koch and Ullman's work.

– After the process of selection of more salient regions, the central representation contains the visual properties of the particular area, and the selected location. The FOA moves to a new area, which inhibits the previous FOA region. Through the WTA mechanism of inhibition, it returns and selects the new most salient region.

The Itti-Koch model

Based on the architecture of the previous model by Koch and Ullman, the most popular visual attention model is proposed by Itti and Koch in 1998 [Itti 1998]. We tested the accuracy of the prediction of this model in Chapter 4, so we describe this model in detail here. The concept maps computed from attribute of massive parallel, and the network WTA "inhibition of return" are taken from the system of Koch and Ullman. This model focuses only on bottom-up process.

Different attribute maps are extracted from the input scene (Fig. 2.6). In the model proposed in 1998, feature maps are combined into three groups of features:

- *Intensity I*

$$I = (r + g + b)/3 \tag{2.1}$$

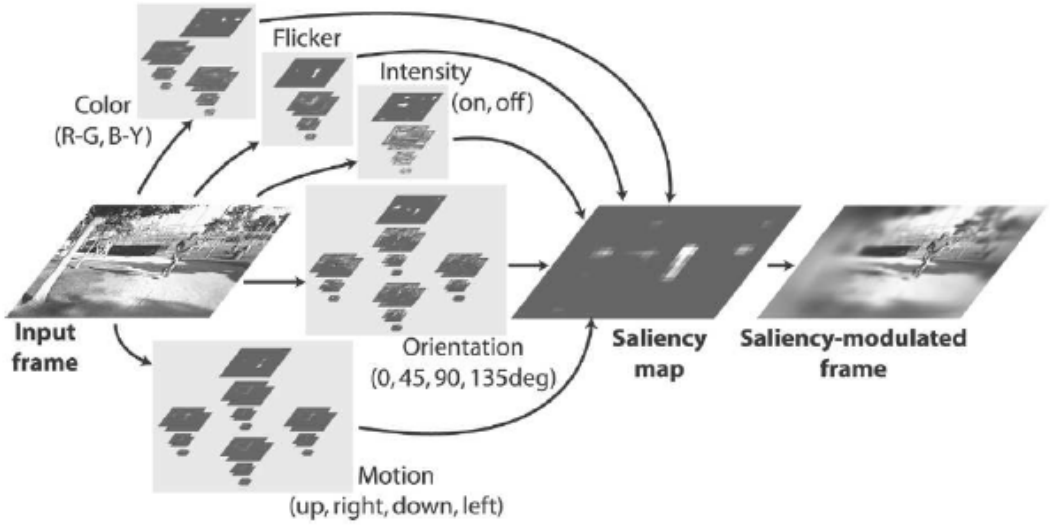


Figure 2.6: The architecture of the visual attention model proposed by Itti, Koch and Niebur [Itti 2005].

Where r, g, and b are the red, green, and blue channels of the input image. In order to decouple hue from intensity, the r, g, and b channels are subsequently normalized by I. I is used to create a Gaussian pyramid $I(\sigma)$.

- *Color C*

  Four broadly-tuned color channels are created:

$$R = r - (g + b)/2 \tag{2.2}$$

$$G = g - (r + b)/2 \tag{2.3}$$

$$B = b - (r + g)/2 \tag{2.4}$$

$$Y = (r + g)/2 - |r - g|/2 - b \tag{2.5}$$

- *Orientation O*

  Local orientation information is obtained from I, using oriented Gabor pyramids with four preferred orientation: 0°, 45°, 90°, and 135°.

Two additional groups of features are added in 2003 [Itti 2003]:

- *Flicker F*

  F is computed from the absolute difference between the luminance I_n of the current frame and that I_{n-1} of the previous frame.

- *Motion R*

  Motion is computed from spatially-shifted differences between Gabor pyramids from the current and previous frames, with the same four Gabor orientations as in the orientation channel, yielding one shifted pyramid S for each Gabor pyramid O.

$$R_n = |O_n * S_{n-1} - O_{n-1} * S_n| \tag{2.6}$$

After the extraction of the features, these maps of attributes are transformed to a "conspicuity map" to locate each location in the visual field by a scalar quantity and to guide the selection of attended locations. This process is based on the spatial distribution of saliency. In order to solve the problem of different dynamic ranges and extraction mechanisms, Itti and Koch proposed a map normalization operator $\mathscr{N}(.)$. It globally promotes maps with a small number of strong peaks of activity (conspicuous locations), while globally suppressing numerous comparable peak responses. The biological motivation behind the design of $\mathscr{N}(.)$ is that it coarsely replicates cortical lateral inhibition mechanisms. The operator $\mathscr{N}(.)$ normalized all the maps in the same range, then each map is multiplied by:

$$(M - \overline{m})^2 \tag{2.7}$$

where, M is the map's global maximum location, and $\overline{m}$ is the average of all the other local maxima. Because only local maxima of activity are considered, $\mathscr{N}(.)$ ignores homogeneous areas and responses associated with meaningful "activitation spots". This process measures the difference between the most active location and the average, through comparing the maximum activity in the entire map to the average overall activation. When this difference is large, the map is strongly promoted with the most active location. Otherwise, the map contains nothing unique and is suppressed.

Feature maps are combined into five separate "conspicuity maps", then normalized and summed into the final input S to the saliency map:

$$S = \frac{1}{5}(\mathscr{N}(\overline{I}) + \mathscr{N}(\overline{C}) + \mathscr{N}(\overline{O}) + \mathscr{N}(\overline{F}) + \mathscr{N}(\overline{R})) \tag{2.8}$$

At any given time, the maximum of the saliency map corresponds to the most salient stimulus that attracts attention. This maximum is selected by a biological implementation of a maximum detector, called WTA neural network.

Siagian and Itti applied this model on a robot localization system in the outdoor environment [Siagian 2007]. Because strong influences on attention and eye movements come from task demands, Baluch and Itti studied the mechanisms of top-down attention [Baluch 2011]. Then, Borji, Sihite and Itti proposed models of top-down visual guidance

using Bayesian Networks [Borji 2011]. Inspired by this classic Itti's attentional model, [Perreira Da Silva 2010] proposed a hierarchical, competitive and non-centralized model without using saliency maps. This new computational model of visual attention was realized on a real-time system by simplifying the calculation processes.

The spatio-temporal saliency model by Marat *et al.*

In our lab, Marat *et al.* proposed a spatio-temporal saliency model to predict eye movements. This biologically inspired model (based on bottom-up process) consists of two pathways: static pathway and dynamic pathway [Marat 2009]. The framework of this model is shown in Fig. 2.7. This model is tested in the following chapters. This model has been completed with a third pathway -face pathway recently [Marat 2012].

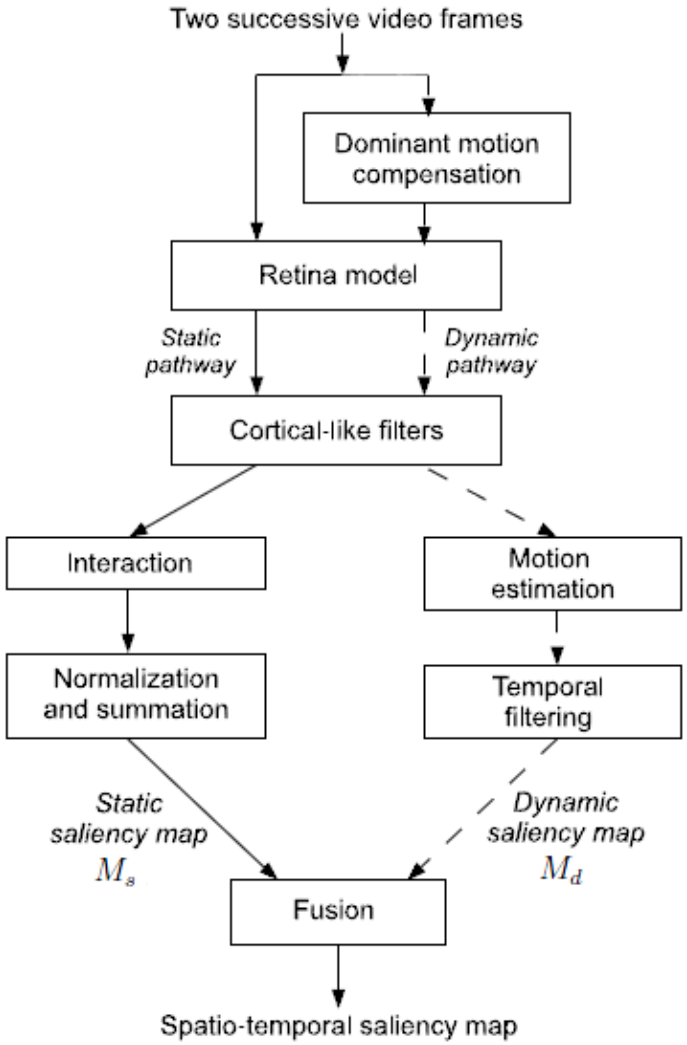


Figure 2.7: The framework of the spatio-temporal saliency model proposed by Marat *et al.* [Marat 2009].

This model is inspired by the first steps of the human visual system: from the retina to

the primary visual cortex. First, it extracts two signals from one frame which correspond to the two main outputs of the retina (parvocellular and magnocellular).

Each signal is then decomposed into elementary features by the cortical-like filters. Here, Gabor filters are used to model V1 cells to extract frequencies, orientations and motion information. These filters extract both static and dynamic information, and provide two saliency maps: a static and a dynamic one, according to their different frequencies.

In static pathway, two types of interactions based on the range of the receptive fields are considered: short and long interactions. The short interactions represent inhibition between neurons of neighboring orientations and overlapping receptivity. The long interaction occurs among collinear neurons beyond the receptive fields.

Dynamic pathway is linked to motion, particularly to the motion which is against the background. A motion estimator is used to calculate the speed of moving region against background. After, a temporal median filter is applied to remove noise.

The last stage is to fuse the two saliency maps (from static and dynamic pathway) to obtain a spatio-temporal saliency map. Four different fusions are proposed, where, M_s represents the static saliency map, while M_d represents the dynamic saliency map:

- *Mean fusion*

  It takes the pixel average of the two saliency maps:

$$M_{mean} = \frac{M_s + M_d}{2} \tag{2.9}$$

- *Max fusion*

  It takes the maximum of the two saliency maps for each pixel:

$$M_{max} = Max(M_s, M_d) \tag{2.10}$$

- *Multiplicative fusion*

  It is a pixel by pixel multiplicative fusion corresponding to a logical *and*:

$$M_{and} = M_s \times M_d \tag{2.11}$$

Then, to be adapted to this fusion, Marat *et al.* proposed the fourth fusion method [Marat 2010]:

- *Reinforced fusion*

The saliency maps weighted by the appropriate characteristic:

$$M_{Rsd} = Max(M_s)\times M_s + Skewness(M_d)\times M_d + (Max(M_s)\times M_s)\times(Skewness(M_d) \times M_d) \tag{2.12}$$

The two last models will be used as references later.

2.4.3 Audio saliency models

Although auditory and visual system have anatomical differences, the mechanism in auditory is similar to that in visual sensory [Shamma 2001]. Several audio attention models are proposed and demonstrated that such models can serve a conceptual basis for comparing the principles underlying of attention across sensory systems.

The Kayser model

The essential concept of auditory saliency modeling was proposed by Kayser *et al.* in 2005 [Kayser 2005]. Compared to the visual attention model proposed by Itti and Koch, the main difference is concentrated on the feature extraction. Feature maps in this model (Fig. 2.8) are clarified by means of a spectrogram, a visual representation of how the frequencies in a sound change over time. The horizontal axis represents time and the vertical one, frequency. The salient event of sound is represented by the brightness of that point in the diagram.

This auditory saliency map extracts individual features, such as spectral or temporal modulation, in parallel way. These features represent various levels of sound feature, analyzed by auditory neurons. After the feature extraction, different sets of filters are used to quantify: sound intensity, frequency contrast, and temporal contrast. Then, all the features are compared across scales with a center-surround mechanism to get the "conspicuity maps" for each feature. To obtain a feature-independent scale, these maps are then normalized with an asymmetric sliding window. In a manner consistent with psychoacoustical masking effects, this window is extended into the past and future. Finally, all the "conspicuity maps" from individual features are combined, in analogy to the idea of the FIT.

The Tsuchida model

Unlike Kayser's model above, which transforms the visual saliency paradigm proposed by Itti *et al.*to the auditory domain, Tsuchida and Cottrell proposed a auditory salience using natural statistics model (ASUN) [Tsuchida 2012]. This ASUN model is an extension of the visual saliency model – salience using natural statistics model (SUN model) [Zhang 2008].

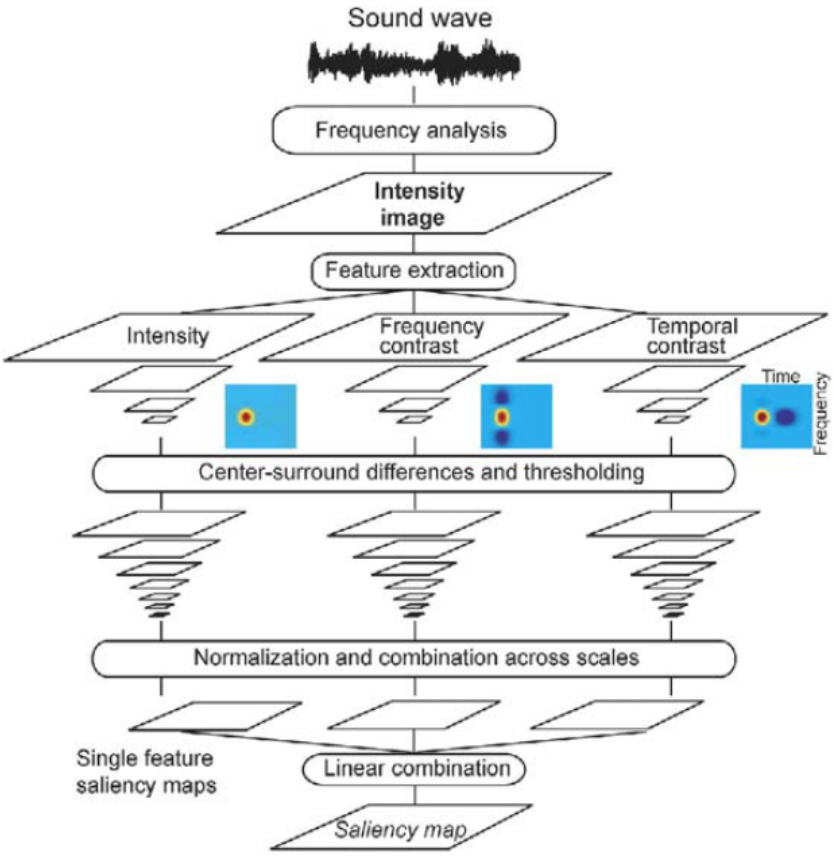


Figure 2.8: The framework of the audio saliency model proposed by Kayser *et al.* [Kayser 2005].

The ASUN model uses a single feature map, which is learned by using independent components analysis (ICA) of natural sounds. The salience at any point is based on the rarity of the realistic auditory feature responses at that point – novelty attracts attention. This model only concentrates on bottom-up portion of SUN. The framework of the audio feature transformation is presented in Fig. 2.9.

Besides the audio saliency model described above, other models are briefly introduced in the following:

- For a specific application on *speech* attention, Kalinli and Narayanan completed the audio saliency model by two additional "conspicuity maps" from the feature extraction of orientations and pitch distribution [Kalinli 2007].

- An auditory attention model is proposed by Duangudom and Anderson to detect what part of a complex auditory scene is most important to analysis [Duangudom 2007]. This model relies on the inhibition of features generated from auditory Spectro-Temporal Receptive Fields (STRF) to compute a saliency map, identifying what is most salient in a complex scene. In their model, the extraction features

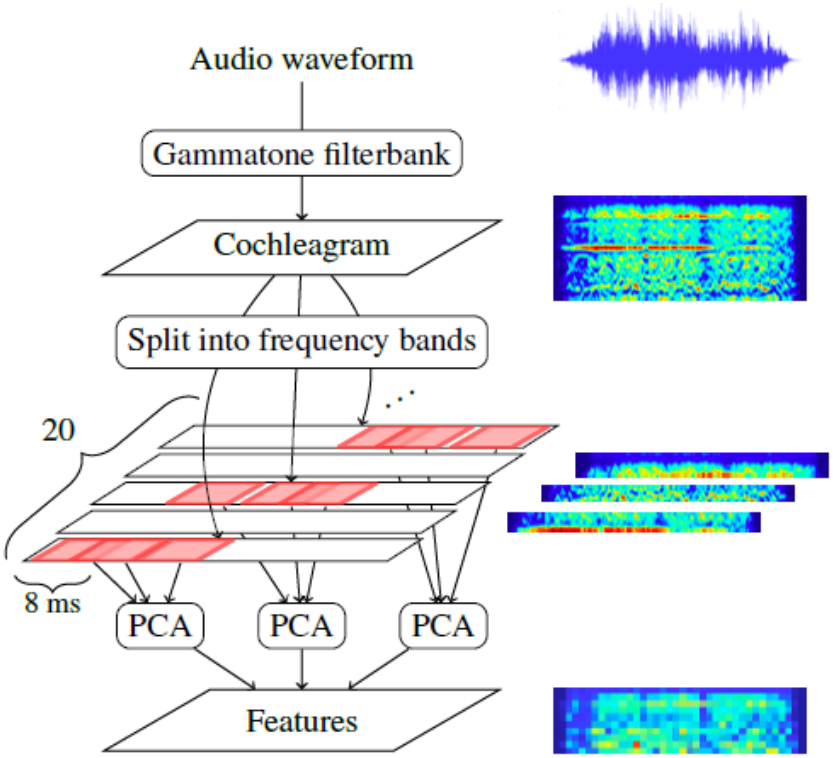


Figure 2.9: The framework of the audio saliency model proposed by Tsuchida and Cottrell [Tsuchida 2012].

are: global energy, temporal modulation, spectral modulation and high temporal spectral modulation.

– Recently, a saliency-based model based on the Discrete Energy Separation Algorithm (DESA) [Litvin 2010], which is used to separate speech and music components from mixed audio stream, is proposed to detect perceptually important audio event [Zlatintsi 2012]. The results from this model are a compact representation of the audio stream by tracking the components with maximal energy contribution across frequencies and time. Note that this model has been tested in [Coutrot 2012b, Coutrot 2013]. This study concluded that with the video dabase used, gaze is weakly influenced by the elementary audio features considered in the model.

2.4.4 Audio-visual saliency models

The saliency map idea is derived from vision, then expanded to audio. Recently, cross-modal interaction of auditory and visual modalities has played an important role in the prediction of human spatial saliency and has been utilized in applications. We present an audio-visual saliency model proposed by [Ma 2005], which is applied on video summarization.

The audio-visual saliency model for keyframe selection

This model consists of three steps: extraction of primary elements of basic channels from the video sequence, generation of saliency curves from a set of attention modeling separately, fusion of these saliency curves to a comprehensive saliency curve. Their work mainly focuses on visual and audio saliency models. The framework of this model is shown in Fig. 2.10.

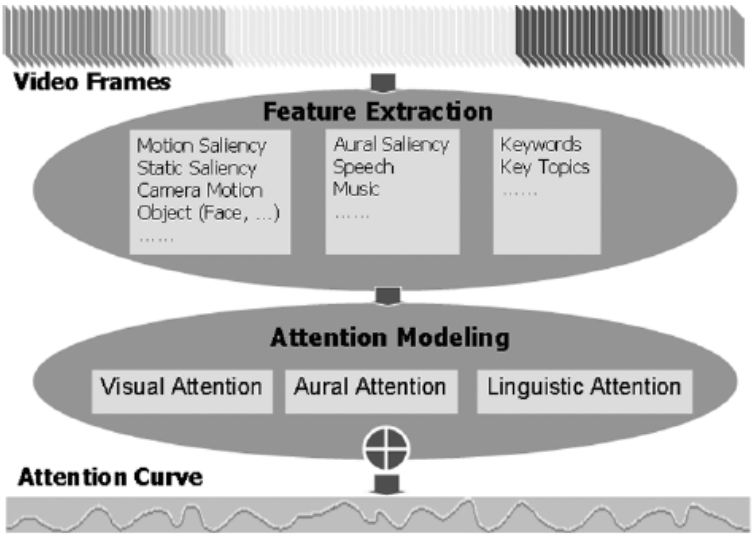


Figure 2.10: The framework of the audio-visual saliency model proposed by Ma *et al.* [Ma 2005].

– ***Visual saliency model***

They model the visual attention by two perceptual models (motion attention model and static attention model), a semantic attention model (face attention model), and a guided attention model (camera motion model).

– ***Audio saliency model***

Audio saliency is based on sound energy, assuming that human may pay attention to a sound if one of the following cases occurs: i) an absolute loud sound measured by average energy of sound; ii) the sudden increases or decreases of the loudness measured by energy peak.

– ***Fusion***

The visual, audio and linguistic saliency curves obtained from different individual channels described above are fused in linear and nonlinear schemes.

There also exists other audio-visual saliency models applied on video summarization with similar architecture, but different extraction of features from visual and audio signals [Evangelopoulos 2008b, Wang 2012], and also applied on video coding [Lee 2011].

The audio-visual saliency model for robot

Another aspect of application of audio-visual saliency model concerns the perception system of robots. In this application, audio saliency model helps the robot to locate the sound source in the space, through turning robot's head and eyes [Ruesch 2008, Kühn 2012]. A very recent study in [Ramenahalli 2013] proposed a simple audio-visual saliency model, which was based on the fusion of visual saliency map and the location of sound source. The sound source space was modeled to be spatially coincident with the visual space.

The work by [Ruesch 2008] presents a multi-modal bottom-up attention system for the humanoid robot where the robot's decisions to move eyes and neck are based on visual and acoustic saliency maps. This model consists of three steps: visual saliency model, spatial auditory saliency model and multi-modal saliency aggregation. In Fig. 2.11, the ego-sphere is a projection surface for spatially related information.

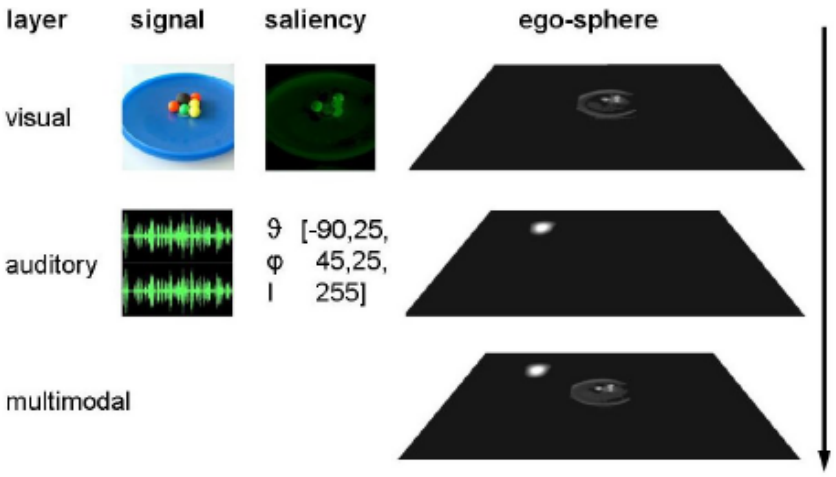


Figure 2.11: The framework of the audio-visual saliency model proposed by Ruesch *et al.* [Ruesch 2008]. Saliency computed from single signal to multi-modal saliency aggregation (left to right). Spatial auditory saliency is shown as a vector containing center location and uncertainty information of direction.

– ***Visual saliency model***

  Visual saliency is composed of the results from filters for intensity, color and motion detection.

– ***Spatial auditory saliency model***

Spatial auditory saliency maps are generated using the position of detected sound sources in the space. The position of a sound source is estimated using interaural spectral differences (ISD) and the interaural time difference (ITD). Spatial auditory saliency is shown as a vector, containing center location, and uncertainty information in longitudinal and latitudinal direction.

– ***Multi-modal saliency aggregation***

After converting the visual and audio saliency information to a common egocentric frame, it is required to combine the visual and audio sensory modalities into an aggregation final map. This is done by taking the maximum value across visual and audio saliency channels at each location.

Due to the fact that auditory output is one dimension, lacking of spatial information, these applications above, only use the audio attention model to select key frame for video scene. There exists no such cross-modal interaction attention model to fuse the two saliency models in two dimensions. However, the cross-modal attention model that predicts the salient regions with the help of audio information is still a gap in the research.

2.5 Conclusion

Human beings have a multitude of senses to receive information from surroundings. Among all senses, two important senses: sight and hearing are introduced. First, human visual and auditory systems are introduced separately. Then, we focus on where and when auditory and visual information integrate in the brain. There exists evidences that audio influences on visual perception and vice versa. Some current researches of the influence of audio-visual interaction on human behavior are then discussed.

In order to select useful information from a huge quantity of information from the environment, we have mechanisms of attention selection to bias attention toward the particularly events both in visual and audio information, which help us react properly and rapidly. These attentional focused regions are influenced by two types of processes: one is a task-independent process, called “bottom-up” and the other is a task-dependent process, called “top-down”.

At last, several principal computational attention models, inspired by bottom-up process and the feature integration theory (FIT), separately in visual and audio are presented. The cross-modal saliency model that predicts the salient regions with the help of audio information is still a gap in the research, meanwhile the sound affects on human gaze when looking at videos with their original soundtrack is less explored.

In the following chapter, an audio-visual psychological experiment on eye movement is designed and the sound effect on human gaze is investigated through the analysis of the eye position data.

Chapter 3

Audio-visual experiment I

In previous chapter, latest studies showed that audio-visual interaction have an influence on human perception. Visual cues can influence audio perception; on the other hand, audio stimuli also can influence visual perception. To evaluate the audio influence on visual attention, investigation of eye movements is a possible way. For static image, the study of [Quigley 2008] indicated that eye movement behavior during the audio-visual condition was the result of an audio-visual interaction process, and the location of the loudspeaker (which carried out the soundtrack) had an influence on eye positions (right or left). However, for videos, does the influence of audio on eye movement still exist? If it exists, how audio stimuli affects visual attention and eye movement for videos?

To answer the questions above, we design an audio-visual experiment to study the sound influence on eye movements when looking freely at videos. In this experiment, we create two sets of short video excerpts. One set has audio and visual information: it consists of video excerpts with their original soundtrack (called AV condition). The other set has only visual information: it consists of the same video excerpts without soundtrack (called V condition). Two groups of participants took part in this experiment: one group watched all the video excerpts with AV condition; and the other group of participants watched the same video excerpts with V condition. Eye positions of the participants from two groups were recorded.

This chapter describes the audio-visual experiment, and analyzes the *difference of eye positions* between the two groups of participants. To complete the analysis, it presents a comparison of the eye position with AV and V conditions separately with a *visual saliency model*. The objective is to evaluate whether the prediction accuracy of the visual saliency model is influenced by sound. Finally, the chapter studies the effect of the sound source localization in the frame on eye movement. This work has been partially presented in two conferences ([Song 2011a, Song 2011b]).

3.1 Eye movement experiment

An audio-visual experiment was designed to investigate the sound effect on gaze when looking at videos, through analyzing the eye positions from the participants. The main idea of this experiment was to divide all the participants to two groups: one group watched the video excerpts without soundtrack (V condition) and the other group watched the same video excerpts with original soundtrack (AV condition). We investigated the sound effect through analyzing the eye position differences between these two groups of participants (with AV and V conditions).

3.1.1 Apparatus

All the eye positions were recorded with an eye tracker, named Eyelink II (SR Research). The EyeLink II system consists of several miniature cameras mounted on a padded headband. These cameras are setting in front of each eye, but they do not shelter against the sight from the participants. Through these cameras, the positions of the two eyes and the pupils are recorded, and the system calculates the positions of the eye on the screen. Two eye cameras allow binocular eye tracking and easy selection of the participant's dominant eye without any mechanical reconfiguration. The head of the participant is fixed on a chin strap that keeps a constant distance between the face of the participant and the screen during the experiment. The principal technical specifications of Eyelink II in this experiment is shown in Table 3.1 [EyelinkWeb].

Table 3.1: The principal technical specifications of Eyelink II

Technical specifications	
Sampling frequency	250 Hz
Average accuracy	< 0.5 °
Spatial resolution (standard deviation)	0.01 °
Saccade event resolution	0.05 ° microsaccades

During this experiment, the sampling frequency is 250 Hz to record the eye positions. These eye position data can be filtered to obtain the duration of saccades and fixations. During one fixation, the eyes rest in the same region. While the eye movement from one fixation to another is considered as a saccade. The stimuli in our experiment to induce eye movements are videos, with or without soundtrack.

3.1.2 Participants

Thirty human participants (10 women and 20 men, aged from 21 to 31) were divided to two groups: fifteen participants viewed video excerpts with their original soundtrack

(AV condition), and the other fifteen participants viewed the same video excerpts without soundtrack (V condition). All participants had normal or corrected to normal vision, and reported normal hearing. They were ignorant of the purpose of the experiment.

3.1.3 Materials

In this experiment, sixty video excerpts lasting 5 to 8 seconds, called clip snippets, were selected from heterogeneous film sources. The sum of all the clip snippets represented 16402 frames. Each clip snippet was converted to the same video format (25 frames per second, 608 × 272 pixels per frame). The sixty clip snippets were then recombined into ten clips, and the structure is shown in Fig. 3.1. Each clip consisted of 6 clip snippets, which came from different film sources. Fig. A.1 shows the content of each snippet in "clip 1".

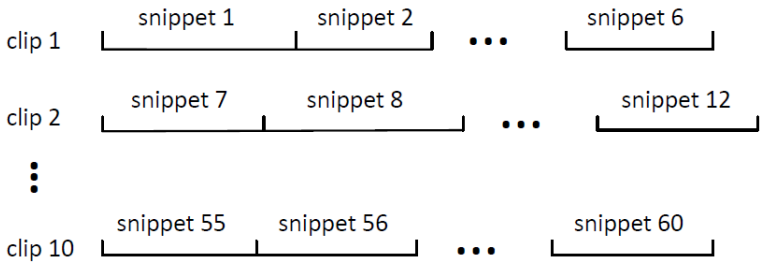


Figure 3.1: The structure of the sixty video excerpts. Six snippets constituted one clip, and there were ten clips in total.

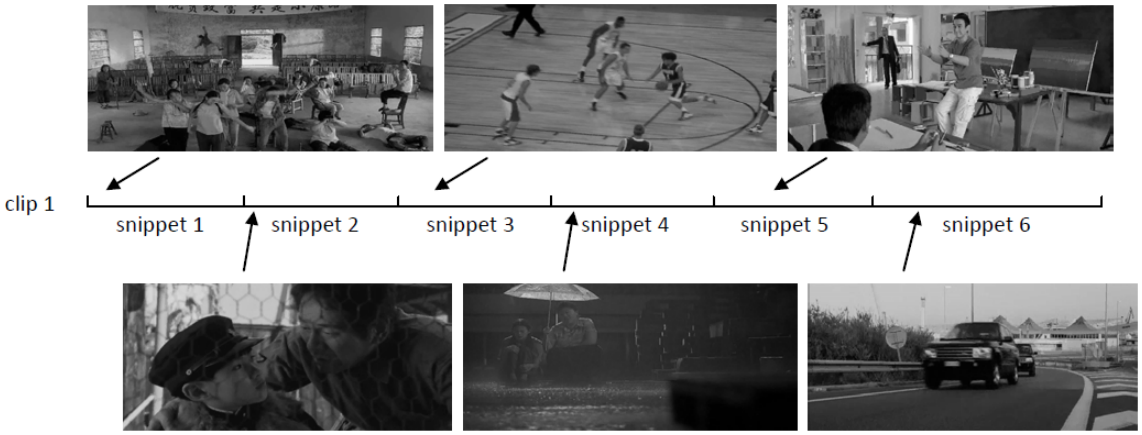


Figure 3.2: The content of each snippet in "clip 1". Each snippet came from different film sources

Because the spatio-temporal saliency model [Marat 2009] calculated in section 3.4 did not consider color information, we used gray level videos as stimuli in the experiment. Two sets of stimuli were built from these clips: one with AV condition (clips with original soundtrack), and the other one with V condition (clips without any soundtrack).

The clip snippet was chosen from films with original soundtrack, which were relevant interesting in both visual and audio fields. In the visual domain, the clip snippets contained various contents, including objects, events, characters, sports and so on. Meanwhile, in the audio domain, the soundtrack contained speech, music, noise and some typical sounds such as rain, knocking a door, etc..

For the reason that we only considered the bottom-up process, the participants viewed the videos without any particular task. Moreover, in order to reduce the effect caused by top-down process, two details of setting were introduced in this experiment. First, as we discussed in chapter 2, the mechanism of top-down process normally influenced attention in the later time [Henderson 1999, Wolf 2000, Tatler 2005]. This influence impacted participants differently over time. Therefore, we concatenated short clip snippets to clips as proposed in [Carmi 2006]. Another aspect, in order to prevent the participants from understanding the language in the video, we chose foreign language films, like Chinese, Indian, Japanese, etc..

3.1.4 Procedure

Human eye positions were recorded by an eye tracker Eyelink II and the clips were presented by SoftEye (a software tool) [Ionescu 2009]. SoftEye is a flexible software tool, synchronized with the eye tracker. All the required data analysis such as eye positions, saccades, fixations detected by the Eyelink II system was recorded in a single file. During the experiment, all the participants were sitting with their chin supported in front of a 19-inch color monitor with 60 Hz refresh rate. The distance between the participant's face and the monitor was 57 cm. The usable field of vision was $20^\circ \times 10^\circ$. The stereo soundtrack was carried by two stereo speakers, placed symmetrically to the monitor. Although the participants were required not to move their head during the experiment, it was hard for them to keep the head completely stable. To reduce the error of eye positions cased by the head movement, a 9-point calibration was carried out every five clips. Before each clip, we presented a drift correction, then a fixation in the center of the screen. Fig. A.2 illustrates the time course of this experimental trials. Participants were asked to look at the ten clips without any particular task. All these ten clips were presented to each participant with random order. Each participant only watched the all clips with one condition – AV or V.

3.1.5 Human eye position density maps

The eye-tracker records eye positions at 250 Hz as mentioned in section 3.1.1. We recorded ten *eye positions* (for the left eye) per frame and per participant. The median of these

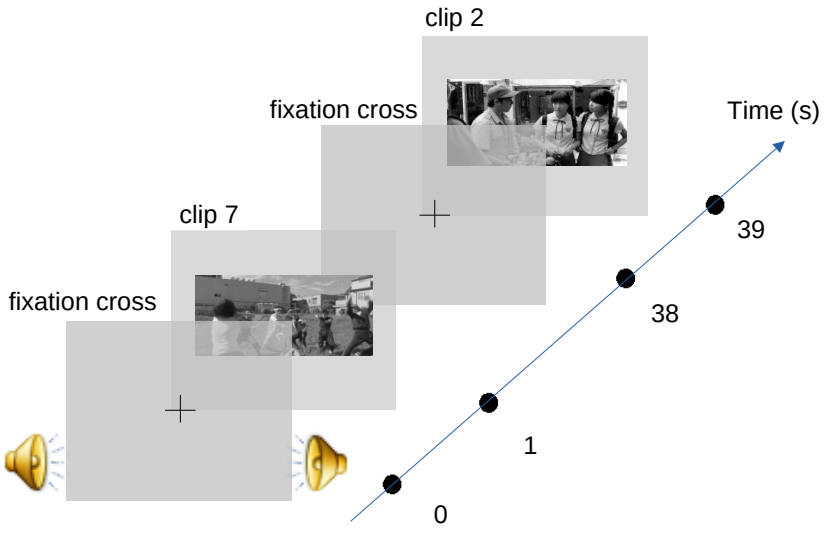


Figure 3.3: Time course of two clips with AV condition. To control the gaze of participant, a fixation cross is presented at the center of the screen before each clip. This sequence is repeated for all the ten clips with random order for each participant.

positions was taken (with X-axis median and Y-axis median) for each frame and for each participant.

For each image k, we obtained a human eye position density map, noted as $M_h(x, y, k)$:

$$\begin{aligned} M_h(x, y, k) &= \sum_{j=1}^{n} \delta(x - x_j, y - y_j) \\ \text{and } \delta(x - x_j, y - y_j) &= \begin{cases} 1 \text{ if } x = x_j \text{ and } y = y_j \\ 0 \text{ if not} \end{cases} \end{aligned} \tag{3.1}$$

where n is the number of participants, (x_j, y_j) is the median eye position of participant j.

In the following calculation, a two-dimension Gaussian was added to each eye position. The standard deviation of the Gaussian was chosen to have a diameter at mid-height equal to 0.5° of visual angle, which is close to the size of the maximum resolution of the fovea.

3.2 Eye position analysis intra each group

Our purpose was to find out whether sound influenced on eye movements when looking at videos. First, the dispersion of the eye positions intra each group (with AV or V condition) was investigated. We tried to evaluate the sound influence through comparing the difference of dispersion from two groups, with AV and V conditions.

In order to investigate the consistency of the eye positions of the participants in each

group (with AV or V condition), we calculated the dispersion intra each group. The dispersion D_p is defined as:

$$Dp = \frac{1}{n^2} \sum_{i,j<i} d_{i,j} \tag{3.2}$$

where, n is the number of participants in one group (with AV or V condition), $d_{i,j}$ is the Euclidean distance of eye positions between participants i and j, and participants i and j are in the same group.

Fig. 3.4 explains how to calculate the "average value of sixty clip snippets". All the sixty snippets were synchronized with the starting frame of each snippet. For example in Fig. 3.5, the value of "dispersion" of frame 1, was the mean of dispersion value of all the snippets for frame 1. This "average value of sixty clip snippets" calculation was not only applied on the "dispersion" calculation, but also on NSS calculation in the following section 3.4.

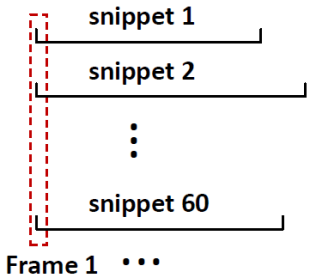


Figure 3.4: Explanation of how to calculate the average value of sixty clip snippets. All the sixty snippets were synchronized with the starting frame of each snippet.

In Fig. 3.5, the low dispersion value represents the eye positions of the participants more focused on the same region. The behavior of dispersion for each group of participants is similar over time. The dispersion is high at the beginning, because eye positions are on the region located at the end of previous clip snippets. From frame 1 to 9, the dispersion decreases sharply and the minimum value appears at frame 9. This decreasing maybe caused by the attractive regions in a new snippet. The situation is the same in the two groups before frame 70. Subsequently, it is stable in the group with V condition, and increases slowly in the group with AV condition, which means with sound, the regions where the participants looked seem more different over a long period. Note that in the study [Coutrot 2012a] with very different experimental conditions (different video content, duration and language), the dispersion with sound is smaller than without sound.

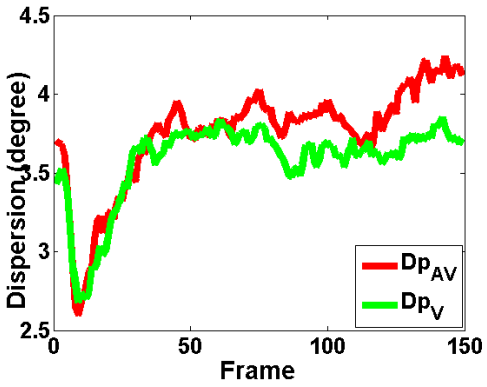


Figure 3.5: Dispersion Dp_{AV} (respectively Dp_V) of eye positions for the group of participants with AV (respectively V) condition over time. For each frame, the dispersion value was an average value of sixty clip snippets (described in Fig. 3.4).

3.3 Eye position analysis inter two groups

In previous section, the difference of dispersion intra each group performed similar in group with AV and V conditions. In this section, in order to investigate the effect of sound on visual gaze, we directly analyzed the difference of eye positions between the two groups of participants with AV and V conditions. Fig. A.3 is an example of eye positions.

Figure 3.6: An example of experimental eye positions of two groups of participants. The red points represent eye position of participants in group with AV condition, and the green points represent eye positions of participants in group with V condition.

3.3.1 Metrics

In order to measure the difference of eye positions from the group with AV condition and the group with V condition, two different metrics were considered: median distance *md* and linear correlation coefficient *cc*.

Median distance *md*

We first calculated the Euclidean distance between eye positions of participants from different groups with AV and V conditions. Then, in order to reduce the influence of the

outliers, we chose "median" (not "mean") to represent the result. This metric is named median distance *md* and defined as:

$$md = median(d_{i,j}), i \in \mathscr{N}, j \in \mathscr{N}' \tag{3.3}$$

where, $\mathscr{N}$ is the group with AV condition and $\mathscr{N}'$ is the group with V condition. $d_{i,j}$ is the Euclidean distance between eye positions of participants i and j, who belong respectively to the group with AV condition and the group with V condition.

Fig. 3.7 (a) shows the median distance *md* values between groups with AV and V conditions of one clip snippet over time. Higher *md* value represented greater difference between eye positions of groups with AV and V conditions. The highest *md* value of this clip snippet appeared at frame 628. Frame 628 with eye positions is shown in (b) of Fig. 3.7. The red points represent eye positions of participants with AV condition, and the green points represent eye positions of participants with V condition. The soundtrack of this snippet is speech, from the adult on the right.

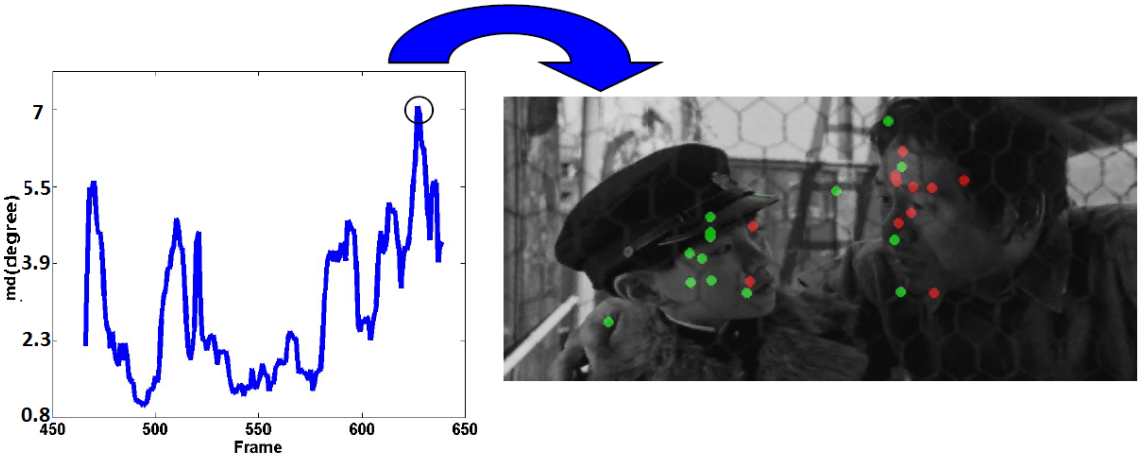


Figure 3.7: (a) Median distance *md* values of one clip snippet over time. (b) Frame 628, which has highest *md* value in this clip snippet, pointed with eye positions of participants. The red points represented eye positions of participants from group with AV condition, and the green points represented eye positions with V condition.

From observing, the red points mainly locate on the adult face (right one) in this video scene, meanwhile, the green points were equally distributed on the two faces. It means, in this example, the participants with AV condition looked at the talking face (sound source), and the participants with V condition looked at the two faces without preference.

Linear correlation coefficient *cc*

Another metric we adopted is the linear correlation coefficient, noted as *cc*. The *cc* describes straight-line relationships between two variables. In our case, the cc assesses

the linearity degree between the two data sets (with AV and V conditions). When the cc value is close to 1, there is an almost perfect linear relationship between the two variables: it indicates low difference between the two data sets. The *cc* is defined as follows:

$$cc(M_{hav}, M_{hv}) = \frac{cov(M_{hav}, M_{hv})}{\sigma_{M_{hav}} \sigma_{M_{hv}}} \tag{3.4}$$

where, M_{hav} (respectively M_{hv}) represents the eye position density maps (mentioned in section 3.1.5) with the AV (respectively V) condition, $cov(M_{hav}, M_{hv})$ is the covariance value between M_{hav} and M_{hv}.

For *cc*, a value of zero indicates no linear relationship between the two maps: there is no correspondence between the eye positions of the two groups with AV and V conditions, and higher values of *cc* indicate higher correspondence between the eye positions of the two groups.

Fig. 3.8 (a) represents *cc* values of one clip snippet over time as an example. We were interested in the frames, where eye positions of the two groups were quite different. Hence, in (b) of Fig. 3.8, we show a frame with zero *cc* value, pointed with eye positions. The red points represent eye positions with AV condition, and the green points represent eye positions with V condition. The soundtrack of this snippet was singers. The singers were the three girls in the front of the image. From observing, the red points mainly located on the singers in the image, meanwhile, most of the green points were focused on the center of the image.

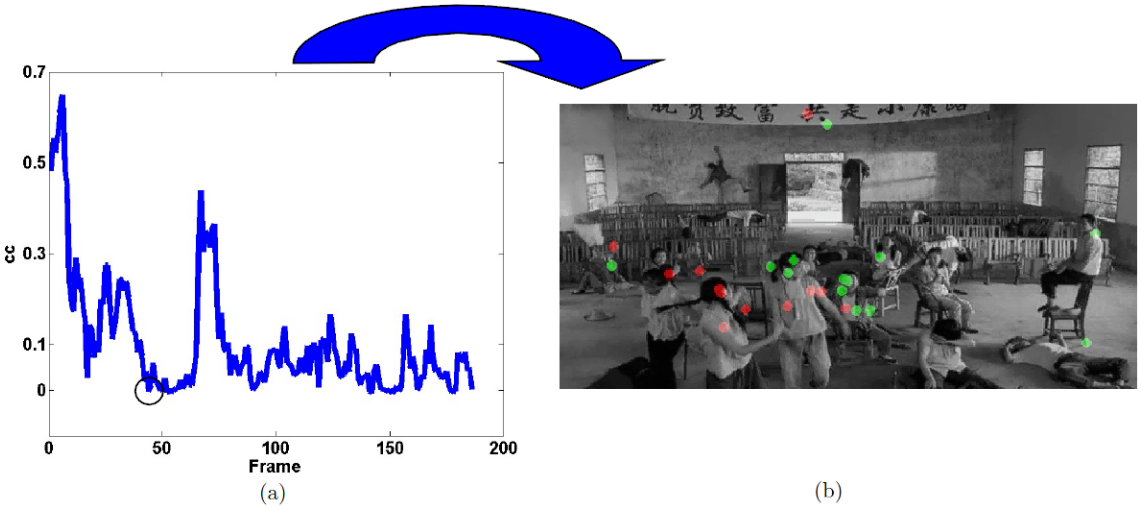


Figure 3.8: (a) Linear correlation coefficient *cc* values of one clip snippet over time. (b) Frame 45 with the lowest *cc* value in this clip snippet. The red points represent eye positions of participants from group with AV condition, and the green points represent eye positions of participants with V condition.

3.3.2 Statistical analysis

Statistics helps us to describe the measurements more precisely. We distinguish two families of statistical tests: the so-called parametric tests (such as ANOVA test) whose conclusions are based on probability that requires the observed distributions or satisfies certain characteristics; non-parametric tests (such as Kruskal-Wallis test), which do not require compliance with these same characteristics.

ANOVA

Analysis of variance (ANOVA) is a collection of statistical models, which compare the means between two or more groups of samples through an estimation of the variance [Gelman 2005]. If the statistically significant probability (p-value) is less than a threshold of significance level (normally considered of 5%), the null hypothesis will be rejected. The simplest form of ANOVA provides a statistical test towards a null hypothesis that several groups are simply random samples from the same population and have equal mean. Rejecting this null hypothesis implies that at least one group has different mean compared to other groups.

In our database, we considered the sound effect as a fixed effect in the ANOVA model, one-way ANOVA form was applied. In the one-way (or single-factor) ANOVA, statistical significance is tested by comparing the $F-test$. The $F-test$ is used in the comparisons of the components of the total deviation, and the definition is presented below:

$$F = \frac{\text{variance between treatments}}{\text{variance within treatments}} \tag{3.5}$$

where the degree of freedom is $n-1$, where n is the number of groups in the comparison.

To apply this widely used statistical method ANOVA, there are several requirements of the data:

- The samples in the data must be independent from each other.
- The data must be from normal distributions.
- All the individuals in the data must be selected from the population randomly.
- Sample sizes should be as equal as possible, but some differences are allowed.

Kruskal-Wallis test

ANOVA is a parametric method, which assumes that the data comes from a normal distribution. In our case, we are not sure the data is a normal population. Thus, we test the Kruskal-Wallis one-way analysis of variance by ranks. It is a nonparametric statistical

method of the classical one-way ANOVA, and compares the medians between two or more samples to determine if the samples come from different populations [Hollander 1999, Corder 2009]. If a significant difference is found in Kruskal-Wallis test, it means that there is a difference between the highest and lowest medians among the groups. Like most non-parametric tests, Kruskal-Wallis performs on ranked data, so the measurement observations are converted to their ranks in the overall data set: the smallest value gets a rank of 1, the next smallest gets a rank of 2, and so on [Howell 1987].

The loss of information involved in substituting ranks from the original values can make this a less powerful test than an ANOVA, so the ANOVA should be used if the data meet the assumptions that the data are normally distributed. In Kruskal-Wallis test, statistical significance is tested by comparing the *Chi-Square test*, and the definition is presented below:

$$\chi^2 = \sum_{i=1}^{n} \frac{(\text{Observed value}_i - \text{Expected value}_i)^2}{\text{Expected value}_i} \tag{3.6}$$

and the degree of freedom is $n-1$, where n is the number of groups in the comparison.

The requirements of the data to apply Kruskal-Wallis test are shown below:

- The samples in the data must be independent from each other.
- The distributions of the data do not have to be normal and the variances do not have to be equal.

3.3.3 Results

Through observing the video pointed with eye positions of the participants, we found that different kinds of sound affect the eye positions differently. Hence, we manually classified all the eye positions, according to the sound type, to three classes: on-screen speech (the speakers appear on screen), non-speech (any kind of audio signal other than speech) and non-sound (intensity below 40 dB) with the software named Praat. [1]

Both ANOVA and Kruskal-Wallis test require the database should be *independent samples*. Our eye position data consider continuous measurement over time, the eye positions for most participants does not change much between two adjacent frames, they could not be considered as independent samples. To solve this problem, we assume a set of continuous frames to be one independent sample. For the size of the set, we choose the average value of one fixation duration (about 8 frames).

[1] Praat is a free program, which provides spectral, pitch, formant and intensity analysis and annotation of sound signal [PraatSite].

Statistical analysis of median distance md

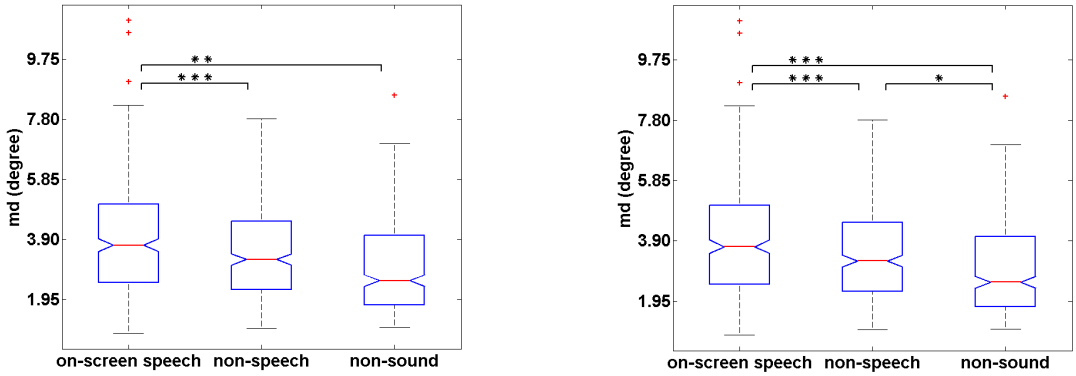
We first calculated the median distance md of eye positions between two groups with AV and V conditions according to equation 3.3 for each frame. Then, all the eye position were classified into three classes: on-screen speech, non-speech and non-sound. Within each class, we took the mean of md of 8 continuous frames as one independent sample. Based on these independent samples, tests of ANOVA and Kruskal-Wallis are shown in Fig. A.4. [2]

In Fig. A.4 (a), with ANOVA test, the result of $F(2, 742) = 9.24$ and $p < 10^{-4}$ indicates that among the three groups: on-screen speech, non-speech and non-sound, at least the mean value of one class is significantly different from the other two classes. Moreover, the mean value of md tends decreasing from on-screen speech to non-speech, finally to non-sound. The mean value of on-screen speech is significantly different from the other two classes: between on-screen speech and non-speech ($F(1, 673) = 12.27$, $p < 10^{-3}$), between on-screen speech and non-sound ($F(1, 420) = 10.44$, $p < 10^{-2}$). It gets the highest mean value among these three classes with median distance md measurement, suggesting the highest difference between the groups with AV and V conditions. Between non-speech and non-sound ($F(1, 391) = 1.99$, $p = 0.16$), the difference is not significant with ANOVA test.

In Fig. A.4 (b), with Kruskal-Wallis test, the result of $\chi^2(2) = 19.63$ and $p < 10^{-5}$ indicates that among the three groups: on-screen speech, non-speech and non-sound, at least the median value of one class is significantly different from the other two classes. Moreover, the median value of md decreases from on-screen speech to non-speech, finally to non-sound. The median value of on-screen speech is significantly different from the other two classes: between on-screen speech and non-speech ($\chi^2(1) = 10.61$, $p < 10^{-3}$), between on-screen speech and non-sound ($\chi^2(1) = 13.82$, $p < 10^{-3}$). It gets the highest median value among these three classes with median distance md measurement, suggesting the highest difference between the groups with AV and V conditions. Between non-speech and non-sound ($\chi^2(1) = 4.15$, $p = 0.04$), the difference is still significant with Kruskal-Wallis test.

In order to determine which statistical analysis (ANOVA or Kruskal-Wallis test) is proper to deal with our data, a Lilliefors test is applied to verify whether the data is from normal distribution or not. In statistics, the Lilliefors test is an adaptation of the Kolmogorov-Smirnov test [Lilliefors 1969], which is used to justify whether the data is from a standard normal distribution or not. The null hypothesis of Lilliefors test is that data are from a normally distributed population. In this null hypothesis, it does not

[2] In the figure, '*' indicates the p value is < 0.05, '**' indicates the p value is < 0.01, '***' indicates the p value is < 0.001.



(a) ANOVA (b) Kruskal-Wallis test

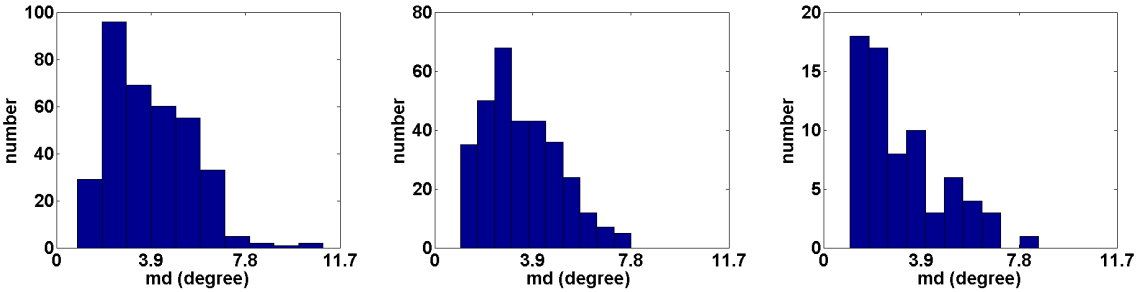
Figure 3.9: Comparison of median distance *md* between the two groups of participants (with AV and V conditions), among three classes of sound: on-screen speech, non-speech and non-sound, by using (a) ANOVA and (b) Kruskal-Wallis test.

specify the expected value and variance of the distribution. The Lilliefors test statistic is the same as for the Kolmogorov-Smirnov test, which is defined below:

$$KS = max_x \left| SCDF(x) - CDF(x) \right| \tag{3.7}$$

where CDF is the normal cumulative distribution function (cdf) with mean and standard deviation equal to the mean and standard deviation of the sample, and $SCDF$ is the empirical CDF estimated from the sample.

Fig. 3.10 shows the distribution of the data for each class. With the Lilliefors test, if it rejects the null hypothesis at the 5% significance level with the logical value $h = 1$, and $h = 0$ if it cannot. That means, if $h = 1$, the data does not come from a normal distribution.



(a) on-screen speech (b) non-speech (c) non-sound

Figure 3.10: The distribution of *md* value for the three classes: on-screen speech, non-speech and non-sound.

The result of Lilliefors test is shown in Table 3.2. All the three classes rejected the null hypothesis that the data came from a normal distribution. In this case, the accuracy of the results from Kruskal-Wallis is better than that from ANOVA test.

From the two tests, we concluded that sound affects human gaze in videos, and this influence is different depending on the sound type. The median value of on-screen speech gets the highest median value among these three classes with median distance *md* measurement, suggesting the highest difference between the groups with AV and V conditions.

Table 3.2: Results of Lilliefors test of *md* value of three classes: on-screen speech, non-speech and non-sound.

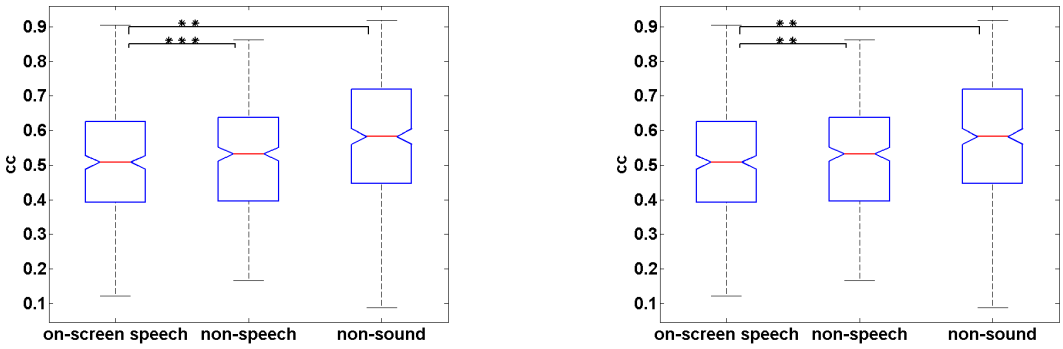
	on-screen speech	non-speech	non-sound
h	1	1	1
p-value	$< 10^{-6}$	$< 10^{-5}$	$< 10^{-3}$

Statistical analysis of linear correlation coefficient *cc*

We repeated the same procedure as shown above to the measurement of linear correlation coefficient *cc*, according to equation 3.4 for each frame. Compared to *md*, lower *cc* values represented higher *md* values. Within each class, we took the mean of *cc* of 8 continuous frames as one independent sample. Based on these independent samples, ANOVA and Kruskal-Wallis tests were calculated.

In Fig. 3.11 (a), with ANOVA test, the result of $F(2, 742) = 7.7$ and $p < 10^{-3}$ indicates that among the three groups: on-screen speech, non-speech and non-sound, at least the mean value of one class is significantly different from the other two classes. Moreover, the mean value of *cc* tends increasing from on-screen speech to non-speech, finally to non-sound. The mean value of on-screen speech is significantly different from the other two classes: between on-screen speech and non-speech ($F(1, 673) = 12.09$, $p < 10^{-3}$), between on-screen speech and non-sound ($F(1, 420) = 7.49$, $p < 10^{-2}$). It gets the lowest mean value among these three classes with linear correlation coefficient *cc* measurement, suggesting the highest difference between the groups with AV and V conditions. Non-speech tends to have lower value than non-sound, but the difference is not significant with ANOVA test ($F(1, 391) = 0.55$, $p = 0.46$).

In Fig. 3.11 (b), with Kruskal-Wallis test, the result of $\chi^2(2) = 14.21$ and $p < 10^{-3}$ indicates that among the three groups: on-screen speech, non-speech and non-sound, at least the median value of one class is significantly different from the other two classes. Moreover, the linear correlation coefficient *cc* tends increasing from on-screen speech to non-speech, finally to non-sound. The median value of on-screen is significantly different from the other two classes: between on-screen speech and non-speech ($\chi^2(1) = 9.83$,

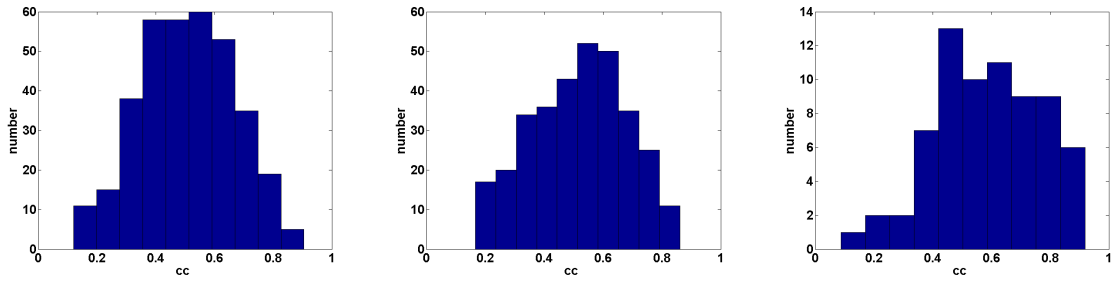


(a) ANOVA (b) Kruskal-Wallis test

Figure 3.11: Comparison of linear correlation coefficient *cc* between the two groups of participants (with AV and V conditions), among three classes of sound: on-screen speech, non-speech and non-sound, by using (a) ANOVA and (b) Kruskal-Wallis test.

$p < 10^{-2}$), between on-screen speech and non-sound ($\chi^2(1) = 8.28$, $p < 10^{-2}$). It gets the lowest median value among these three classes with linear correlation coefficient *cc* measurement, suggesting the highest difference between the groups with AV and V conditions. Non-speech tends to have lower value than non-sound, but the difference is not significant with Kruskal-Wallis test ($\chi^2(1) = 1.15$, $p = 0.28$).

Also, a Lilliefors test was applied to verify whether the data of *cc* values is from normal distribution or not. Fig. 3.12 shows the distribution of the data for each class. With the Lilliefors test, if it rejects the null hypothesis at the 5% significance level with the logical value $h = 1$, and $h = 0$ if it cannot. That means, if $h = 1$, the data does not come from a normal distribution.



(a) on-screen speech (b) non-speech (c) non-sound

Figure 3.12: The distribution of *cc* value for three classes: on-screen speech, non-speech and non-sound.

The result of Lilliefors test of *cc* is shown in Table 3.3. On-screen speech class accepts the null hypothesis that the data came from a normal distribution, and the other two classes reject the null hypothesis.

The results from ANOVA and Kruskal-Wallis are similar. We concluded that sound affects human gaze in videos, and this influence is different depending on the sound type. The mean value of on-screen speech gets the lowest mean value among these three classes with linear correlation coefficient *cc* measurement, suggesting the highest difference between the groups with AV and V conditions. Non-speech class tends to have lower *cc* value than non-sound class, suggesting higher difference between the groups with AV and V conditions.

Table 3.3: Results of Lilliefors test of *cc* value of three classes: on-screen speech, non-speech and non-sound.

	on-screen speech	non-speech	non-sound
h	0	1	1
p-value	0.20	0.01	0.50

3.3.4 Conclusion

We analyzed the sound influence on human gaze through observing the difference of eye positions between groups with AV and V conditions. This difference of eye positions were measured by two metrics: median distance *md* and linear correlation coefficient *cc*. Through the statistical analysis of ANOVA (parametric method) and Kruskal-Wallis (nonparametric method), both *md* and *cc* confirmed the existence of sound influence on gaze when looking at videos.

Moreover, different types of sound influenced eye movement differently. Three classes: on-screen speech, non-speech and non-sound were classified manually and analyzed separately. Both *md* and *cc* showed the difference between the groups with AV and V conditions was highest in on-screen speech class, and this difference is significant at a level of 5%.

3.4 Effect of sound on a visual saliency model

To complete the analysis, we investigate whether there is an effect of sound on a visual saliency model in this section. To evaluate whether the prediction accuracy of a visual saliency model decrease, we compare the experimental eye positions from group with AV condition (respectively with V condition) with a visual saliency model.

The visual saliency model we chose is a spatio-temporal saliency model developed in our laboratory by S. Marat et al. [Marat 2009]. This model is introduced in section 2.4.2. It is inspired by the biology of the first steps of the human visual system, extracting two signals from a video stream corresponding to the two main outputs of the retina:

parvocellular and magnocellular. Then, both signals are split into elementary feature maps by cortical-like filters. These feature maps are used to form two saliency maps: a static (output of the static pathway) and a dynamic one (output of the dynamic pathway).

The static pathway of the visual saliency model consists of two types of interactions based on the range of the receptive fields. The static saliency map mainly represents the edge of the objects, which have large contrast from the background. The dynamic pathway is tightly linked to motion and particularly to the motion of a region against the background. The dynamic saliency map is sensible to the motion amplitude against the background, not the orientation of the motion.

3.4.1 Criterion

For the evaluation, we chose the Normalized Scanpath Saliency (NSS) criterion, which was proposed by Peters and Itti [Peters 2005]. It is especially designed to compare eye positions with the salient areas emphasized by a saliency model.

The NSS metric corresponds to a Z-score, which computed by comparing a computational saliency map from the model to eye positions of participants. The larger the value of Z-score is, the less probable it is that the experimental results are due to chance. We computed the NSS metric as follows:

$$NSS(k) = \frac{\overline{M_h(x,y,k) \times M_m(x,y,k)} - \overline{M_m(x,y,k)}}{\sigma_{M_m(x,y,k)}} \tag{3.8}$$

where, (x, y) are the coordinates of the eye position, and k is the frame number. $M_h(x, y, k)$ 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.

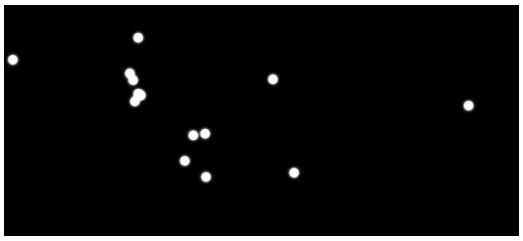
Zero $NSS(k)$ value indicates no correspondence between saliency map and eye positions. High $NSS(k)$ value (maybe above one) suggests a greater correspondence. First, we calculated the NSS, successively from static pathway and dynamic pathway, for the two groups with AV and V conditions separately. Then we analyzed the difference of NSS for each frame between two groups in three classes. Fig. 3.13 shows the $M_h(x, y, k)$ and $M_m(x, y, k)$ for groups with AV and V conditions of one frame as an example.

3.4.2 Approach to calculate prediction accuracy

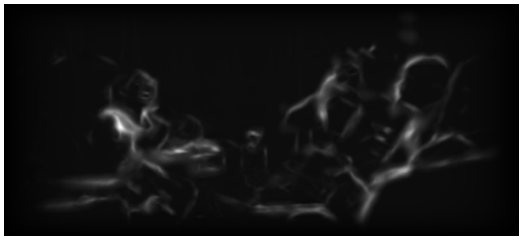
To analyze the prediction accuracy of the visual saliency model, we investigated static and dynamic pathways separately. To calculate the prediction accuracy of static pathway (respectively dynamic pathway), we compared it with the eye positions from two groups with AV and V conditions separately. The procedure of calculation was described below:

(a) original image

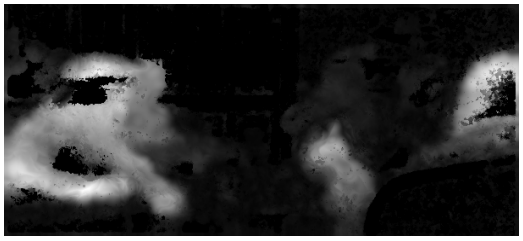
(b) human eye position density map from group with AV condition



(c) human eye position density map from group with V condition



(d) static saliency map



(e) dynamic saliency map

Figure 3.13: (a) An example of frame, (b) and (c) the experimental eye positions from groups of participants with AV and V conditions, (d) and (e) the saliency maps from static and dynamic pathways of the visual saliency model.

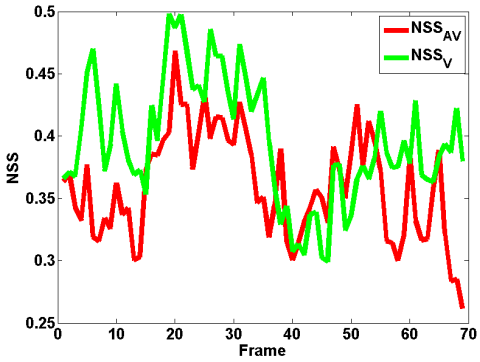
– First, for each frame, we calculate the NSS value between saliency map and eye positions from group with AV condition (respectively with V condition).

– Then, the average value of sixty clip snippets (described in Fig. 3.4) of $NSS(k)$ values are calculated over time. The NSS values are synchronized with the beginning of each clip snippet over time.

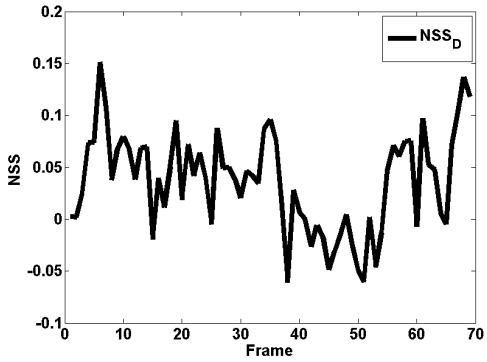
3.4.3 Comparison with static pathway

In Fig. 3.14 (a), NSS value of both with AV and V condition increased around frame 13, and decreased around frame 35. Before the decreasing, NSS_V seemed higher than NSS_{AV}. After frame 40, NSS_V and NSS_{AV} were similar. For both NSS_V and NSS_{AV}, the NSS values were small. Perhaps because of the image quality in the video data was not high, the edge of the objects in the scene was blurred. In Fig. 3.14 (b), if the NSS_D difference ($NSS_V - NSS_{AV}$) is above 0, it means the prediction accuracy for group with

AV condition is lower than the group with V condition.



(a) NSS_V and NSS_{AV} between groups of participants with AV and V conditions over time.



(b) NSS_D difference ($NSS_V - NSS_{AV}$) between groups of participants with AV and V conditions over time.

Figure 3.14: Results of prediction accuracy for static pathway, evaluated by NSS: NSS_V, NSS_{AV}, and NSS_D difference ($NSS_V - NSS_{AV}$) over time.

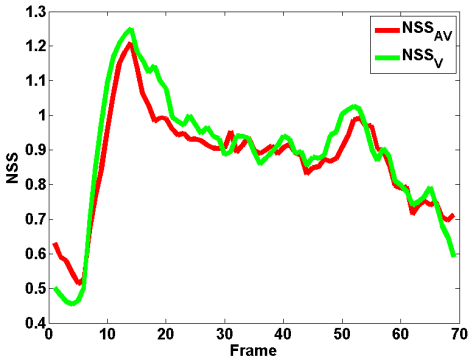
In section 3.3, we concluded that, sound effect human gaze on videos differently depending on the sound type. Hence, we calculated the NSS_D difference ($NSS_V - NSS_{AV}$) between groups with AV and V conditions in three sound classes: on-screen speech, non-speech and non-sound. To solve the problems of independent samples, also the mean of 8 continuous frames of NSS_D difference was calculated as one independent sample in the following calculation.

To determine whether the NSS_D difference is equal to 0 or not, first, the Wilcoxon signed-rank test was applied. It is the non-parametric formula alternative to the t-test for independent samples. Wilcoxon signed-rank test performs a two-sided signed rank test of the null hypothesis that data in the vector x comes from a continuous, symmetric distribution with zero median, against the alternative that the distribution does not have zero median [Wilcoxon 1945, Gibbons 2003]. The median of the on-screen speech class is significantly above 0, with the Wilcoxon signed-rank test $p < 10^{-8}$. The medians of the non-speech class ($p = 0.15$) and non-sound ($p = 0.15$) are not significantly different from 0.

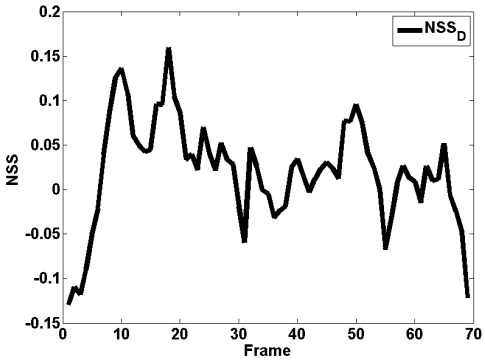
From the results above, we conclude that the accuracy of prediction from static pathway decreases in a group with AV condition compared to a group with V condition, for the on-screen speech class. (There is no significant difference for the non-speech class and for the non-sound class).

3.4.4 Comparison with dynamic pathway

In Fig. 3.15 (a), temporal behavior of NSS_V and NSS_{AV} were similar. They decreased slightly in the beginning, and then increased sharply till frame 13. This delay of increasing was caused by the reaction time of the participants. After, they decreased slowly, except a small peak around frame 50. In Fig. 3.15 (b), visually, the NSS difference is above 0 from frame 6 to 56 after the beginning of the clip snippets. The prediction accuracy seemed decreased in group with AV condition.



(a) NSS_V and NSS_{AV} between groups of participants with AV and V conditions over time.



(b) NSS_D difference ($NSS_V - NSS_{AV}$) between groups of participants with AV and V conditions over time.

Figure 3.15: Results of prediction accuracy for dynamic pathway, evaluated by NSS: NSS_V, NSS_{AV}, and NSS_D difference ($NSS_V - NSS_{AV}$) over time.

Then, we calculated the NSS_D difference ($NSS_V - NSS_{AV}$) between groups with AV and V conditions in three classes: on-screen speech, non-speech and non-sound. Also the mean of 8 continuous frames of NSS_D difference was taken into account as one independent sample.

The median of the on-screen speech class is significantly above 0, with the Wilcoxon signed-rank test $p < 10^{-6}$. The medians of the non-speech class ($p = 0.12$) and non-sound ($p = 0.18$) are not significantly different from 0. The results are very similar to those obtained from the static pathway and the conclusion is identical. Then, in the case of video with soundtrack, it would be interesting to complete the visual saliency model by a 'sound pathway'.

3.4.5 Conclusion

The prediction accuracy decreased of the saliency model proposed by Marat *et al.*, when tested on the videos with original soundtrack. The decreasing of prediction accuracy appeared both in static and dynamic pathway. Moreover, the decreasing of prediction

accuracy was different on different types of sound. For on-screen speech class, the decreasing of prediction accuracy was significant at a level of 1%. However, for other two classes: non-speech and non-sound, the prediction accuracy was not significantly different between with AV and V conditions.

3.5 Interest of a 'sound localization pathway'

In previous section, we concluded that the prediction accuracy of a visual saliency model decreased, when it applied on the video data with original soundtrack. We try to find a method to complete this visual saliency model to increase the prediction accuracy, when it used on video with soundtrack.

From our observation, the sound source in the video seems to attract human attention. To simplify the problem, we only consider the clip snippets with only one sound source in each frame. Hence, we located the coordinates of the sound source manually and called it "sound localization pathway". Then, we apply a two-dimension Gaussian to the position of the sound source to obtain a sound saliency map M_{ms}. At last, we compare with NSS the experimental data of the eye positions (groups with AV and V conditions) and the sound saliency maps (M_{ms}).

3.5.1 Selection the size of two-dimension Gaussian

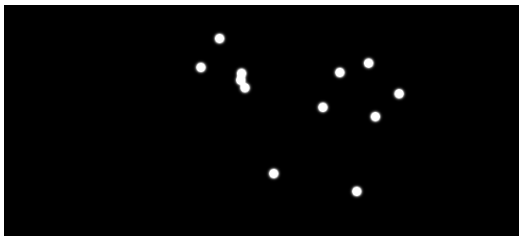
Sound saliency map M_{ms} was created by adding a two-dimension Gaussian on the sound source. First, we chose the size of standard deviation of the Gaussian with a diameter at mid-height equal to 0.5° of visual angle, which is close to the size of the maximum resolution of the fovea. This Gaussian size was the same as for human density map (in section 3.1.5).

In Fig. 3.16, we first compared the experimental eye positions from group with AV condition (respectively with V condition), which was shown in (b) (respectively in (c)), with the sound saliency map M_{ms} with the Gaussian size of a diameter at mid-height equal to 0.5° of visual angle (shown in (d)), to evaluate the prediction accuracy of this sound saliency map. The evaluation metrics was still NSS.

In Fig. 3.17 (a), curve of NSS_V and NSS_{AV} were similar, without obvious decreasing or increasing over time, and the value was much smaller than the NSS value from dynamic pathway (in Table 3.4). It maybe caused by the improper selection of Gaussian size, which did not represent the sound source region correctly. Hence, we increased the size of the two-dimension Gaussian, which was added on the sound source to create sound saliency map M_{ms}, to the diameter at mid-height equal to 1/3 of the image height. The center of this two-dimension Gaussian was the position of the sound source. An example of sound

(a) original image

(b) human eye position density map from group with AV condition



(c) human eye position density map from group with V condition

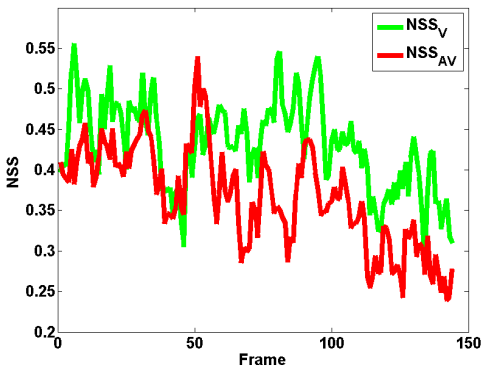
(d) sound saliency map M_{ms} with the Gaussian size of a diameter at mid-height equal to 0.5° of visual angle

(e) sound saliency map M_{ms} with the Gaussian size of a diameter at mid-height equal to 1/3 of the image height

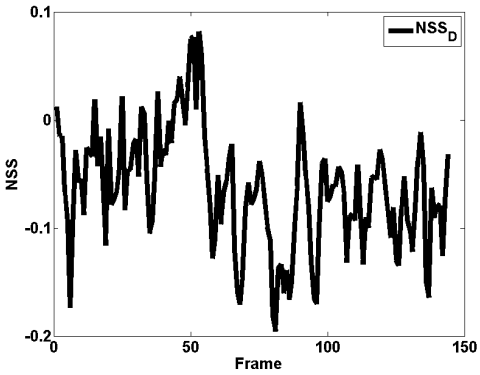
Figure 3.16: (a) An example of original frame from the video data, (b) and (c) the experimental eye positions from groups of participants with AV and V conditions, (d) and (e) the sound saliency maps M_{ms} with different size of Gaussian.

saliency map M_{ms} with the Gaussian size of a diameter at mid-height equal to 1/3 of the image height was shown in Fig. 3.16 (e).

In Fig. 3.18 (a), NSS value of group of participants with AV condition (NSS_{AV}) first decreased from the beginning to frame 6. This decreasing maybe caused by "center bias", that participants tended to watch the center of image to gather more information. After, this curve increased sharply till frame 18. It means participants moved their eyes to the salient regions predicted by sound saliency map. Then the curve decreased slowly. However, it tended to increase again after frame 80. It may be caused by "top-down" process. NSS value of group of participants with V condition (NSS_V) performed similar to NSS_{AV}, except two differences: in (b), the NSS_D showed that NSS_V was smaller than NSS_{AV} after frame 6; for NSS_V, it seemed to reach the peak at frame 12, which was a little earlier than NSS_{AV} (frame 18).

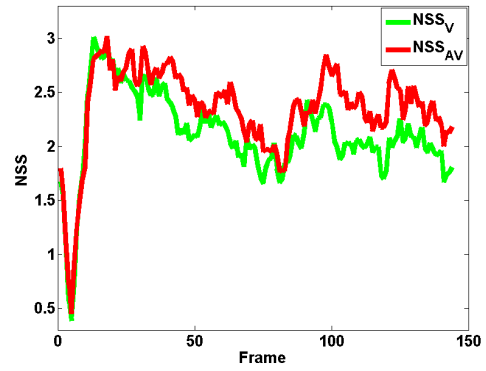


(a) NSS_V and NSS_{AV} between groups of participants with AV and V conditions over time.

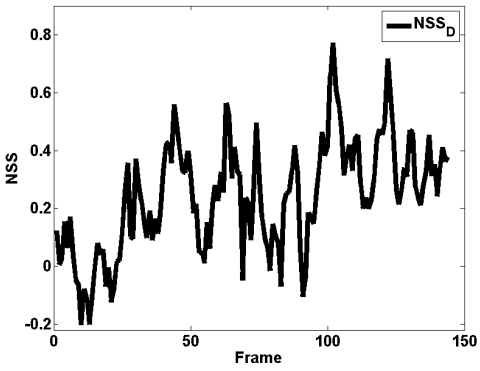


(b) NSS_D difference ($NSS_V - NSS_{AV}$) between groups of participants with AV and V conditions over time.

Figure 3.17: Results of prediction accuracy of sound saliency maps M_{ms} with the Gaussian size of a diameter at mid-height equal to 0.5° of visual angle, evaluated by NSS: NSS_V, NSS_{AV}, and NSS_D difference ($NSS_V - NSS_{AV}$) over time.



(a) NSS_V and NSS_{AV} between groups of participants with AV and V conditions over time.



(b) NSS_D difference ($NSS_V - NSS_{AV}$) between groups of participants with AV and V conditions over time.

Figure 3.18: Results of prediction accuracy of sound saliency maps M_{ms} with the Gaussian size of a diameter at mid-height equal to 1/3 of the image height, evaluated by NSS: NSS_V, NSS_{AV}, and NSS_D difference ($NSS_V - NSS_{AV}$) over time.

In Table 3.4, the NSS_{AV} for M_{ms} with Gaussian size of a diameter at mid-height equal to 1/3 of the image height increased, compared the NSS_{AV} for M_{ms} with Gaussian size of a diameter at mid-height equal to 0.5° of visual angle, from 0.4 to 2.34. For NSS_V, this increasing also existed. Because M_{ms} with Gaussian size of a diameter at mid-height equal to 1/3 of the image height was closer to the real size of object, which was considered as sound source, this size of Gaussian was proper. Most of the time, the sound source is also the moving or face region on the screen, the group without sound also obtains a high value in this model. Nevertheless, this result shows that locating the sound source is a

possible way of increasing the prediction accuracy.

Table 3.4: The mean value of NSS from frame 1 to 150 presented Fig. 3.17 (a) and Fig. 3.18 (a).

	mean of NSS_{AV}	mean ofNSS_V
M_{ms} with Gaussian size of a diameter at mid-height equal to 0.5° of visual angle	0.40	0.44
M_{ms} with Gaussian size of a diameter at mid-height equal to 1/3 of the image height	2.34	2.11

In order to test the difference of prediction accuracy of 'sound localization pathway', we calculated the NSS_D difference ($NSS_{AV} - NSS_V$) between groups with AV and V conditions in on-screen speech class and non-speech class. Because non-sound class had no sound source in the screen, we did not consider this class.

In Fig. 3.19, the median of the on-screen speech class is significantly above 0, with the Wilcoxon signed-rank test $p < 10^{-6}$. The median of the non-speech class (p=0.18) is not significantly different from 0. From this, we conclude that the accuracy of prediction from 'sound localization pathway' increases in the group with AV condition compared to the group with V condition, for the on-screen speech class.

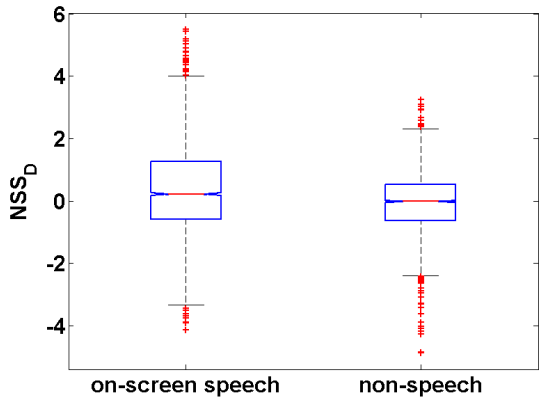


Figure 3.19: Kruskal-Wallis test of NSS_D difference ($NSS_{AV} - NSS_V$) between groups with AV and V conditions for 'sound localization pathway' in two classes (on-screen speech and non-speech).

3.5.2 Conclusion

The prediction accuracy of saliency model can be increased by adding a sound pathway, by locating the sound source. The sound saliency map, created by adding a proper size

of two-dimension Gaussian on the sound source, increased the prediction accuracy in AV condition than in V condition. This increasing was significant, when sound is on-screen speech. However, for non-speech class, the prediction accuracy of this sound saliency map was not significantly different between AV and V conditions.

3.6 General conclusion

This chapter presented an audio-visual experiment to investigate the sound effect on human gaze when looking freely at videos. Through the analysis of difference of the eye positions between group with AV condition and with V condition, we concluded that sound affected human gaze differently depending on the sound type, and the effect was greater for the on-screen speech class. When tested with the visual saliency model proposed by Marat *et al.*, the accuracy of prediction decreased in group with AV condition compared to group with V condition in on-screen speech class. However for the other two classes: non-speech and non-sound, the accuracy of prediction did not decrease between group with AV and V conditions. Locating the sound source as a 'sound localization pathway', the prediction accuracy was higher in group with AV condition than in group with V condition.

Chapter 4

Audio-visual experiment II

In the previous chapter, through the analysis of the first audio-visual experiment, we observed that sound influences human gaze in videos differently depending on the sound type, and the effect is greater for the on-screen speech class. We only considered three sound classes and no strict control of sound event over time.

To investigate the sound influence on gaze in videos deeply, a second audio-visual experiment is presented in this chapter to answer the question of which type of sound influences human gaze. We compare the behavior of human gaze in relation to thirteen more refined sound classes. The videos excerpts are chosen so that the onset of each relevant sound occurs in the middle of a visual scene. In this way, we avoid the content of visual scene and soundtrack changing at the same time. The aim is to isolate the effect of sound by comparing the eye positions in AV and V conditions.

Hence, we designed a new audio-visual experiment of two groups of participants with audio-visual (AV) and visual (V) conditions. Then, we compared the difference of eye positions from the group with AV condition and the group with V condition of the thirteen sound classes separately. To find out where humans look after the onset of auditory stimuli, we analyzed the distance between sound source and eye positions. Then fixation duration between groups with AV and V conditions are analyzed. Finally, the experimental eye positions are compared with the prediction regions of two visual saliency models (Marat's *et al.* and Itti's *et al.*). A part of the results has been published in [Song 2012].

4.1 Audio-visual experiment design

This audio-visual experiment is designed to investigate which type of sound influences gaze in videos through observing the eye positions from the participants. The principle of the experiment design is that each participant watch half of the video excerpts with original soundtrack (AV condition) and the other half video excerpts without soundtrack

(V condition). Then, we investigate the sound effect through analyzing the eye positions differences between these two groups of participants.

The same apparatus –Eyelink II was used to track and record the eye positions of the participants as in previous experiment. The principal technical specifications of Eyelink II in this experiment are shown in Table 3.1. During the experiment, the sampling frequency of the system recording the eye positions is 250 Hz.

4.1.1 Participants

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

In the first audio-visual experiment, a participant viewed all the clips with only one condition – AV or V. However in this new experiment, a participant viewed the clips with two conditions – AV and V. With this design, the effect of the difference of fixation duration caused by individual participant has been reduced. Moreover, for each individual participant, we obtain related eye positions – with AV and V conditions. This type of data can be measured by paired t-test and mixed-effect model (section 4.5).

4.1.2 Materials

In this experiment, eighty video excerpts are chosen from heterogeneous sources of films (with original soundtrack and visual scene). Each clip snippet lasts around 200 frames (8 seconds). The sum of all the clip snippets is 16402 frames (around 11 minutes). All the clip snippets are converted to the same video format (25 fps, 842×474 pixels/frame). In the visual domain, each clip snippet consists of just one shot. In the audio domain, the sound signal is divided into two parts. The first sound lasts to about the middle of the clip snippet, and is then followed by the second sound. To reduce the difference of sound amplitude between clip snippets, while keeping the difference of sound amplitude into each clip snippet, we increase of 25% the sound amplitude of the clip snippets, which are lower than the mean of all the clip snippets. Respectively, we decrease of 25% the sound amplitude of the clip snippets, which are higher than the mean.

In order to prevent the participants from understanding the language in the video, we chose foreign languages for each participant, like Chinese, Indian, Japanese, etc.. A clip

snippet example is presented in Fig. A.6. The eighty clip snippets are then recombined into ten clips [Carmi 2006], each clip being the concatenation of eight clip snippets from different film sources and different sound classes of the second sound. We use gray level stimuli. Two sets of stimuli are built from these clips, one with AV condition (frames + soundtrack), and the other one with V condition (frames only).

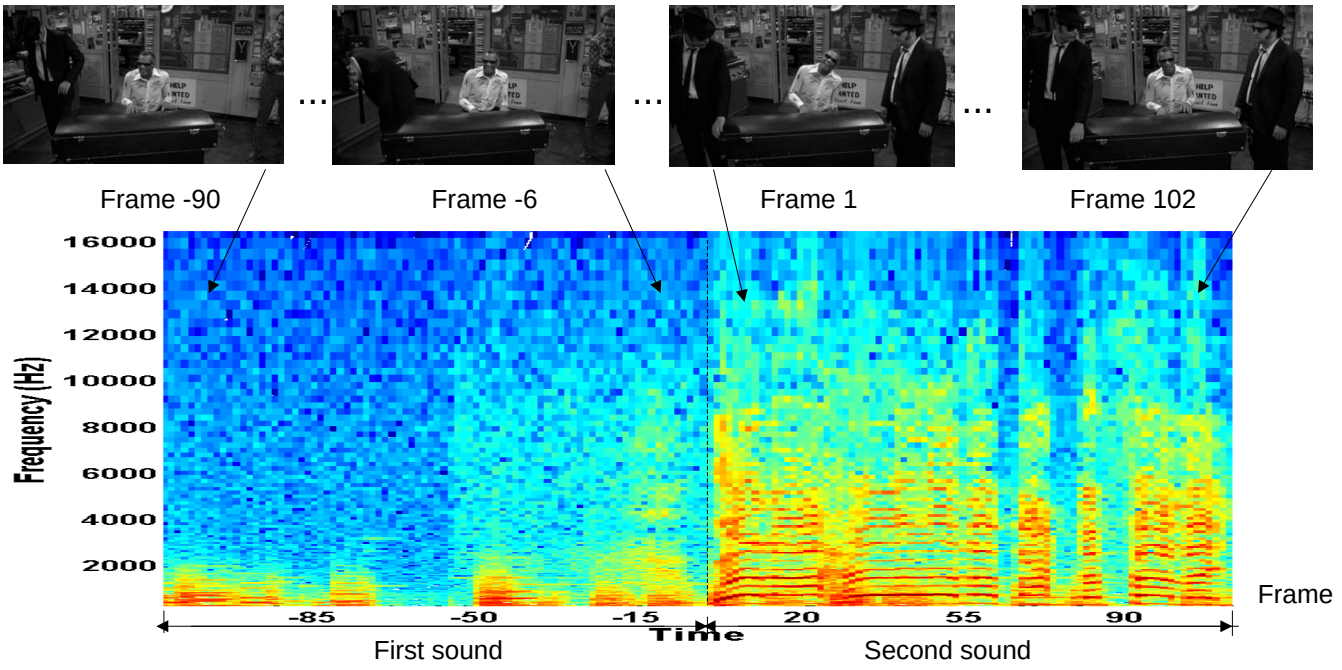


Figure 4.1: An example of some frames of a clip snippet with the associated soundtrack. The soundtrack is a succession of two types of sound. In this example, the first sound is the man in the center playing piano, and the second sound is the man in the center singing.

Here, we only observe the behavior of human gaze after the second sound. The aim is to analyze the effect of an audio change unrelated to the visual changes that occur when a new clip snippet starts. Compared to the first experiment, in this second experiment, the first sound lasted at least two seconds before the second sound occurs, which was enough to avoid *center bias*.

Presentation of center bias

Some studies using static images showed that the center of the image was an attractor for the participants on the first fixations [Parkhurst 2002, Tseng 2009]. This so called "center bias" also influences eye movement when viewing dynamic scenes [Dorr 2010]. This bias is generally explained by several factors:

- The video excerpts used in the experiment were selected from films, which was made by professionals, who tend to put interesting objects in the center of the images.
- The video excerpts were presented in central vision of the participants, whose eye movements were recorded by an eye-tracker.
- Before the presentation of each clip, the participants were asked to watch the fixation cross in the center of the screen.

4.1.3 Procedure

Human eye positions were tracked by an eye tracker-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 vision was $35° \times 20°$. A headphone carried the monophonic sound. Compared to two loudspeakers used in experiment I, the headphone can reduce the influence of possible noise from the environment, which may distract the participants during the experiment. A 9-point calibration was carried out every five clips. 10 clips were 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.

4.2 Pre-experiment: validation of sound classification

Based on another research [Niessen 2008], we classified the second sound into thirteen classes (see Fig. A.5). For each class, there were 5 to 11 clip snippets. Fig. 4.3 shows the examples of clip snippet in each sound class. Numbers of clip snippets and frames in each class are given in Table 4.1.

The difference between clusters of classes "on-screen with one sound source" and "on-screen with more than one sound source" was the number of sound sources on the screen. Here, we called one *sound source* a visual event in the scene associated with the soundtrack. In this instance the sound can be associated with a spatial location. The "off-screen sound source" group was different from the other two in that there was no sound source on the screen when the second sound appeared. This classification was carried out manually by the author. In order to validate whether this classification was proper or not, we proposed a pre-experiment.

Table 4.1: Number of clip snippets and frames in each class

sound class	number of snippets	number of frames
speech	11	2729
singer	5	790
human noise	6	1087
animal	5	1054
music	7	1140
action	6	1309
impact and explosion	8	1832
vehicles and mechanics	6	1119
singers	5	928
animals	5	898
actions	6	1110
voice-over	5	1352
background music	5	1054
total	80	16402

4.2.1 Pre-experiment design

Apparatus

The sound database was carried out by a headphone (the same condition as in experiment II), and the presentation of the sound classifications and the recording of the answer from the participants were done by a psychology software tool named E-prime. [1]

Participants

Five participants heard the eighty sound excerpts with a random order in this pre-experiment. All participants had reported normal hearing.

Materials

One sound excerpt in this pre-experiment is extracted from one clip snippet. Because we want to classify the second sound in each clip snippet, we cut the audio stream of the clip snippet from the second sound as one sound excerpt. To reduce the impact of the transform from mute to second sound, we keep 200 ms of first sound before the second sound. The crescendo of this 200 ms duration is continuous till the appearance of the second sound. Hence, the total duration for each sound is more than 800 ms.

When only the sounds are presented (without images), it is impossible to know if they are off-screen sound source. We decided to propose eleven sound classes to the participants

[1] E-Prime is a suite of applications to fulfill computerized experiment needs, which provides an easy-to-use environment for computerized experiment design, data collection, and analysis [EprimeSite].

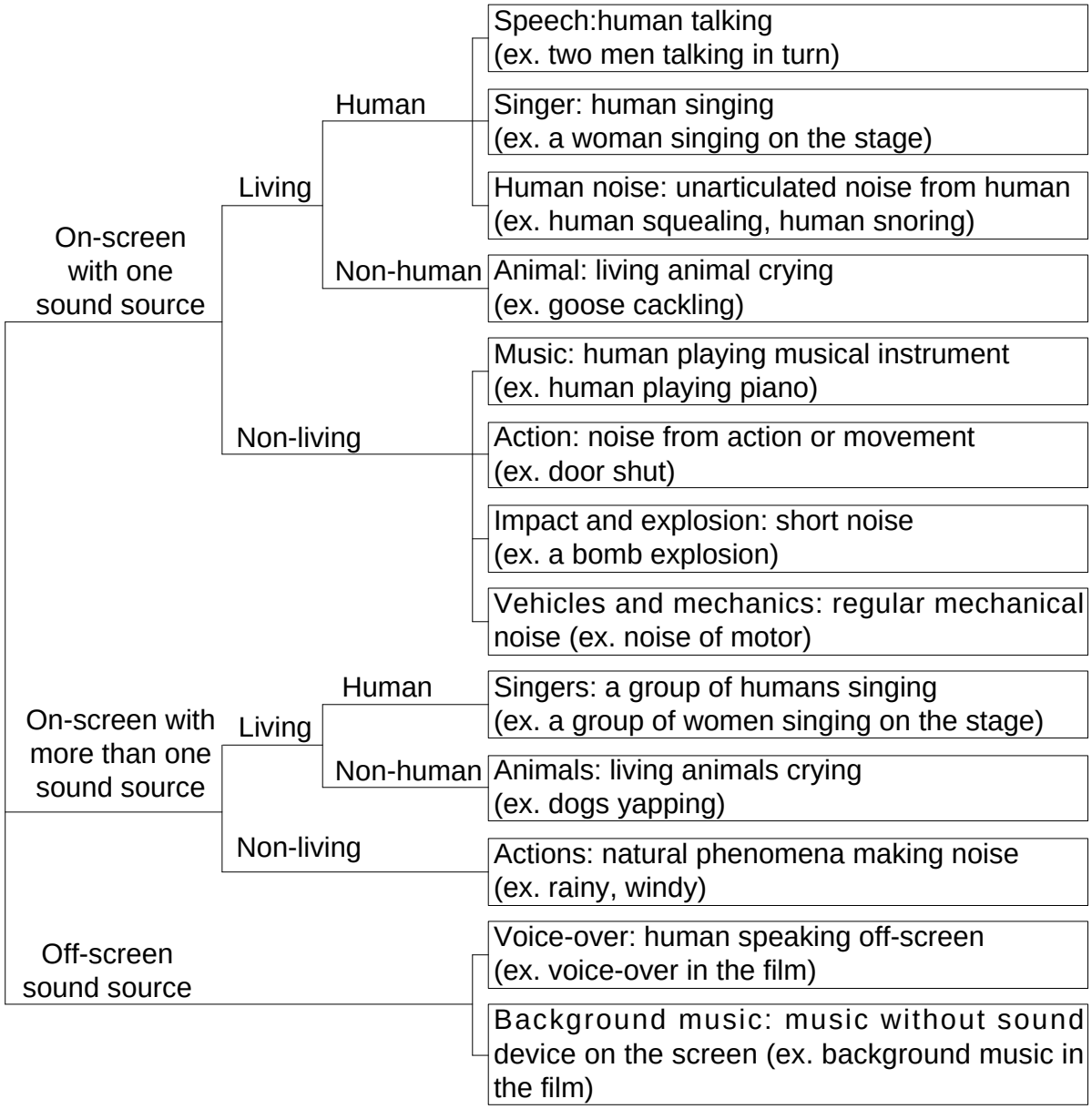


Figure 4.2: Classification of the second sound

to select from. Compared to the classes shown in Fig. A.5, classes of "speech" and "voice-over" composed class "speech", and classes of "music" and "back-ground music" composed class "music".

Procedure

Before the participants started the experiment, they were informed of the description of the eleven sound classes proposed in the list. The participants first heard a sound excerpt

(a) speech

(b) singer

(c) human noise

(d) animal

(e) music

(f) action

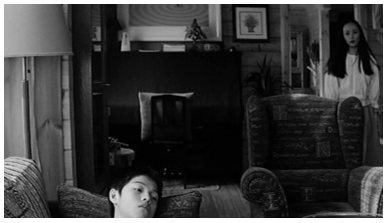
(g) impact and explosion

(h) vehicles and mechanics

(i) singers

(j) animals

(k) actions



(l) voice-over

(m) background music

Figure 4.3: Example of frames of the thirteen sound classes.

from a headphone, then chose one of the classes proposed on the screen. This procedure repeated until all the sounds have been presented with a random order.

4.2.2 Result

In order to measure whether the classification proposed by the author was proper or not, we calculated the *correct classification rate* of each sound class, which was described as:

$$\text{Correct classification rate} = \frac{n_c}{n_p \times n_s} \tag{4.1}$$

where, n_c is the number of "correct classification" in this sound class. If the participants classify the sound excerpt to the same sound classification as the author proposed, it is considered as one "correct classification". n_p is the number of participants, here the number is 5, and n_s is the number of sound excerpts in the sound class (from 5 to 16).

In Fig. 4.4, the minimal correct classification rate is 80% ("impact and explosion", "vehicles and mechanical" and "animals" classes), and the mean correct classification rate is 90%. Hence, we can conclude that the classification is suitable for the audio-visual experiment.

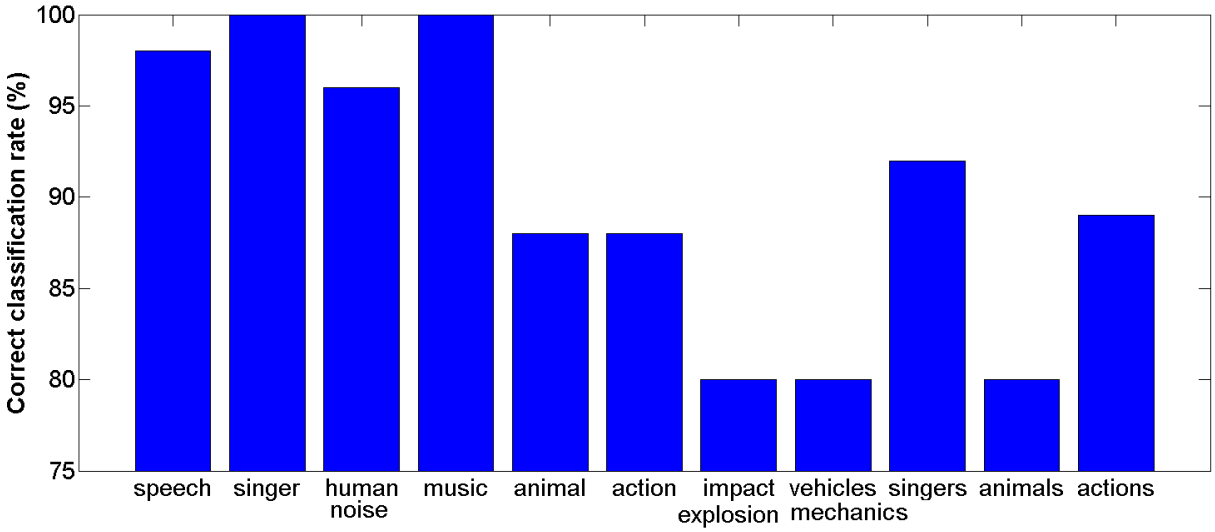


Figure 4.4: Correct classification rate for each sound class.

4.3 Analysis of eye position difference between groups with AV and V conditions

In order to investigate the effect of sound on visual gaze, we compared the eye positions of participants with AV condition to the participants with V condition. In the following, we will present the details of this comparison. Fig. 4.5 illustrates by an example the different eye positions of the two groups of participants for the clip snippet shown in Fig. A.6.

(a) Frame 86 (b) Frame 94

Figure 4.5: Eye positions are illustrated for participants, with AV condition (red points) and V condition (green points) from singer class, presented previously in Fig. A.6. Frame 86 is just before the appearance of second sound.

4.3.1 Criteria

In group of participants with AV condition (respectively with V condition), if one eye position is far from the others, we consider it as an outlier in this group. To measure whether one eye position is an outlier, we introduce squared Mahalanobis distance to remove the outliers of the eye positions in each group.

Presentation of Mahalanobis distance

First, we applied squared Mahalanobis distance to the eye positions in each group (with AV condition, respectively V condition) to remove the outliers of each frame. The Mahalanobis distance is a measure of how much the value of a case differs from the average of all other cases. Large Mahalanobis distances signify potential outlier cases [Ntoumanis 2005]. This distance is described as:

$$D_M(i) = (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \tag{4.2}$$

where for the participant i, μ is the mean of the data (except participant i), and Σ is the covariance of the data (except participant i).

An image point has two degrees of freedom, so a threshold of 9.21 on the Mahalanobis distance sets the confidence level to 99%. It means that if the Mahalanobis distance between one eye position and the others is more than 9.21, this eye position has 99% probability of being an outlier. We removed the potential outlier eye positions before the following calculations.

In the following, three metrics are presented: distance d, Kullback-Leibler divergence KLD and linear correlation coefficient cc.

Criterion of distance d

Then, in order to measure the distance of eye positions between the two groups (with AV and V conditions) for each frame, we adopted the criterion named mean distance d, which is defined as:

$$d = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n'} d_{i,j}}{n \times n'}, i \in \mathscr{N}, j \in \mathscr{N}' \tag{4.3}$$

In the above equation, $\mathscr{N}$ is the group with AV condition, and the number of participants in this group is n. $\mathscr{N}'$ is the group with V condition, and the number of participants in this group is n'. $d_{i,j}$ is the Euclidean distance between eye positions of participants i and j, who belong respectively to the group with AV condition and the group with V condition.

Criterion of Kullback-Leibler Divergence KLD

To confirm the measurement, another criterion, which is used to estimate the difference between two probability distributions, named Kullback-Leibler divergence is calculated. The Kullback-Leibler divergence criterion was already adopted to compare distributions of eye positions between groups by other researchers, such as [Tatler 2005]. With the Kullback-Leibler divergence calculation, we compared the experimental eye position density maps (calculated by 3.1) between group of participants with AV and V conditions. For a given frame, a 2-D Gaussian (equal to 2° wide) was added to each eye position in the density map of group of participants with AV condition (M_{hav}), respectively with V condition (M_{hv}). Here, we use symmetric Kullback-Leibler Divergence (KLD). For each frame, we calculated the following equation:

$$KLD(M_{hav}, M_{hv}) = \frac{1}{2}\left(\sum_{i=1}^{p} M_{hav} log\frac{M_{hav}}{M_{hv}} + \sum_{i=1}^{p} M_{hv} log\frac{M_{hv}}{M_{hav}}\right) \tag{4.4}$$

where p represents the same size of video frame (842×474). High KLD values represent high differences between two distributions of eye positions.

Criterion of linear correlation coefficient cc

The third metric we adopted is the linear correlation coefficient, noted as cc. cc describes straight-line relationships between two variables. cc was introduced in detail in section 3.3). For cc, a value of zero indicates no linear relationship between the two maps: there is no correspondence between the eye positions of the two groups with AV and V conditions, and higher values of cc indicate higher correspondence between the eye positions of the

two groups.

4.3.2 Comparison among different clusters of sound classes

We analyzed the mean distance d (then KLD and cc) between the eye positions of the participants in the two groups with AV and V conditions, among three clusters of classes (see Fig. A.5): "on-screen with one sound source", "on-screen with more than one sound source" and "off-screen sound source".

In this section, for each clip snippet, we took 25 frames (from frame 6 to 30, to eliminate reaction time of about 5 frames) after the beginning of the second sound. We used the ANOVA test to compare distance d among different clusters of classes. This test requires the samples in each cluster to be independent samples. Because we consider continuous measurement over time, the eye positions for most participants does not change much between two adjacent frames, they could not be considered as independent samples. To solve this problem, we assume a set of continuous frames to be one independent sample. For the size of the set, we choose the average value of one fixation duration (about 7 frames, see in Section. 4.5). We compute a distance d for each frame, and subsample by computing the mean of 8 adjacent frames (for margin) as an independent sample.

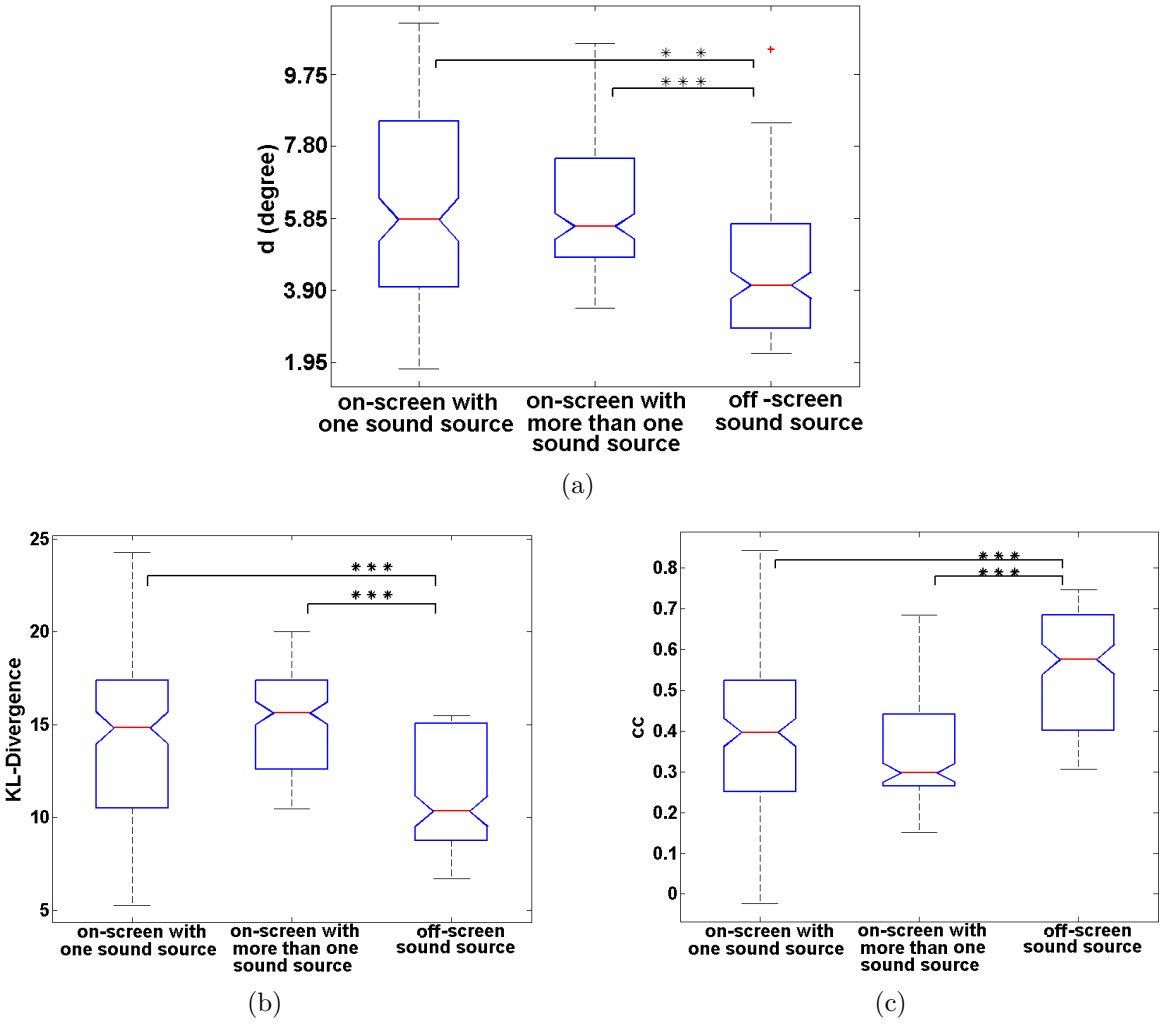
In Fig. A.7 (a), with ANOVA test, "off-screen sound source" presents the lowest d among the three clusters of classes. The difference is significant, between "on-screen with one sound source" and "off-screen sound source" ($F(1, 175) = 7.94$, $p < 10^{-2}$), and also significant between "on-screen with more than one sound source" and "off-screen sound source" ($F(1, 73) = 8.69$, $p < 10^{-3}$). The difference between "on-screen with one sound source" and "on-screen with more than one sound source" is not significantly different ($F(1, 184) = 0.12$, $p = 0.73$).

This result is confirmed by other two criteria: KLD (Fig. A.7(b))and cc (Fig. A.7(c)). For KLD the difference is significant, between "on-screen with one sound source" and "off-screen sound source" ($F(1, 175) = 13.97$, $p < 10^{-3}$), and also significant between "on-screen with more than one sound source" and "off-screen sound source" ($F(1, 73) = 35.56$, $p < 10^{-7}$). The difference between "on-screen with one sound source" and "on-screen with more than one sound source" is not significantly different ($F(1, 184) = 1.77$, $p = 0.19$).

For cc, the difference is significant between "on-screen with one sound source" and "off-screen sound source" ($F(1, 175) = 22.53$, $p < 10^{-5}$), and also significant between "on-screen with more than one sound source" and "off-screen sound source" ($F(1, 73) = 31.24$, $p < 10^{-6}$). The difference between "on-screen with one sound source" and "on-screen with more than one sound source" is not significantly different ($F(1, 184) = 0.86$, $p = 0.36$).

These results indicated that a localizable sound source leading to a greater distance between the groups with AV and V conditions, suggesting that the presence of a localizable

sound source influences human gaze owing to the auditory and visual interaction.



(a)

(b)

(c)

Figure 4.6: Criteria of mean distance d, Kullback-Leibler Divergence KLD and linear correlation coefficient cc 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 "off-screen sound source". Larger d, KLD and smaller cc values represent greater difference between groups with AV and V conditions.

4.3.3 Analysis of thirteen sound classes separately

Secondly, we analyzed the thirteen sound classes separately. We did not analyze 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.

Presentation of randomization tests

Randomization tests are a subclass of statistical tests called permutation tests. In permutation tests, the *p-value* is the proportion of data permutations or configurations providing a statistical test as large as (or as small as) the value for the research results. Randomization tests are permutation tests for randomized experiments. They test null hypotheses about the treatment effects on a random assignment of research units [Edgington 2007].

Randomization tests are superior to parametric tests (such as ANOVA) in several aspects:

- There is no requirement that the random samples are from one or more populations, and there is no need to assume the normality of the distribution of the sample data.
- Because randomization tests are not concerned on populations, it is not necessary to concentrate on estimating (or even testing) characteristics of those populations.
- The null hypothesis of randomization test has nothing to do with parameters. Under this kind of null hypothesis, the score that is associated with a participant is independent of the treatment that person received.
- Even more, compare to parametric tests, randomization tests emphasize the importance of random assignment of participants to treatments.

Approach of a randomization test

The approach to realize a randomization test on the eye position data is presented below:

- First, we extracted 18 participants (total number of participants is 36) randomly from groups with AV and V conditions to create a new group called G1. The rest of the participants formed another new group, called G2.
- Afterwards, we calculated the d between G1 and G2 for each frame.
- We repeated this procedure 5000 times, we obtained:
  - For each frame, a distribution of 5000 random d values ($d_i, i = 1, 2, ...5000$).
  - The mean of the 5000 d values as the reference (d_R): this is an estimate of the distance that can be expected between the two random groups of participants.
- Finally, we calculated the difference ($d_{AVV} - d_R$) where d_{AVV} represents the distance between participants with AV and V conditions.

Results of the randomization test for distance d

Fig. 4.7 explains how to calculate the average value of eighty clip snippets, synchronized with the beginning of the second sound. For example, the value of d_{AVV} in Fig. A.8 of frame 1, was the mean of d_{AVV} value of all the first frames of the second sound in snippets.

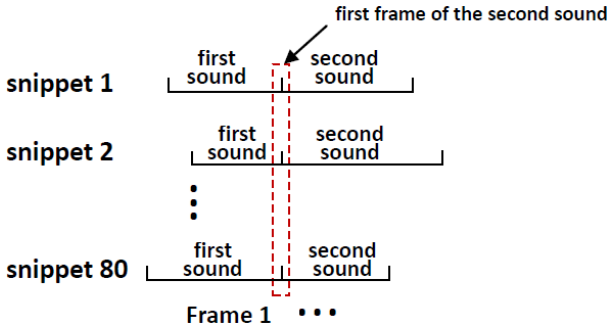


Figure 4.7: Explanation of how to calculate the average value of eighty clip snippets, synchronized with the starting frame of the second sound in each snippet.

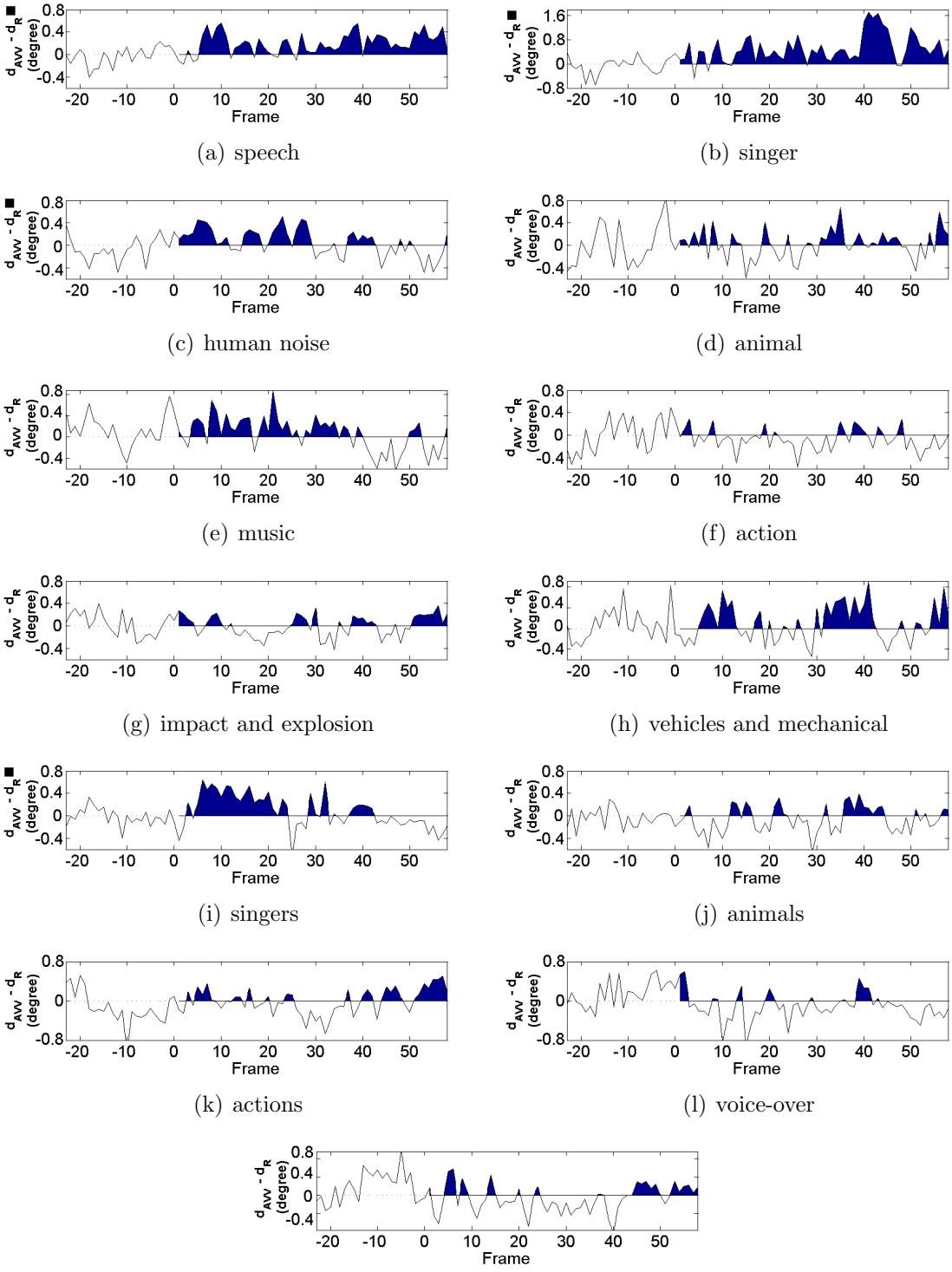
Because d_{AVV} is caused by the effect of both image and sound, and d_R is caused by the effect of image only, the difference $(d_{AVV} - d_R)$ is mainly caused by the effect of sound. Fig. A.8 shows the difference over time between d_{AVV} and d_R for the thirteen sound classes. If $(d_{AVV} - d_R)$ is over 0, the distance between AV and V groups is greater than that between the two random groups. Visually, different sound classes performed differently.

To find out which classes give the highest difference between d_{AVV} and d_R:

- First, we analyze a duration of one second (25 frames). To deduct the reaction time of the participants, we took a duration from frame 6 to 30 after the beginning of the second sound.
- Then, we compared $\overline{d}_{AVV}$ (the temporal mean of d_{AVV} over the 25 frames) to the distribution of $\overline{d}_i$, where $\overline{d}_i$ is the temporal mean of d_i between G1 and G2 over the 25 frames for the random trial i.
- To estimate the probability of $\overline{d}_i$ being greater than $\overline{d}_{AVV}$, we calculated $p = n/5000$ where n is the number of $\overline{d}_i$ which are greater than $\overline{d}_{AVV}$.

Table 4.2 shows the results from frame 6 to 30 after the beginning of the second sound. The high $\overline{d}_{AVV}$ values (therefore low p values) for the marked classes (with ■): speech, singer, human noise, and singers, show that human voice had the greatest effect on visual gaze.

To verify that the effect measured above is really due to the second sound, we perform the same calculation for a period of one second (25 frames) before the beginning of the



(a) speech (b) singer

(c) human noise (d) animal

(e) music (f) action

(g) impact and explosion (h) vehicles and mechanical

(i) singers (j) animals

(k) actions (l) voice-over

(m) background music

Figure 4.8: Difference $(d_{AVV} - d_R)$ over time for the thirteen sound classes. Frame 1 is the beginning of the second sound.

second sound. Results of probability estimations of $\overline{d}_i$ values higher than $\overline{d}_{AVV}$ of all the sound classes from frame -24 to 0 are higher than 0.1, suggesting that before the second

Table 4.2: Probability estimations of $\overline{d}_i$ values higher than $\overline{d}_{AVV}$ from frame 6 to 30 after the beginning of the second sound

sound class	p	sound class	p
speech ■	0.003	singers ■	0.006
singer ■	0.002	animals	0.789
human noise ■	0.023	actions	0.476
animal	0.763	voice-over	0.996
music	0.061	background music	0.606
action	0.857		
impact and explosion	0.892		
vehicles and mechanics	0.161		

sound, eye position of participants between groups with AV and V conditions are not significantly different for all the sound classes.

Results of the randomization test for Kullback-Leibler Divergence KLD and linear correlation coefficient cc

To confirm the result from distance d, Kulback-Leibler Divergence KLD and linear correlation coefficient cc were calculated with the same randomization method as introduced above. Table 4.3 and Table 4.4 show the results from frame 6 to 30 after the beginning of the second sound. The high $\overline{KLD}_{AVV}$ values or the lower $\overline{cc}_{AVV}$ values (therefore low p values) for the marked classes (with ■): speech, singer, human noise, and singers, confirmed that human voice had the greatest effect on visual gaze.

Table 4.3: Probability estimations of $\overline{KLD}_i$ values higher than $\overline{KLD}_{AVV}$ from frame 6 to 30 after the beginning of the second sound

sound class	p	sound class	p
speech ■	0	singers ■	0.002
singer ■	0.001	animals	0.138
human noise ■	0.001	actions	0.261
animal	0.113	voice-over	0.779
music	0.394	background music	0.895
action	0.215		
impact and explosion	0.792		
vehicles and mechanics	0.137		

4.3.4 Conclusion

Through the statistical analysis of ANOVA and randomization test, both results suggested that the difference of eye positions from group with AV and V conditions were different

Table 4.4: Probability estimations of $\overline{cc}_i$ values lower than $\overline{cc}_{AVV}$ from frame 6 to 30 after the beginning of the second sound

sound class	p	sound class	p
speech ■	0	singers ■	0.01
singer ■	0.022	animals	0.430
human noise ■	0.006	actions	0.366
animal	0.067	voice-over	0.592
music	0.126	background music	0.558
action	0.088		
impact and explosion	0.232		
vehicles and mechanics	0.194		

among thirteen classes. The differences were significant only for human voice cluster: speech, singer, human noise and singers classes.

4.4 Analysis of distance between sound source and eye positions

In the previous section, we showed that the distance d between eye positions of participants with AV and V conditions is greater for speech, singer, human noise and singers classes than others. In this section, we will verify the assumption that participants with AV condition moved their eye to the sound source after the beginning of the second sound.

4.4.1 Analysis of eight sound classes separately

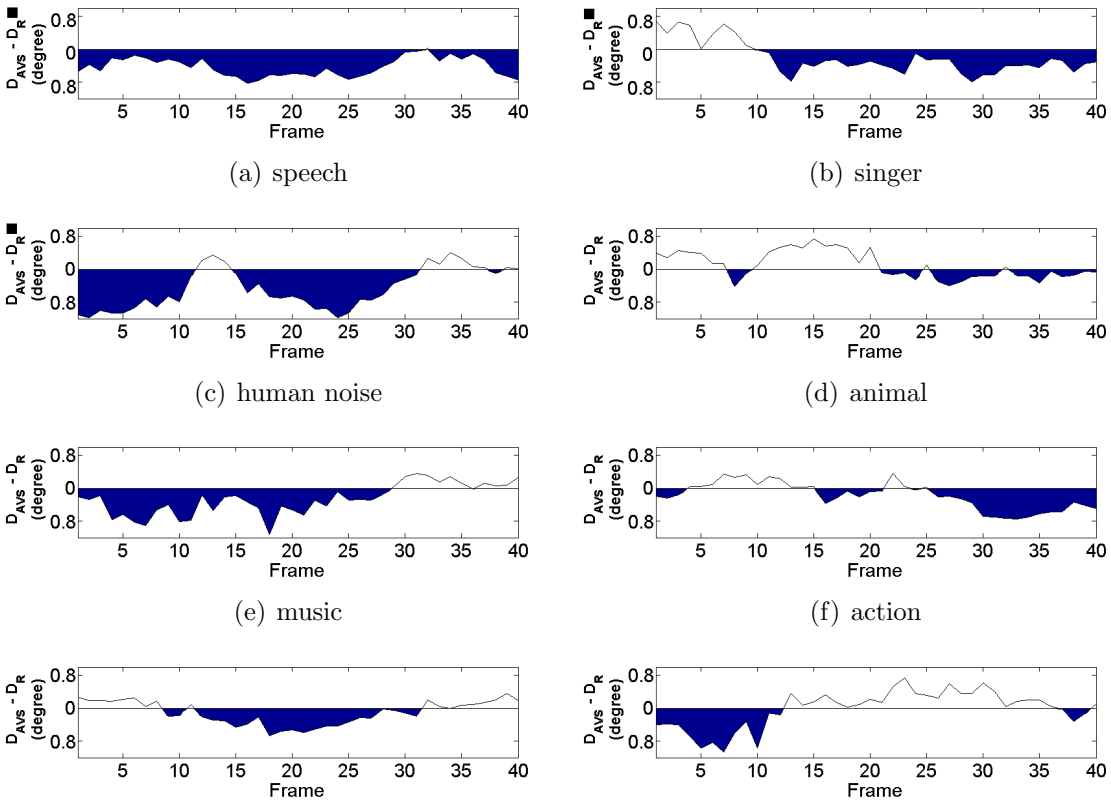
For the reason that it is hard to locate sound source in "on-screen with more than one sound source" and "off-screen sound source" clusters, we only analyze the eight sound classes in "on-screen with one sound source" cluster. To this end, the procedure was described below:

- First, we located the approximate coordinates of the center of the sound source manually.
- Then, for each frame, we calculated the Euclidean distance between the eye position of each participant with AV condition and the sound source. The mean of the Euclidean distances for all the participants gives the D_{AVS} value.
- Similarly, as before, for each randomization, we considered the mean Euclidean distance between eye position of participants of G1 and sound source ($D_i, i = 1, 2...5000$).

– We took the mean of 5000 distance values as the reference (D_R), which was affected only by image information.

– Afterwards, for each frame, we calculated $D_{AVS} - D_R$ for all the classes with one sound source. This difference reflects the influence of the sound information.

The difference of D_{AVS} and D_R for all the classes with one sound source over time are shown in Fig. 4.9. If the values of $D_{AVS} - D_R$ are negative, the group with AV condition is closer to the sound source than the random group.



(a) speech (b) singer

(c) human noise (d) animal

(e) music (f) action

(g) impact and explosion (h) vehicles and mechanics

Figure 4.9: Difference ($D_{AVS} - D_R$) over time for eight sound classes in "on-screen with one sound source" cluster.

Visually, in Fig. 4.9, different sound classes perform differently. To find out which classes give the highest difference between D_{AVS} and D_R, we took the same duration of one second (25 frames) in the previous analysis, from frame 6 to 30 after the beginning of the second sound:

– We compared $\overline{D}_{AVS}$ (the mean of D_{AVS} over the 25 frames) to the distribution of $\overline{D}_i$ ($i = 1, 2, ...5000$), where $\overline{D}_i$ was the mean of D_i between G1 and sound source over the 25 frames for the random trial i.

– To estimate the probability of $\overline{D}_i$ being smaller than $\overline{D}_{AVS}$, we calculated $p = n/5000$ where n is the number of $\overline{D}_i$ which are smaller than $\overline{D}_{AVS}$.

In Table 4.5, $\overline{D}_i$ is smaller than $\overline{D}_{AVS}$ ($p < 0.05$), from frame 6 to 30 after the beginning of the second sound, for speech, singer, human noise classes (marked with ■) suggesting that participants tend to move their eyes to the sound source only when they hear human voice.

Table 4.5: Probability estimation of $\overline{D}_i$ being smaller than $\overline{D}_{AVS}$ from frame 6 to 30 after the beginning of the second sound

sound class	p	sound class	p
speech ■	0.041	music	0.058
singer ■	0.039	action	0.292
human noise ■	0.002	impact and explosion	0.062
animal	0.283	vehicles and mechanics	0.849

4.4.2 Qualitative analysis of music class

The result in Table 4.5 shows that the probability of $\overline{D}_i$ being smaller than $\overline{D}_{AVS}$ for music class is low (0.058), suggesting that participants have a tendency to move their eyes to the sound source, but not as significant as human voice classes. We have 5 music clip snippets, among these clip snippets, 4 snippets are humans playing musical instruments. Example frames of 4 clip snippets are shown in Fig. 4.10.

In the previous calculation, we considered the musical instruments to be the sound source. However, we may also assume that the instrument player's face is more attractive compared to the instrument itself because of its importance for visual attention [Langton 2008]. So, here is the question to be investigated: between the face of the person who is playing the instrument and the instrument, which attracts human gaze?

We took the same duration of one second (25 frames) as above, from frame 6 to 30 after the beginning of the second sound. The probability of $\overline{D_{Mi}}$ (D_{Mi} is the mean Euclidean distance between G1 and Musical instrument) being smaller than $\overline{D}_{AVM}$ (D_{AVM} is the Euclidean distance between eye position of participants with AV condition and sound source — Musical instrument) is $p = 0.164$. The probability of $\overline{D_{Fi}}$ (D_{Fi} is the mean Euclidean distance between G1 and Face of the player) being smaller than $\overline{D}_{AVF}$ (D_{AVF} is Euclidean distance between eye position of participants with AV condition and the attention-getting region — Face of the player) is $p = 0.042$.

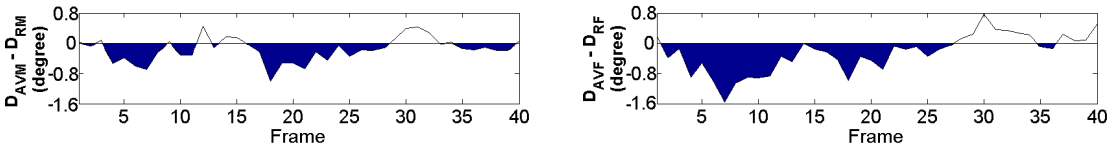
Fig. 4.11 illustrates the distances from group with AV condition to musical instrument (a) and to player's face (b) over time. Here, the dark regions below zero represent smaller

(a) (b)

(c) (d)

Figure 4.10: Example frames of 4 clip snippets in music class.

distances from a face or a musical instrument, that is, participants looked at both of them. Furthermore, a face was reached more at the scene onset, as shown in Fig. 4.11 (b). Afterwards, both the face and the instrument were reached somewhat equally.



(a) Musical instrument is salient region. (b) Player's face is salient region.

Figure 4.11: Average distance ((a) distance from Musical instrument (D_{AVM}-D_{RM}), and (b) distance from Face of the player (D_{AVF}-D_{RF})) for 4 clip snippets of music class over time.

4.4.3 Conclusion

Previous section concludes that the difference of eye positions from group with AV and V conditions is significant only for human voice cluster: speech, singer, human noise and singers classes. For this human voice cluster, the participants with AV condition tend to move their eyes to the sound source in the scene after hearing human voice. For music class, participants tend to move their eyes first to the player's face rather than the sound source –musical instrument for the first gaze after hearing the music.

4.5 Analysis of fixation duration

4.5.1 Using paired t-test

We also investigated the effect of sound on fixation duration. For each participant, we calculated the mean of fixation duration for each clip. Besides of the influence of sound (AV and V condition), there were other two influence factors: different participants and different clip content. In order to reduce the influence of these two factors, first model we adopted was paired t-test (by clip and by participant).

Presentation of paired t-test

A paired t-test is used to compare the means of two population, that the observations in one sample can be paired with observations in the other sample [Zimmerman 1997]. There are two examples of where this might occur:

- Before-and-after observations on the same participants (by participant).
- A comparison of two different methods of measurement or two different treatments, when the measurements (or treatments) are applied to the same participant.

Compare to typical t-test, which requires the samples in the data should be independent from each other, paired t-test consist of a sample of matched pairs of similar units, or one group of units that has been tested twice (such as the examples above). This paired t-test statistic is calculated as:

$$t = \frac{\bar{d}}{\sqrt{s^2/n}} \tag{4.5}$$

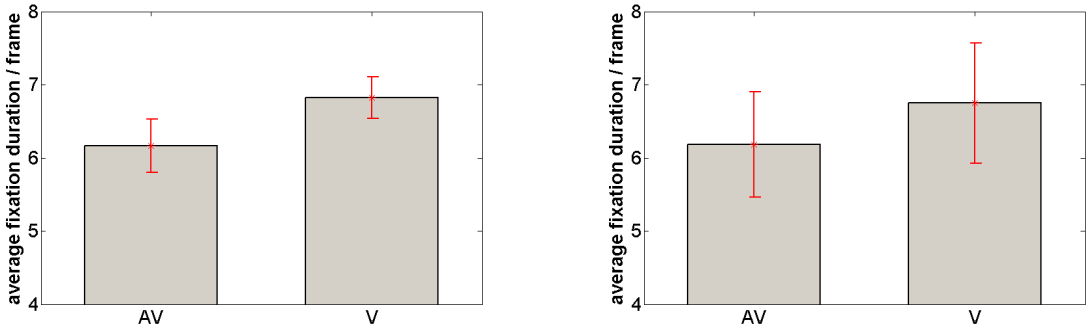
where $\bar{d}$ is the mean difference, s^2 is the variance of the samples, n is the size of the samples, and t is a Student t quantile with n-1 degrees of freedom.

Results

We investigated sound effect on fixation duration of human gaze of all the database of AV and V conditions with paired t-test. Two influence factors: participant and clip were considered separately. Each participant viewed five clips with AV condition and five with V condition, hence the fixation duration with AV and V conditions were relevant. In this situation, it was suitable to apply paired t-test.

In Fig. 4.12, by clip, AV condition has a shorter average duration of fixation (6.17 frames, 247 ms) than V condition (6.17+0.65 frames, 273 ms), and the difference is significant ($t(9) = 2.479, p = 0.035$). By participant, AV condition still has a shorter average

duration of fixation (6.19 frames, 248 ms) than V condition (6.19+0.56 frames, 270 ms), and the difference is significant ($t(35) = 2.697, p = 0.011$). That means, participants move eyes more frequently with AV condition than with V condition with paired t-test. Note that in [Coutrot 2012a] with very different experimental conditions, sound induces a tendency to increase the fixation duration.



(a) by clip (b) by participant

Figure 4.12: Distribution and mean of average fixation duration for AV and V conditions: (a) by clip (b) by participant).

4.5.2 Using mixed-effect model

According to a recent research, the mixed model is well suited for the analysis of biological data to deal with the variation between individuals (intersubject variance) and the variation within an individual (intrasubject variance) [Demidenko 2004]. In our case, intersubject variance is represented by the different participants and intrasubject variance is represented by the different clips. We only wanted to investigate the influence of conditions (AV and V), but would not take into account the differences between participants and between clips. By using mixed-effect models, we analyzed the fixed effect of conditions with crossed random effects of participants and clips.

Presentation of mixed-effect model

A *mixed model* is a statistical model that contains both fixed effects and random effects. These models are based on maximum likelihood methods and are in common use in many areas of science, medicine, and engineering. A linear mixed model can be written as:

$$Y = X\beta + Z\gamma + \epsilon \tag{4.6}$$

where the vector β is a vector of fixed effects parameters, whose elements are unknown constants to be estimated from the data. The vector of random effects is denoted by γ

and ϵ is a residual error of random vector.

Fixed effect factor is defined with a finite set of levels, and when interest lies in the estimation of each particular level effect. In our data, the effect of different conditions (AV and V) is considered as a fixed effect, which is the same for all observations in the calculation.

Random effect factor is defined with an infinite set of levels, with only a finite subset presents in the data collection. The interest lies more in the variance induced by these levels than in the estimation of the levels themselves. In our data, the effect of different participants and clips are both considered as random effects in the calculation.

Application of the model to the experiment, the random effects – subjects and items (in our case, they are participants and clips) are not independent, but related. [Baayen 2008] proposed to consider these two random effects as crossed random effects, with formula:

$$y_{ij} = X_{ij}\beta + S_i s_i + W_j w_j + \epsilon_{ij} \tag{4.7}$$

In our context, the vector y_{ij} represents the responses of participant i to clip j. X_{ij} is the design matrix for the fixed effect factor – condition (AV-V). X_{ij} consists of an initial column of ones and followed by columns representing factor contrasts and covariates. The number of rows in X_{ij} matrix is as much as the number of trials with participant i and clip j. β is the vector of the fixed effect coefficients. Like in [Baayen 2008], the S_i matrix is the structure for participant i, and W_j matrix is the structure for clip j. The participant matrix S and the clip matrix W can be combined into a single matrix written as Z, and the participant and clip random effects s and w can be combined into a single vector written as γ (in Eq. 4.6).

The formula in Table 4.6 means that the variable "Fixation" (fixation duration) depends on several terms. The fixed effect is "Condition" (condition AV-V). The random effects for "Participants" are specified as $(1 + Condition|Participants)$. This notation indicates that we introduce by-participant adjustments to the intercept (denoted by 1) as well as by-participant adjustments to "Condition". The random effects for "Clips" are specified in the same way $(1 + Condition|Clips)$.

In the simplified formula in Table 4.7, the random effects for "Clips" are specified as $(1|Clips)$. This notation indicates that we introduce adjustments to the intercept (denoted by 1) conditional on "Clips".

In the past, it was difficult to fit mixed models with multiple and crossed factors to large and possibly unbalanced data sets. The methods in the *lme4* package from **R** [2] are particularly designed to fit models with several crossed random effect factors. The following calculation is based on the *lme4* package, proposed by [Bates b].

Results

In Table 4.6, AV condition has a shorter average duration of fixation (6.18 frames, 247 ms) than V condition (6.18+0.56 frames, 270 ms). It means that the participants with AV condition tend to move their eyes more frequently compared to the participants with V condition.

Table 4.6: Analysis of mixed-effect models with fixed effect – condition and crossed random effects – participants and clips.

Formula: $Fixation \sim Condition + (1 + Condition \vert Participants) + (1 + Condition \vert Clips)$				
AIC	1232	BIC	1266	
Random effects:				
Groups	Name	Variance	Std.Dev.	Corr
Participants	(Intercept)	4.59	2.14	
	Condition V	1.18	1.08	0.095
Clips	(Intercept)	0.08	0.29	
	Condition V	0.001	0.04	-1.00
Residual		1.07	1.03	
Fixed effect:				
	Estimate	Std. Error	t value	
(Intercept)	6.18	0.37	16.38	
Condition V	0.56	0.21	2.64	

The variance of intercept for "Clips" (0.08) is much lower than that for "Participants" (4.59) in Table 4.6, suggesting the effect of clips was much lower than the effect of participants. Hence, we simplified the formula of the random effects for "Clips" from $(1 + Condition|Clips)$ to $(1|Clips)$. The results shown in Table 4.7 suggest that this simplification is correct, since the t-value in Table 4.7 (2.65) is quite similar to that in Table 4.6 (2.64).

Calculation by *lmer* provided estimations of the fixed-effects parameters, standard errors for these parameters and a t-value, but no p-values. The reason was explained by Bates –the author of *lmer* [Bates a]. The denominator degrees of freedom used to penalize certainty are unknown with unbalanced, multilevel data in our calculation. Without this degree of freedom, it is impossible to find related p-value to the t-value.

[2] **R** is an open source programming language and software environment for statistical computing and graphics [R-project Web].

Table 4.7: Analysis of mixed-effect models (simplified model) with fixed effect – condition and random effects – participants and clips.

Formula: $Fixation \sim Condition + (1 + Condition \vert Participants) + (1 \vert Clips)$				
AIC	1228	BIC	1255	
Random effects:				
Groups	Name	Variance	Std.Dev.	Corr
Participants	(Intercept)	4.59	2.14	
	Condition V	1.18	1.08	0.096
Residual		1.07	1.03	
Fixed effect:				
	Estimate	Std. Error	t value	
(Intercept)	6.18	0.37	16.45	
Condition V	0.56	0.21	2.65	

4.5.3 Conclusion

The analysis through two statistical models (paired t-test and mixed-effect model) show that AV condition has a shorter average duration of fixation than V condition. It means that the participants with AV condition tend to move their eyes more frequently, compared to the participants with V condition. Besides the fixed effect – condition (AV-V), between the random effects – participants and clips, effect of clips was much lower than that from participants.

4.6 Discussion

The study in this chapter demonstrates that not only has human speech a strong effect on human gaze when looking freely at videos, but also singer(s) and human noise. The mean distance d between the groups with AV and V conditions is lower for "off-screen sound source" cluster than for the two "on-screen sound source" clusters. The result indicates that a change in auditory information affects human gaze, when the information is linked to a visual event in the video [Hidaka 2010, Gordon 2011]. The reason might be that synchronized audio-visual events capture attention rather than unpaired audio-visual stimuli [Van der Burg 2010].

By calculating the difference between $\overline{d}_{AVV}$ (the temporal mean of distances between the two groups of participants) and randomization distribution $\overline{d}_i$, we conclude that the distance between participants with AV and V conditions is greater for four human classes (speech, singer, human noise, and singers). The database excerpts about humans are mostly made up of human voices, where the sound source is the speaker's face. The result of $\overline{D}_{AVS}$ (mean of distance between participants with AV condition and sound source) is

smaller than $\overline{D}_i$ (randomization distribution), and implies that after the auditory stimuli, humans searched for the sound source, associated with auditory information in the scene.

Interestingly, this kind of behavior is obvious when the auditory stimuli are a human voice. Acoustic and visual speeches are strongly integrated only when the perceiver interprets the acoustic stimuli as speech [Kim 2004, Tuomainen 2005]. Human voice has a strong effect on audio-visual interaction. In Fig. A.8 (a), the distance between participants with AV and V conditions of "speech" class increases after 6 frames. However, in Fig. 4.9 (a), the eye positions of participants with AV condition seem to reach the sound source after 14 frames. It takes 8 frames on average (320 ms) for a participant to move their eyes to the sound source after detecting the second sound. Therefore, an increase in uncertainty of one modality, in turn, increases the influence of another [Heron 2004]. In our case, the human speech is foreign to the participants, thus, the uncertainty of acoustic information increased: that is, visual information is processed to locate the human sound source.

In the "music" class, the distance between the eye positions of participants with AV condition and the human face (only the person playing the musical instrument) is smaller than the distance between the eye positions of participants with AV condition and the musical instrument. The visual event linked to the acoustic stimuli is the instrument, not the face. The result shows that after the participants hear music, first they tend to move their eyes to the player's face. After a while, both the human face and musical instrument are reached. One possible explanation of this behavior is that participants responded faster to social stimuli (like faces) as compared to non-social stimuli (like houses) [Escoffier 2010]. Hence, after the stimuli of the second sound (music), participants first move their eyes to the player's face, then to the musical instrument.

The comparison of fixation duration between the groups of participants with AV and V conditions was carried out. We observed that the group with AV condition had shorter fixation duration than the group with V condition. One possible reason is that auditory information brings additional information about sound source to participants. They explore the scene more quickly with AV condition. It may also be caused by the fact that the responses of the participants to bimodal audio-visual stimuli were significantly faster than unimodal visual stimuli [Sinnett 2008]. Recent research from [Zou 2012] also confirms that synchronous audio-visual stimuli facilitate visual search performance, and has shorter reaction time than visual stimuli only. Another aspect, the influence of random effect on conditions is mainly caused by the participants, not the clips.

4.7 Comparison with visual saliency models

In the previous chapter, the comparison of the eye positions with Marat's *et al.* visual saliency model, showed that the accuracy of prediction decreased when the video data in experiment I with original soundtrack. Is the performance of this visual saliency model adapted with other video data with soundtrack, if we synchronized all the clip snippets with the beginning of second sound of each clip snippet, not the beginning of the starting frame of each clip snippet (such as described in section 3.4)? Hence, we repeated the calculation of comparison of the data base in experiment II, but synchronized all the clip snippets with the beginning of second sound of each clip snippet.

4.7.1 Criteria

First criterion we chose, was still the Normalized Scanpath Saliency (NSS) criterion, which was described in section 3.4. It is especially designed to compare eye fixations with the salient areas emphasized by a model saliency map.

An additional criterion, which was proposed by Torralba (TC) [Torralba 2006] *et al.*, was applied to reinforce the results. This method simply estimates the ratio of the eye positions predicted by the saliency map over all experimental eye positions. A eye position is considered to be predicted, if it is projected on the most salient region. The most salient region is 20% of the whole salient map surface. TC value is calculated in the equation below:

$$TC = 100 \times \frac{N_{inside}}{N_{all}}\% \tag{4.8}$$

where, N_{inside} is these positions inside salient regions, and N_{all} is the total experimental eye positions.

4.7.2 Procedure

With the purpose of testing the prediction accuracy of the model, the procedure was proposed below:

- First, we calculated the NSS for each clip snippet from the onset of the second sound for both AV and V conditions.
- We then calculated NSS_V (respectively NSS_{AV}), which is the average value of NSS values for all the clip snippets, which were synchronized with the beginning of second sound of each clip snippet.

– Finally, we considered the NSS_D difference ($NSS_V - NSS_{AV}$) between groups of participants with AV and V conditions.

This procedure to calculate for TC was the same as calculating NSS. This procedure was applied on Marat's *et al.* visual saliency model and Itti's *et al.* saliency model separately.

4.7.3 Comparison with Marat's *et al.* visual saliency model

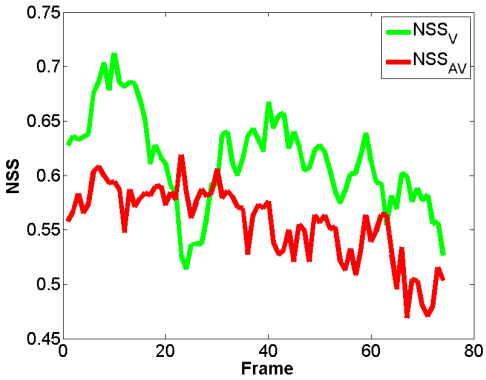
The visual saliency model we tested first was a spatio-temporal saliency model developed in our laboratory by S. Marat et al. [Marat 2009]. This model was introduced in section 2.4.2. This visual saliency model has two pathways: static and dynamic pathways. Static pathway consists of two types of interactions based on the range of the receptive fields: short interactions, which reinforce objects belonging to a specific orientation; long interactions, which are used for contour facilitation. Static saliency map is sensible to the contrast of the edge of the objects in the scene. Dynamic pathway is tightly linked to motion and particularly to the motion of a region against the background. Dynamic saliency map is sensible to the motion amplitude against the background, not the orientation of the motion.

Comparison with static pathway

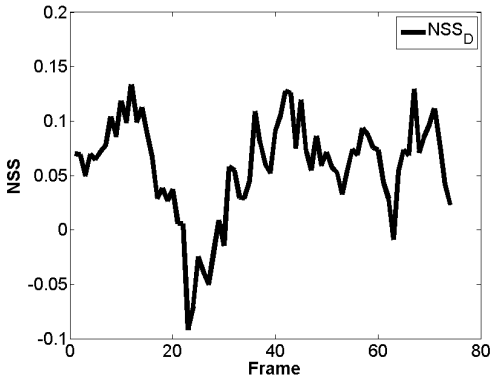
Compared with static saliency maps, the NSS (and TC) values are calculated with groups of participants with AV (NSS_{AV}) and V (NSS_V) conditions separately. In Fig. 4.13 and Fig. 4.14, visually, the NSS and TC perform similar. Hence we only describe in detail NSS curve.

In Fig. 4.13 (a), in a global view, NSS_V decreases slowly. NSS_{AV} also decreases slowly with lower value than NSS_V. In (b), if NSS_D is above 0, suggesting higher NSS_V value than NSS_{AV}. Visually, NSS_D value is above 0, for a long duration (more than 1 second). Hence, longer duration is chosen for analysis, from frames 6 to 56 (2 seconds) after the second sound. Moreover, this duration from frames 6 to 56 is taken in the following calculation of comparison with saliency models.

In order to verify whether the difference which is above 0, is significant or not, we applied t-test and Wilcoxon signed-rank test here. In order to satisfy the requirement of independent samples in t-test and Wilcoxon signed-rank test, we still took the mean of NSS (or TC) difference of 8 continuous frames as one independent sample. Small $p - value$ (less than 5 %) indicates a rejection of the null hypothesis that the samples have mean 0.

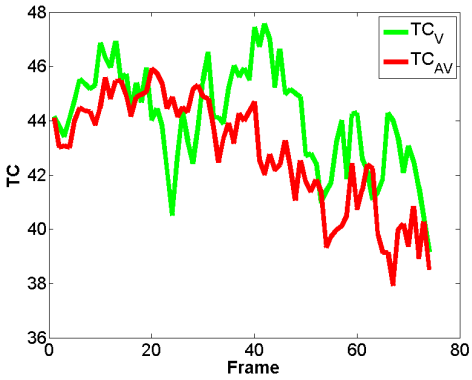


(a) NSS_V and NSS_{AV} between groups of participants with AV and V conditions over time.

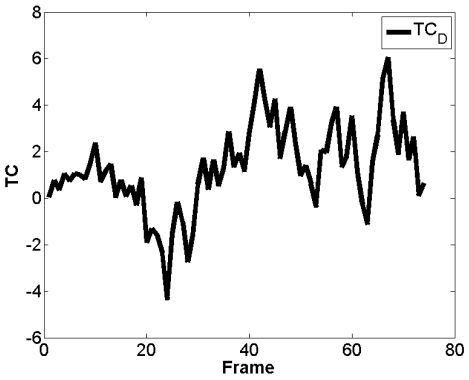


(b) NSS_D difference ($NSS_V - NSS_{AV}$) between groups of participants with AV and V conditions over time.

Figure 4.13: Results of prediction accuracy for static pathway, evaluated by NSS: NSS_V, NSS_{AV}, and NSS_D difference ($NSS_V - NSS_{AV}$) over time. The frame 0 is at the starting frame of the second sound in each clip snippet.



(a) TC_V and TC_{AV} between groups of participants with AV and V conditions over time.



(b) TC_D difference ($TC_V - TC_{AV}$) between groups of participants with AV and V conditions over time.

Figure 4.14: Results of prediction accuracy for static pathway, evaluated by TC: TC_V, TC_{AV}, and TC_D difference ($TC_V - TC_{AV}$) over time. The frame 0 is at the starting frame of the second sound in each clip snippet.

From the results of the statistical analysis in Table 4.8, we conclude that the accuracy of prediction from the static pathway of the model decreases in a group with AV condition compared to group with V condition during frames 6 to 56 after the appearance of the second sound both in criteria NSS and TC.

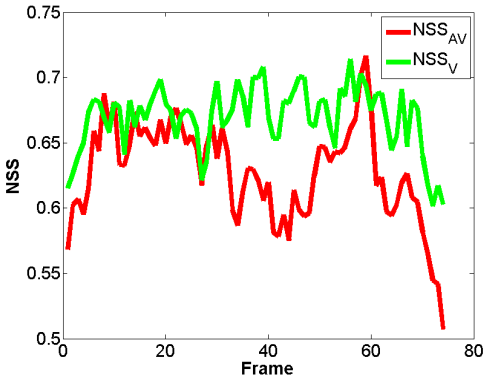
Table 4.8: NSS and TC difference between groups with AV and V conditions for static pathway with t-test and Wilcoxon signed-rank test.

	NSS	TC
p-value from t-test	0.02	0.05
p-value from Wilcoxon signed-rank test	0.007	0.04

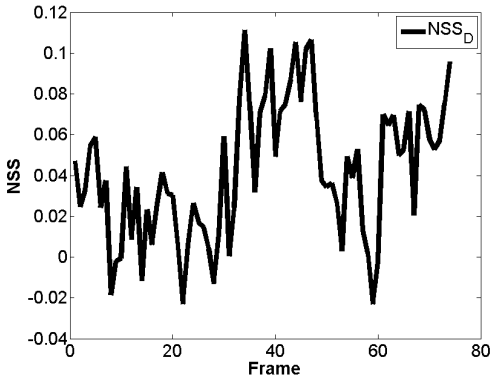
Comparison with dynamic pathway

Then, comparison of eye positions and the dynamic saliency maps was calculated with the same procedure as static pathway. In Fig. 4.15 and Fig. 4.16, globally, the performance of NSS and TC are similar.

In 4.15 (a), NSS_V increased shapely from the beginning till about frame 15, then it was stable. It was caused by the sound source in the screen normally corresponded to motion. The motion was attractable. For NSS_{AV}, it also increased at the beginning, then, less stable than NSS_V. In (b), NSS_D which was above 0, lasts for a long duration (more than 1 second). Same longer duration from frames 6 to 56 (2 seconds) after the second sound and the mean of NSS (and TC) difference of 8 continuous frames as one independent sample, were applied in statistical analysis t-test and Wilcoxon signed-rank test.



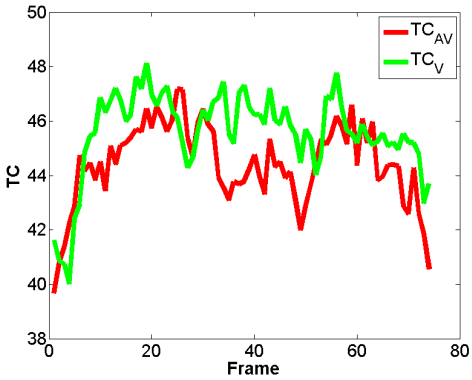
(a) NSS_V and NSS_{AV} between groups of participants with AV and V conditions over time.



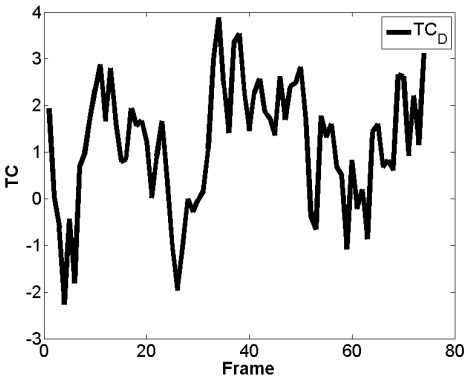
(b) NSS_D difference ($NSS_V - NSS_{AV}$) between groups of participants with AV and V conditions over time.

Figure 4.15: Results of prediction accuracy for dynamic pathway, evaluated by NSS: NSS_V, NSS_{AV}, and NSS_D difference ($NSS_V - NSS_{AV}$) over time. The frame 0 is at the starting frame of the second sound in each clip snippet.

From the results in Table 4.9, we concluded that the accuracy of prediction from the dynamic pathway of the model also decreased in a group with AV condition compared to group with V condition during frames 6 to 56 after the appearance of the second sound both in criteria NSS and TC, and this decrement was significant at a level of 5%.



(a) TC_V and TC_{AV} between groups of participants with AV and V conditions over time.



(b) TC_D difference ($TC_V - TC_{AV}$) between groups of participants with AV and V conditions over time.

Figure 4.16: Results of prediction accuracy for dynamic pathway, evaluated by TC: TC_V, TC_{AV}, and TC_D difference ($TC_V - TC_{AV}$) over time. The frame 0 is at the starting frame of the second sound in each clip snippet.

Table 4.9: NSS and TC difference between groups with AV and V conditions for dynamic pathway with t-test and Wilcoxon signed-rank test.

	NSS	TC
p-value from t-test	0.02	0.04
p-value from Wilcoxon signed-rank test	0.01	0.03

4.7.4 Comparison with Itti's *et al.* saliency model

Results of comparison of experimental eye positions with Marat's *et al.* saliency model showed that the prediction accuracy of both static and dynamic decreased in the group with AV condition compared to group with V condition during a certain duration – from frame 6 to 56 after the second sound. Besides this saliency model, we tested the most popular and well-known visual attention model, which was proposed by Itti and Koch in 1998 [Itti 1998], and is described in section 2.4.2 in details.

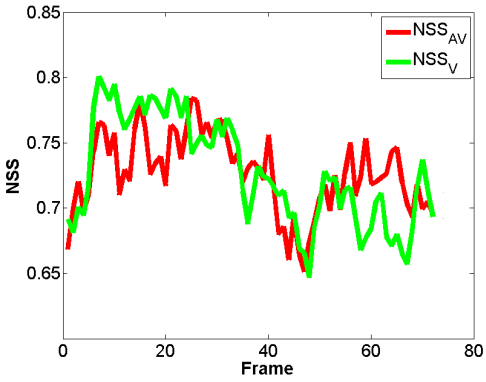
This model focuses only on bottom-up process. Different attribute maps: intensity, color and orientation features are extracted from the input scene. Then, a motion feature is added in 2003 [Itti 2003] for video.

Comparison with model proposed in 1998

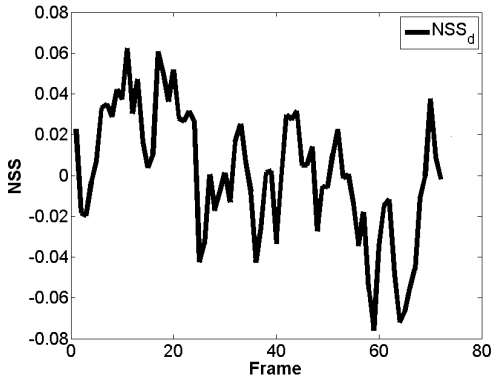
Comparison of eye positions and the saliency maps was calculated with the same procedure as for Marat's *et al.* visual saliency model. The results of NSS (and TC) values shown in Fig. 4.17 (and Fig. 4.18) suggest there was no big difference between NSS and TC. In Fig. 4.17 (a), visually, NSS_V increases a little from the beginning till frame 10, then

decreases slowly. Because the beginning of the second sound appears about 3 second (75 frames) after the beginning of each clip snippet, this increasing seems the second peak in Fig. 3.14 (around frame 50). NSS_{AV} performs similar to NSS_V, but with lower value.

Same duration for NSS_D and TC_D, which is selected in comparison of Marat's *et al.* model, is chosen for analysis, from frames 6 to 56 (2 seconds) after the second sound. Also, t-test and Wilcoxon signed-rank test are applied to verify whether the difference which is above 0, is significant or not. We still take the mean of NSS (or TC) difference of 8 continuous frames as one independent sample.



(a) NSS_V and NSS_{AV} between groups of participants with AV and V conditions over time.



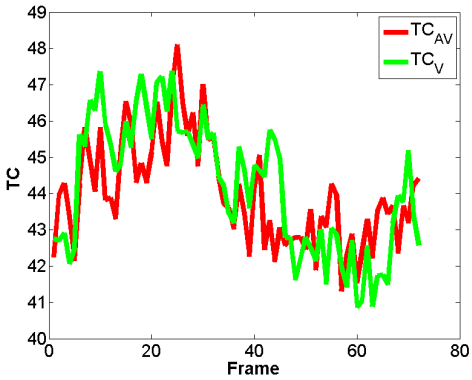
(b) NSS_D difference ($NSS_V - NSS_{AV}$) between groups of participants with AV and V conditions over time.

Figure 4.17: Results of prediction accuracy for Itti's *et al.* saliency model, evaluated by NSS: NSS_V, NSS_{AV}, and NSS_D difference ($NSS_V - NSS_{AV}$) over time. The frame 0 is at the starting frame of the second sound in each clip snippet.

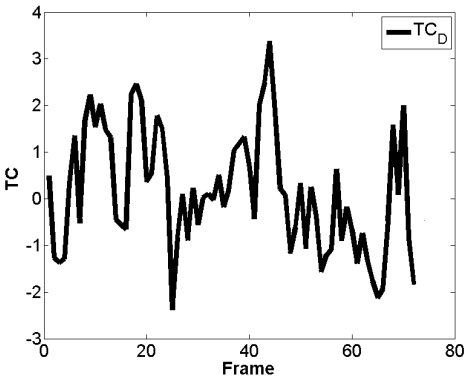
The results of the statistical analysis are shown in Table 4.10, suggesting that the prediction accuracy had a tendency to decrease in a group with AV condition compared to group with V condition, both in criteria NSS and TC. However, this decrement, during frames 6 to 56 after the appearance of the second sound, was not significant at a level of 5%.

Table 4.10: NSS and TC difference between groups with AV and V conditions for Itti's *et al.* saliency model, proposed in 1998, with t-test and Wilcoxon signed-rank test.

	NSS	TC
p-value from t-test	0.08	0.1
p-value from Wilcoxon signed-rank test	0.06	0.09



(a) TC_V and TC_{AV} between groups of participants with AV and V conditions over time.



(b) TC_D difference ($TC_V - TC_{AV}$) between groups of participants with AV and V conditions over time.

Figure 4.18: Results of prediction accuracy for Itti's *et al.* saliency model, evaluated by TC: TC_V, TC_{AV}, and TC_D difference ($TC_V - TC_{AV}$) over time. The frame 0 is at the starting frame of the second sound in each clip snippet.

Comparison with motion pathway

To complete the initial saliency model to process videos, Itti's *et al.* [Itti 2003] added a "motion" pathway. Corresponding to dynamic pathway of Marat's *et al.* model, we tested this "motion" pathway separately.

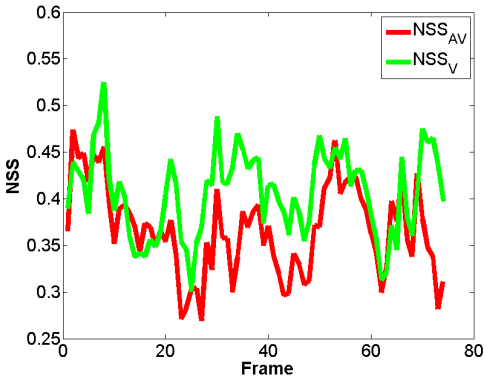
Comparison of eye positions and the motion saliency maps was calculated in the same procedure as above. In Fig. 4.19 and Fig. 4.20, visually, the NSS and TC perform similar.

In Fig. 4.19 (a), NSS_V performs stably all along time. NSS_{AV} seems decrease slowly from the beginning till frame 42, then increase a little. In (b), NSS_D suggests the NSS difference (NSS_V-NSS_{AV}) is above 0 during frames 6 to 56 after the appearance of the second sound. Hence, same duration is calculated for t-test and Wilcoxon signed-rank test to verify whether this difference is significant.

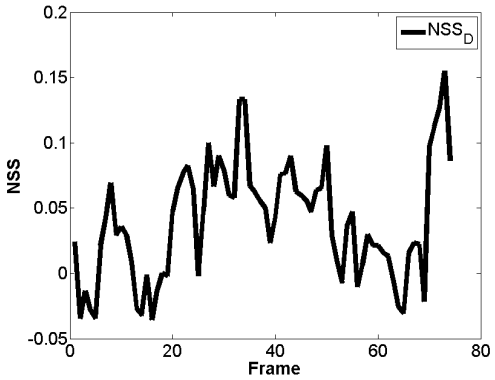
From the results in Table 4.11, the prediction accuracy decreased in a group with AV condition compared to group with V condition, both in criteria NSS and TC in a level of significance of 5%, during frames 6 to 56 after the appearance of the second sound.

Table 4.11: NSS and TC difference between groups with AV and V conditions for motion pathway with t-test and Wilcoxon signed-rank test.

	NSS	TC
p-value from t-test	0.02	0.009
p-value from Wilcoxon signed-rank test	0.03	0.01

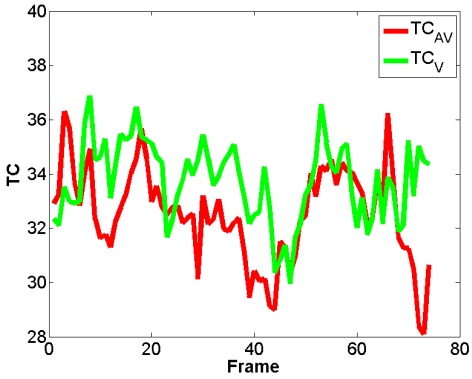


(a) NSS_V and NSS_{AV} between groups of participants with AV and V conditions over time.

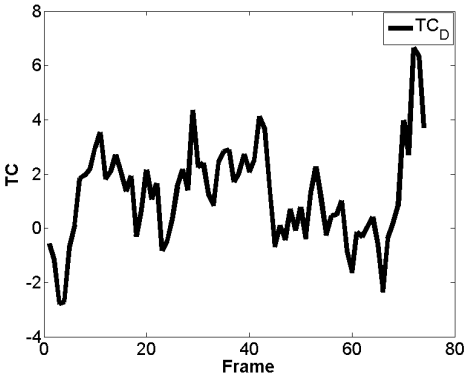


(b) NSS_D difference ($NSS_V - NSS_{AV}$) between groups of participants with AV and V conditions over time.

Figure 4.19: Results of prediction accuracy for motion pathway of Itti's *et al.* saliency model, evaluated by NSS: NSS_V, NSS_{AV}, and NSS_D difference ($NSS_V - NSS_{AV}$) over time. The frame 0 is at the starting frame of the second sound in each clip snippet.



(a) TC_V and TC_{AV} between groups of participants with AV and V conditions over time.



(b) TC_D difference ($TC_V - TC_{AV}$) between groups of participants with AV and V conditions over time.

Figure 4.20: Results of prediction accuracy for motion pathway of Itti's *et al.* saliency model, evaluated by TC: TC_V, TC_{AV}, and TC_D difference ($TC_V - TC_{AV}$) over time. The frame 0 is at the starting frame of the second sound in each clip snippet.

4.7.5 Conclusion

Comparison of the experimental eye positions with static and dynamic pathways in Marat's *et al.* saliency model, and Itti's *et al.* saliency model proposed in 1998 and additional pathway of motion in 2003, the results of NSS and TC both suggest that prediction accuracy tended to decrease in a group with AV condition than in group with V condition, during frames 6 to 56 after the appearance of the second sound. This decreasing of prediction accuracy is significant at a level of 5% for Marat's *et al.* saliency

model and the motion pathway from Itti's *et al.* saliency model.

4.8 General conclusion

This chapter presented the audio-visual experiment II, which was also designed for the purpose of investigating the sound effect on human gaze when looking at videos. Thirteen different types of sound had been proposed. Through our analysis of difference between the eye positions from group with AV condition and with V condition, we concluded that not only human speech had a strong effect on human gaze when looking freely at videos, but also singer(s) and human noise.

After the participants heard the sound of human voice cluster (speech, singer and human noise), they moved their eyes to the sound source – talking face in the scene. However, for music class, when there was a human playing a musical instrument in the scene, participants tended to move their eyes first to the player's face, rather than the sound source – musical instrument.

Participants with AV condition had a significantly shorter average duration of fixation than V condition, through the statistical analyses (paired t-test and mixed-effect model). It suggested that the participants with AV condition move their eyes more frequently compared to those participants with V condition.

When compared experimental eye positions to visual attention models proposed by Itti *et al.* [Itti 2003] and Marat *et al.* [Marat 2009], we observed that the accuracy of eye movement predictions decreased for the group with AV condition compared to the group with V condition.

Chapter 5

Preliminary audio-visual saliency model

We investigated sound influence on visual gaze when looking at videos through audio-visual experiments, which were described in details in previous two chapters. In audio-visual experiment I (chapter 3), through the analysis of difference of the eye positions between two groups of participants: with AV condition and with V condition, we observed that sound affected human gaze differently depending on the sound type, and the effect was greater for the on-screen speech class.

In audio-visual experiment II (chapter 4), a deeper investigation of the influence of audio-visual interaction on eye movement was carried out. We compared the behavior of human gaze in relation to thirteen more refined sound classes between two groups of participants: with AV condition and with V condition. After the participants heard the sound of human voice cluster (speech, singer and human noise), they moved their eyes to the sound source (talking face) in the scene. However, for musical instrument class, participants tended to move their eyes first to the player's face, rather than the sound source (musical instrument).

In this chapter, a state of the art of audio-visual fusion schemes and methods are first briefly presented. Then, based on the results obtained from two audio-visual experiments above, a preliminary audio-visual saliency model is proposed. This preliminary audio-visual saliency model mainly concentrates on two different fusion strategies of speech and musical instrument sound classes.

5.1 State of the art of audio-visual fusion

Current psychophysical studies on audio-visual interaction concentrate on two areas: the influence of visual input on auditory perception and the influence of acoustic input on

visual perception (described in section 2.2). To take into account the influence of audio-visual interaction, when deal with multimedia analysis tasks, many researchers pay attention to multimodal fusion. A multimedia analysis task involves processing of multimodal data, such as video with its original soundtrack, to obtain valuable insights about the data. These multimedia data consist of related features, which are represented in multiple modalities. The fusion of these multiple modalities can provide complementary information than single modality, and also increase the accuracy of the overall decision. For example, fusion of audio-visual features with the incorporation of web casting text analysis can significantly improve the event detection in the sport video [Xu 2008]. In this thesis, we focused on human attention, which is multimodal in nature, with senses of vision and hearing. Hence, we concentrated on the fusion of visual and audio information.

In the analysis process, audio-visual fusion comes with a certain computational cost and complexity. This is due to the different characteristics of audio and visual information, which are briefly described below:

- Visual and audio signals are captured in different formats and at different rates. In a video, visual information is captured at a certain frame rate, which is different from the audio sampling rate obtained from audio signal. Therefore, the fusion of visual and audio information should consider this asynchrony.

- The processing time of visual and audio information are different. The complexity of the signals are different: visual signal is two-dimension and audio signal is one-dimension.

- The visual and audio events may be independent or related. For example, in cluster of sound classes "on-screen with one sound source" described in chapter 4, visual and audio events are related. In cluster of sound classes "off-screen sound source", visual events are independent from the audio events. When fusing visual and audio information, the correlate or independent events may be fused differently.

5.1.1 Audio-visual fusion schemes

In order to solve these varying characteristics above in audio-visual fusion, different fusion strategies are proposed. Four levels of fusion of visual and audio information are presented in the following part: feature level, classifier level, decision level and hybrid level [Shivappa 2010]. Fig. 5.1 shows different fusion strategies of feature, classifier and decision levels.

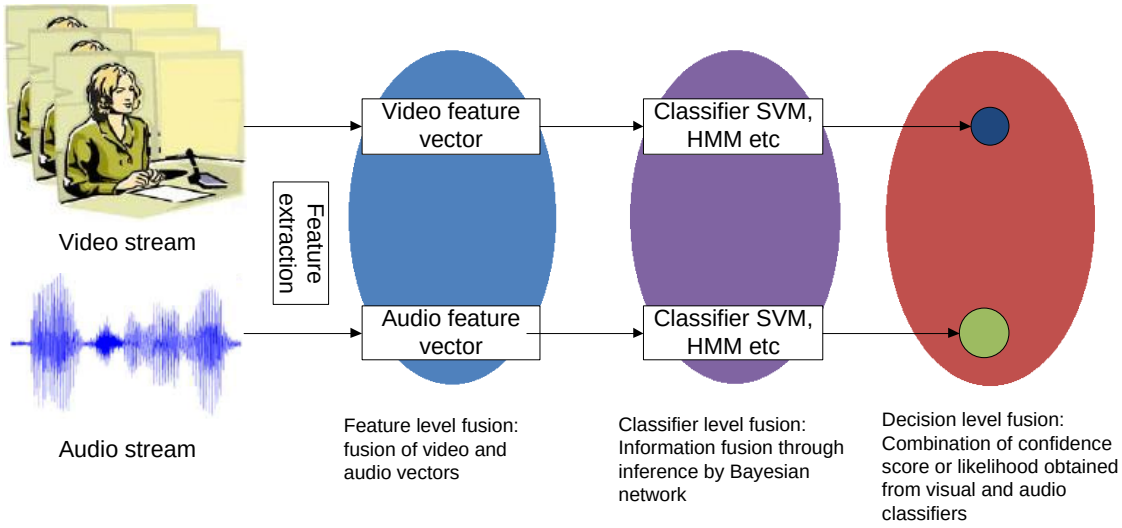


Figure 5.1: Different level of fusion strategies of visual and audio signals.

Feature level audio-visual fusion strategies

The feature level, also called early fusion approach, comes after the feature extraction of input visual and audio streams. In the feature level fusion, numbers of features extracted from visual and audio signals are numerous, and summarized as [Wang 2000]:

- *Visual features:* The visual features can be extracted from color, contrast, motion etc.. The motion for example can be represented in the form of motion direction and magnitude.

- *Audio features:* It may include features of energy, non-silence ratio, zero crossing rate (ZCR), mel-frequency cepstral coefficient (MFCC), etc..

This feature level fusion has the advantage that it uses the correlation between visual and audio features at an early stage. However, the combined features from visual and audio signals have a large dimensionality. For further calculation, dimensionality reduction techniques are required, such as principal component analysis (PCA), quadratic discriminant analysis (QDA), linear discriminant analysis (LDA), etc.. However, this fusion approach is hard to represent the time synchronization between relevant visual and audio events. Hence it can not be applied to the tasks, which has strict temporal synchronization requirement.

Several different audio-visual analysis tasks are using this early fusion. For example, [O'Donovan 2007] proposed to fuse visual and audio information in a early stage for audio-visual tracking. In their work, they treated the signal of microphone arrays as generalized

camera arrays, not as two separate geometry sensors.

Classifier level audio-visual fusion strategies

The classifier level fusion comes after processing the features extracted from visual and audio streams separately. Visual and audio information are fused within the classifier. To model visual and audio streams separately, for example, dynamic Bayesian networks (DBNs) are adopted.

This classifier fusion does allow the weighted combination of visual and audio modalities, which is based on their reliability. This fusion scheme is widely applied on audio-visual speech recognition systems. For example, [Shivappa 2008] proposed a multimodal information fusion scheme based on iterative decoding theory with its application to audio-visual speech recognition system.

Decision level audio-visual fusion strategies

The decision level, also called late fusion approach, involves the combination of probability scores or likelihood values obtained from separate classifiers to obtain a final decision of the task. Compared to feature and classifier levels of fusion, there are many advantages for the decision level. Unlike feature level, where visual and audio signals have different representation, in the decision level, they usually have the same representation. It will be easier to fuse. On the other hand, in the decision level of fusion, local decisions are obtained from different classifiers, it becomes time-consuming because of the learning process in the classifiers.

A decision level fusion can be applied on improving speech recognition, by using audio-visual models [Glotin 2001]. It can also be applied on detection of monologues in video shots. [Iyengar 2003] fused the decisions obtained from a face detector and a speech recognizer based on their synchrony score.

Hybrid level audio-visual fusion strategies

A combination of the feature level and decision level fusion strategies is studied by several researchers to exploit the advantages of the fusion strategies above, named hybrid level fusion.

In this hybrid level fusion, the features extracted from visual and audio signals separately, are first fused in feature level and at the same time, the individual features are analyzed in decision level. At last, all the decisions obtained are fused to obtain the final decision.

Several researchers have successfully adopted the hybrid fusion strategy in their applications. For example, [Keller 2009] proposed a spectral diffusion framework to provide a spectral embedding of multimedia data, which was applied to audio-visual speech recognition.

5.1.2 Audio-visual fusion methods

There are different methods to fuse visual and audio information. These methods are suitable under different settings and can be classified to three main categories: rule-based, classification-based and estimation-based methods [Atrey 2010].

Rule-based methods

These rule-based fusion methods include a variety of basic rules, such as linear weighted rule, majority voting rule, custom-defined rule, etc.. Visual and audio information can be combined in different levels of fusion with rule-based methods.

In this category, linear weighted fusion is widely used for the reason that it is one of the simplest methods. For example, [Jaffré 2006] applied a linear fusion of audio and video indexes in person identification from audio-visual source.

Classification-based methods

In this category, a variety of classification techniques such as support vector machine (SVM), dynamic Bayesian networks, neural networks, etc. are used to classify visual (or audio) information into one of the predefined classes. Visual and audio information can be combined in different levels of fusion with classification-based methods.

For example, [Bredin 2007] proposed a biometric modality based scheme to identify talking face. The main idea was to use the audio-visual speech synchrony measure between the voice of the talking face and its related video frames. They adopted decision level audio-visual fusion strategies, to calculate scores for speaker verification, face recognition and synchrony. Then, these scores were sent and combined in a SVM model. This SVM model provided the final decision of the identification of the talking face.

Estimation-based methods

In this estimation-based category, methods are mainly used to better estimate the state of moving objects in audio-visual data, such as Kalman filter. [Talantzis 2006] proposed a system for tracking people in three dimensions, adopting a decentralized Kalman filter, which fused visual and audio information to better locate estimation in real time. Also, they applied a decision level of fusion in their work.

5.2 Preliminary audio-visual saliency model of speech class

In previous analysis of audio-visual experiment I (section 3.5), we proposed to locate the coordinates of the sound source manually in the scene. The sound source saliency maps increased prediction accuracy of group of participants with AV condition, suggesting sound source in the scene was attractive for participants with AV condition. Because there is no spatial information contained in the sound signal of the database, it is a difficult task to locate sound source in the visual scene automatically.

On the other hand, there is evidence that faces in the scene are preferred by the visual system compared to other object categories [Rossion 2000], and can be processed at the earliest stage after stimulus presentation [Ro 2001]. More precisely, a recent work by Rahman [Rahman 2013] developed in our lab, shows that different faces in the scene do not attract attention equally, eye movements are influenced by the location, number and size of the faces. For speech class, if there is *only one face*, and this face is talking face, it is well-known that this talking face region attracts attention for participants with AV and V conditions. In the present study, we focus on a more complex situation that besides the sound source (talking face), there are other faces in the scene, do other non-talking faces still attract attention? To find out whether the sound source of speech class (talking face) is more attractive than other faces of participants with AV condition. In this section, we first present an investigation of eye movement behavior of the participants, when they hear a sound of speech.

Database

To better investigate the eye movement behavior, we concentrated on clip snippets, which satisfy two conditions :

- In audio aspect, soundtrack of the clip snippet contains speech period.
- In visual aspect, there is only one talking face in the frame, and there are other non-talking faces in the same frame.

Hence, five clip snippets are selected from speech class (eleven clip snippets) in audio-visual experiment II. Fig. 5.2 shows frame examples extracted from the five clip snippets.

5.2.1 Eye movement behavior of speech class

To investigate for participants whether sound source (talking face) is more attractive than other non-talking faces during speech period, we consider each face region in the scene

(a) Clip snippet I

(b) Clip snippet II

(c) Clip snippet III

(d) Clip snippet IV

(e) Clip snippet V

Figure 5.2: Frame examples extracted from five clip snippets in speech class: clip snippets I to V.

as an individual saliency map. By comparing the experimental eye positions with each individual face map separately, we can find out which face is more attractive. Detailed description of this approach is presented below:

- First, we labeled the position of each face in the visual frame manually. The face position is located by a bounding box. A frame example is shown in Fig. 5.3 (a).

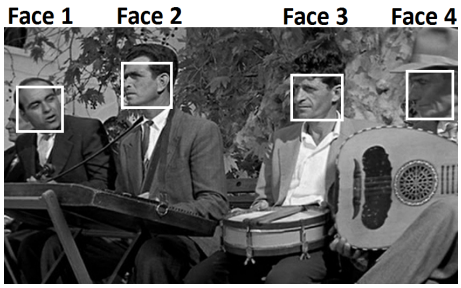
- Then, in each face region, a 2-D Gaussian is applied in the bounding box. The variance of this 2-D Gaussian is determined by the dimensions of the bounding box. In each bounding box, from origin in both horizontal and vertical axis, the amplitude of the 2-D Gaussian function remains the same value. This hand-labeled face position covered with a 2-D Gaussian map is called face map M_f. Face map examples are shown in Fig. 5.3 (b), (c), (d) and (e).

To evaluate the comparison between eye positions and face map, we still chose NSS and TC criteria, which are detailed in section 4.7.

In the following, we present the eye movement behavior of participants with clip snippet I by using the approach above. First of all, the content of clip snippet I is described below:

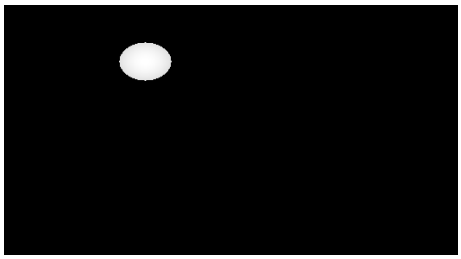
- In the visual scene, there are four faces: one talking face and three non-talking faces. From left to right in the frame, four faces are indexed from 1 to 4, and the corresponding hand-labeled face maps are M_{f1}, M_{f2}, M_{f3} and M_{f4}. Fig. 5.3 shows a frame example in clip snippet I and M_{f1}, M_{f2}, M_{f3} and M_{f4} for the same frame.

- The soundtrack of clip snippet I is manually labeled: from frame 12 to 30 is speech period, and face 1 is talking face; from frame 58 to the end is musical instrument period (detailed analysis is presented in section 5.3), and face 4 is player's face.

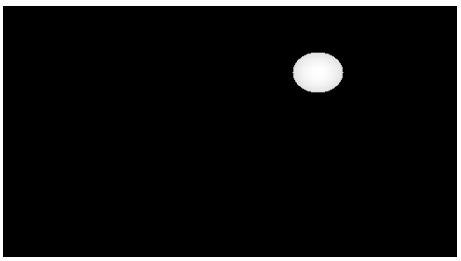


(a) A frame with hand-labeled face positions.

(b) Face map M_{f1}



(c) Face map M_{f2}



(d) Face map M_{f3}

(e) Face map M_{f4}

Figure 5.3: A frame example in clip snippet I with hand-labeled face positions and corresponding face maps M_{f1}, M_{f2}, M_{f3} and M_{f4}.

To find out which face is more attractive during speech period, we compared experimental eye position density map M_h from groups with AV condition (respectively with V condition) with face map M_{f1}, M_{f2}, M_{f3} and M_{f4} separately. This comparison is evaluated by the criteria of NSS and TC. High NSS and TC indicate more participants looking at the saliency regions.

Fig. 5.4 shows the NSS values of groups of participants with AV condition (respectively with V condition) over time for M_{f1}, M_{f2}, M_{f3} and M_{f4} separately. In (a), visually, NSS_{AV} increases sharply after frame 15, resting at a high value before decreasing sharply from frame 35. From frame 15 to 35, NSS_{AV} is much higher than NSS_V, suggesting with AV condition, after the speech stimuli, participants tend to move their eyes to talking face rather than other faces. In (b), both NSS_{AV} and NSS_V get a peak around frame 10. It is because of the influence of previous clip snippet. At the end of previous clip snippet, there exists salient regions around face 2. In (d), from frame 60 to 80, NSS_{AV} is higher than NSS_V, we will analyze it in next section of musical instrument class.

To confirm the results, another criterion TC is calculated. Fig. 5.5 shows the TC values of groups with AV and V conditions over time for M_{f1}, M_{f2}, M_{f3} and M_{f4} separately. The performance of TC is similar to NSS (in Fig. 5.4), suggesting that talking face is more attractive for participants with AV condition than other faces in the frame, during the sound of speech period.

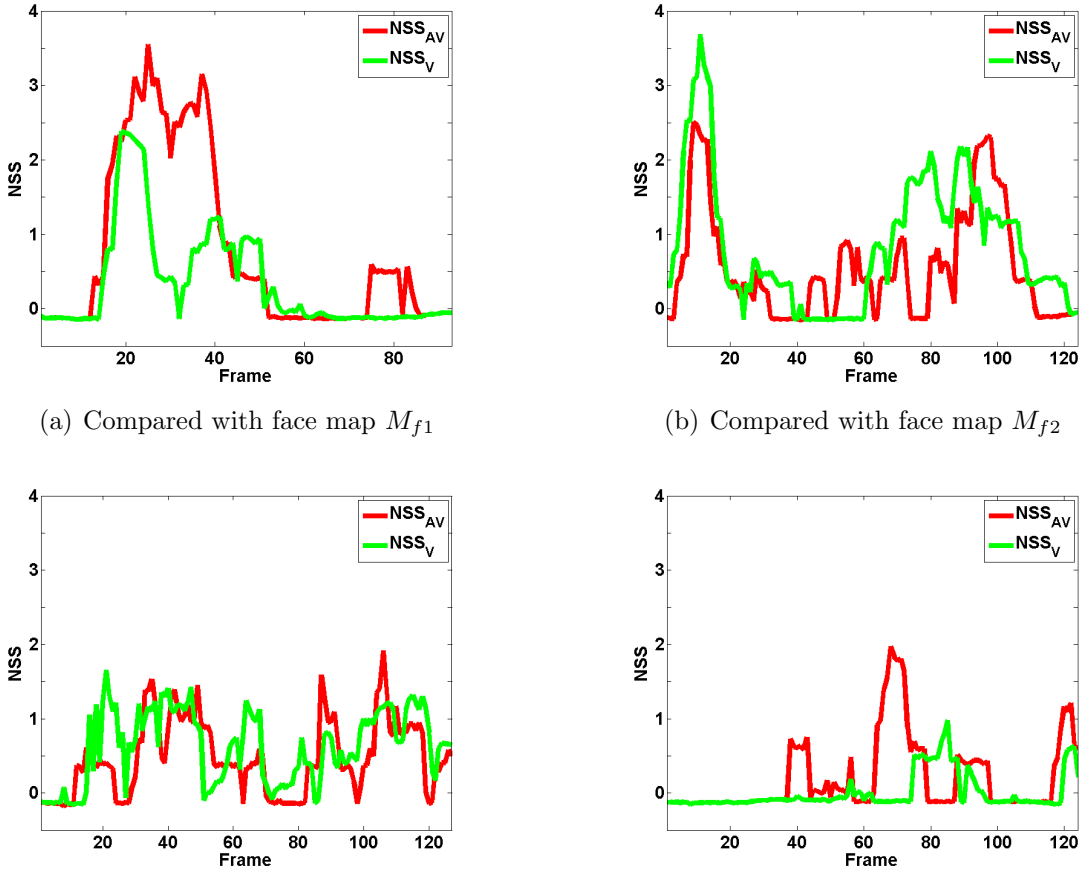
Both results of NSS and TC show that during the soundtrack of speech period, talking face gained more attention than other faces in the frame in group of participants with AV condition.

5.2.2 Proposal of an audio-visual saliency model

From investigation of clip snippet I above, we found that talking face was more attractive for participants with AV condition than other faces in the frame. Hence, for a saliency model, the principle is to locate talking face as salient region after the sound stimulus of speech. Based on other studies, two *hypotheses* are proposed:

- Several researches pointed out that visual motion was in conjunction with associated soundtrack [Kidron 2005]. Another important characteristic is that for talking face, the movement of face (especially speaker's lip) is synchronized with speech sound. Hence, we propose as *hypothesis* that talking face is moving.

- Other studies pointed out that event boundaries corresponded to points of maximum quantitative change of physical features [Zacks 2001, Evangelopoulos 2008a]. In a video, during a period of speech, the maximum changes both in visual and audio will be at the beginning or at the end of the speech period. Hence, we propose as



(a) Compared with face map M_{f1}

(b) Compared with face map M_{f2}

(c) Compared with face map M_{f3}

(d) Compared with face map M_{f4}

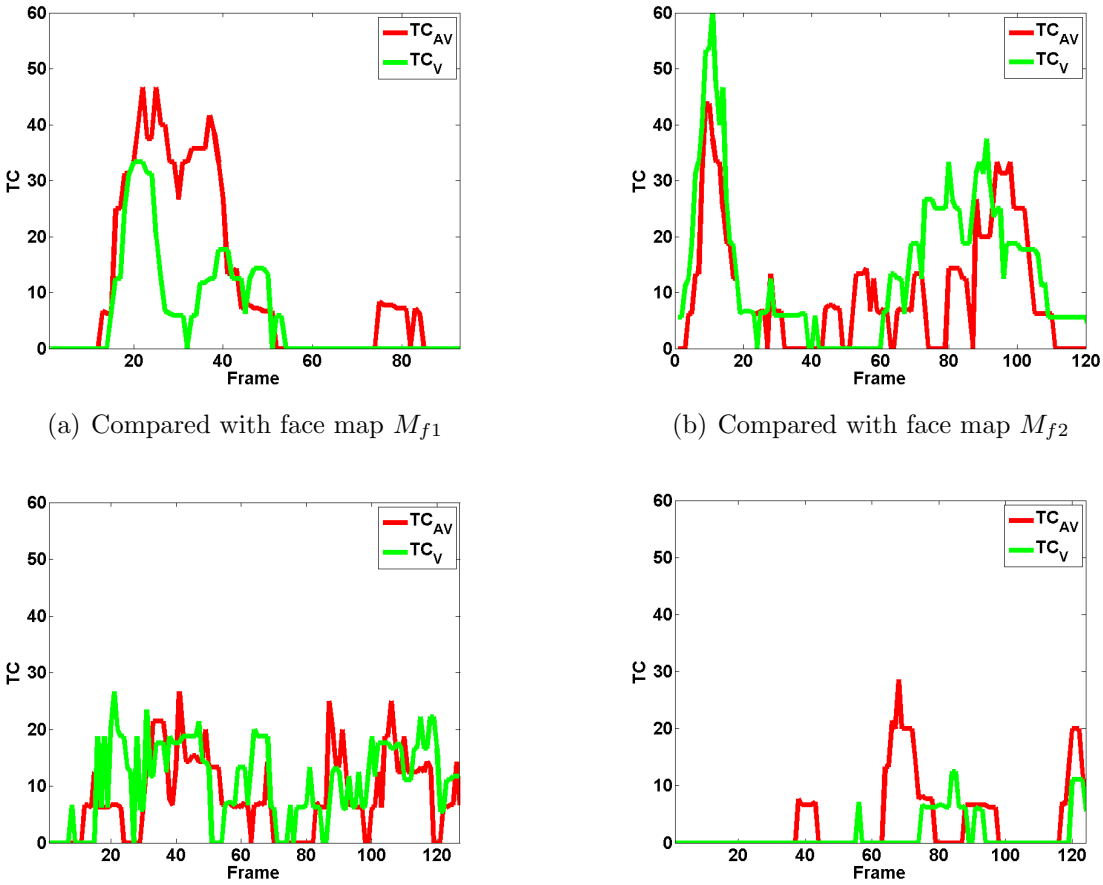
Figure 5.4: Results of prediction accuracy of clip snippet I for face maps: M_{f1}, M_{f2}, M_{f3} and M_{f4}, evaluated by NSS. When face maps are compared with group with AV condition (respectively with V condition), results are called NSS_{AV} (NSS_V). The frame 1 is the starting frame of clip snippet I.

hypothesis that movement of talking face is temporally more salient at the beginning of speech period.

Based on these hypotheses, we propose a kind of fusion of locations of all the faces and motion information to identify talking face. Motion information is provided by dynamic/motion pathway of a spatio-temporal saliency model, which was developed in our laboratory by S. Marat et al. [Marat 2009]. This dynamic/motion pathway is tightly linked to motion and particularly to the motion of a region against the background.

To achieve the objective of locating talking face automatically, two main procedures should be done:

- Soundtrack should be classified to speech and non-speech periods. Each period has its corresponding starting frame and ending frame in visual over time. Original soundtrack in clip snippets contains different types of sound, such as speech, music,



(a) Compared with face map M_{f1}

(b) Compared with face map M_{f2}

(c) Compared with face map M_{f3}

(d) Compared with face map M_{f4}

Figure 5.5: Results of prediction accuracy of clip snippet I for face maps: M_{f1}, M_{f2}, M_{f3} and M_{f4}, evaluated by TC. When face maps are compared with group with AV condition (respectively with V condition), results are called TC_{AV} (TC_V). The frame 1 is the starting frame of clip snippet I.

animal, etc.. Moreover, most of the time, the soundtrack is multiple audio bands. It means that at a certain time, the audio signal is mixed by more than one type of sound. It increases the difficulty to classify the sound to speech and non-speech periods. To avoid the error from misclassification, speech and non-speech periods are labeled manually.

- Locations of all the faces in the scene should be given. For the reason that the clip snippets have complex background and turning faces (not frontal face, which is hard for face detector, like the well-known face detector proposed by Viola *et al.* [Viola 2004]) in the frame. We hand-labeled the positions of all the faces in each frame to avoid the error from mislabeled faces, before the selection of talking face.

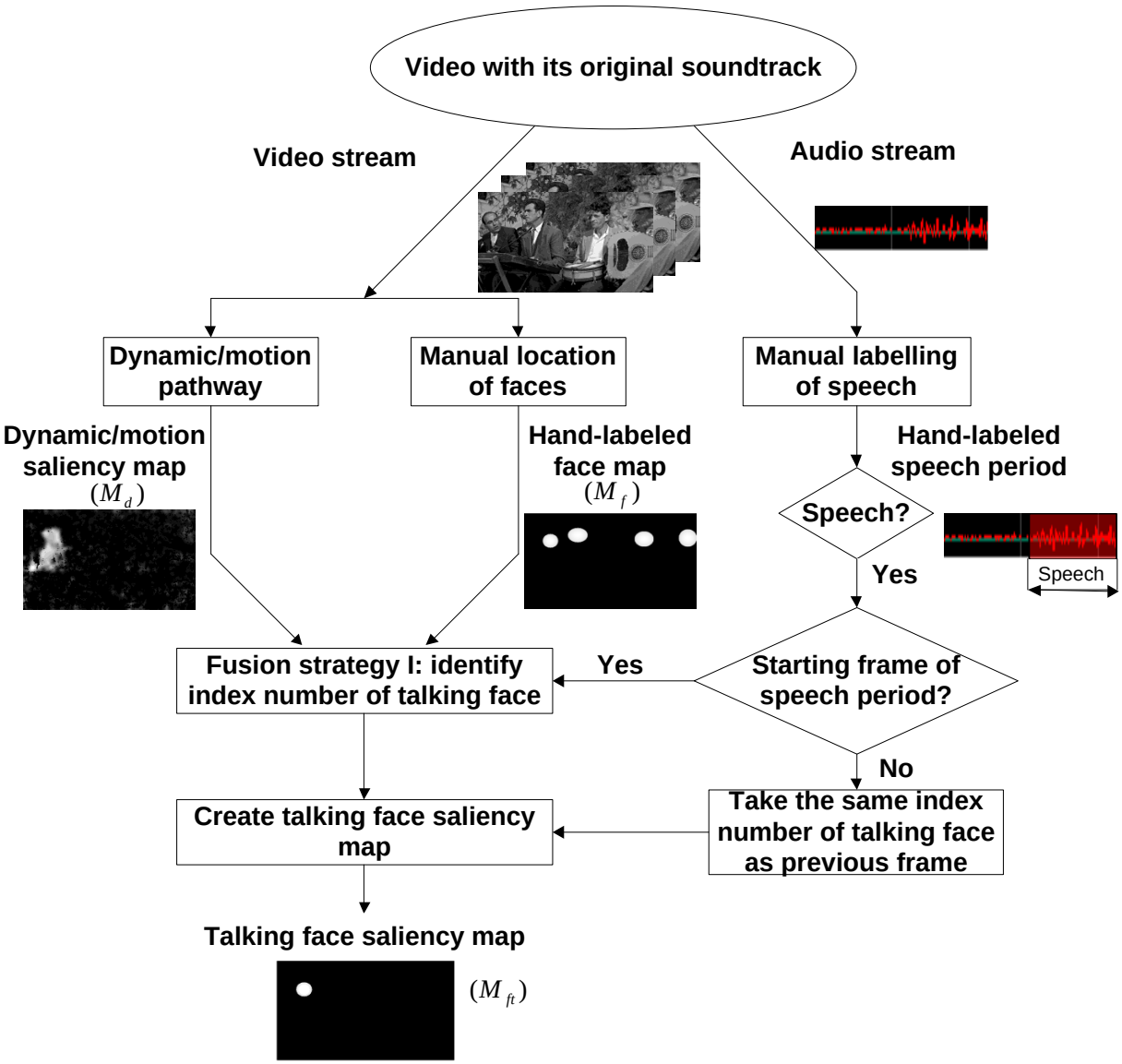


Figure 5.6: Flow chart of proposed saliency model for speech class.

The flow chart of the audio-visual saliency model is shown in Fig. A.10, and the algorithm of fusion strategy I is described briefly below:

1. If it is the starting frame of speech (the boundary of audio event), we calculate the spatial mean of dynamic/motion saliency value of each individual hand-labeled face region, recorded as $M_{df_i}(k)$ (i is index number of the face, and k is frame number).

2. After, we consider a short duration as the boundary of audio event, not just one frame. Hence, we calculate the temporal mean of $M_{df_i}(k)$ (k=1,2...7) (recorded as $\overline{M}_{df_i}$) for each individual face. Duration of seven frames is chosen for the reason that it is equal to the mean fixation duration.

3. The boundary of visual event which corresponds to this speech period should have maximum quantitative change of motion. Dynamic saliency map is sensible to this change of motion and represents this big change as high value of $\overline{M}_{df_i}$. Hence, we considered the face region, which has highest value of $\overline{M}_{df_i}$ as talking face region. Once the talking face is selected at the beginning of speech period, this selection will be kept over all this speech period.

4. At last, we added a 2-D Gaussian to the center of selected talking face region to create *talking face saliency map* M_{ft}.

This fusion strategy I of selecting talking face from all the faces performs well on the database of five speech clip snippets chosen before. The talking faces in these five clip snippets are 100% correctly selected.

5.2.3 Comparison with visual saliency model

In order to test the accuracy of prediction of proposed talking face saliency map M_{ft}, we compare M_{ft} with other two visual saliency maps: dynamic/motion saliency map M_d and hand-labeled face map M_f. To evaluate the prediction accuracy of each saliency map, the same criteria are chosen: NSS and TC. Fig. 5.7 shows a frame example of M_d, M_f and M_{ft}.

(a) Original frame

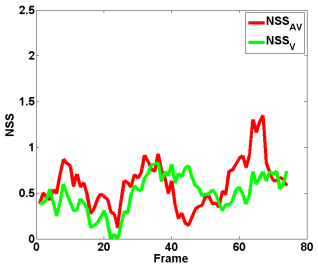
(b) Dynamic/motion saliency map M_d

(c) Hand-labeled face map M_f

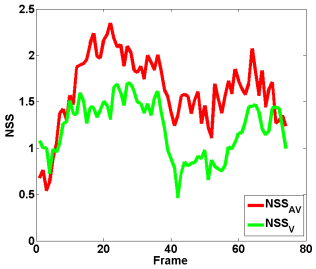
(d) Talking face saliency map M_{ft}

Figure 5.7: A frame example of dynamic/motion saliency map M_d, hand-labeled face map M_f and talking face saliency map M_{ft}.

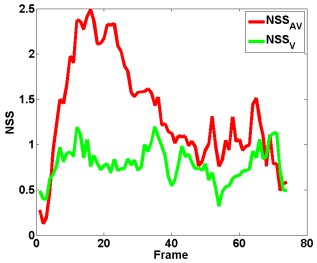
First, M_d, M_f and M_{ft} are evaluated by criterion NSS. Fig. 5.8 shows the average value of NSS of testing data of five clip snippets. All the clip snippets are synchronized with the starting frame of speech. In (a), the performance of NSS_{AV} and NSS_V are similar: stable over time with low NSS value. In (b), NSS_{AV} increases sharply after the stimulus of speech sound, then keeps a high value over time. In (c), NSS_{AV} increases sharply after the stimulus of speech sound, suggesting that talking face attracts attention for participants with AV condition. However, NSS_V is stable over time, suggesting that talking face has no particular attraction for participants with V condition.



(a) Dynamic/motion saliency map M_d evaluated by NSS

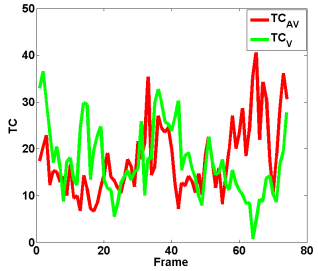


(b) Hand-labeled face map M_f evaluated by NSS

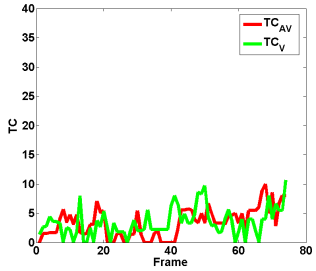


(c) Talking face saliency map M_{ft} evaluated by NSS

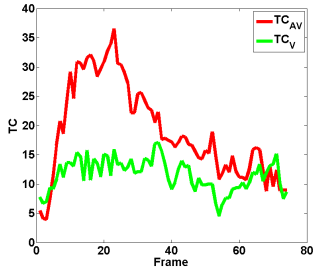
Figure 5.8: Results of prediction accuracy of mean of five clip snippets for M_d, M_f and M_{ft}, evaluated by NSS. When maps are compared with group with AV condition (respectively with V condition), results are called NSS_{AV} (NSS_V). Frame 1 is the starting frame of speech.



(a) dynamic/motion/motion saliency map M_d evaluated by TC



(b) Hand-labeled face map M_f evaluated by TC



(c) Talking face saliency map M_{ft} evaluated by TC

Figure 5.9: Results of prediction accuracy of mean of 5 clip snippet for M_d, M_f and M_{ft}, evaluated by TC. When maps are compared with group with AV condition (respectively with V condition), results are called TC_{AV} (TC_V). Frame 1 is the starting frame of speech.

Then, to confirm the result, M_d, M_f and M_{ft} are evaluated by criterion TC. Fig. 5.9 shows the average value of TC of testing data of the same five clip snippets. All the clip snippets are synchronized with the starting frame of speech. In (c), both TC_{AV} and

TC_V perform similar to NSS_{AV} and NSS_V (in Fig. 5.8). Hand-labeled face map M_f (in (b)) gets highest difference between NSS and TC. For TC, it only takes into account 20% of most salient regions. Hence, in M_f, only small region of the center of each face is considered.

At last, we observe in detail the same period as in chapter 4 that from frame 6 to 30 (1 second after speech stimuli). Table 5.1 shows the temporal mean value of NSS and TC from frame 6 to 30 after the starting frame of speech of group of participants with AV condition (respectively with V condition). Compared to dynamic/motion saliency map and hand-labeled face map, talking face saliency map performs best in group of participants with AV condition both evaluated in NSS and TC. However, when compare to group with participants with V condition, this talking face saliency map does not perform better than the other two maps. It suggests that participants with V condition does not pay more attention to talking face.

Table 5.1: The temporal mean value (from frame 6 to 30 after the starting frame of speech) of NSS and TC when compared with dynamic/motion saliency map M_d, hand-labeled face map M_f, and talking face saliency map M_{ft} presented in Fig. 5.8 and 5.9

Saliency map	M_d		M_f		M_{ft}	
Condition	AV	V	AV	V	AV	V
NSS	0.53	0.29	1.87	1.43	2.00	0.84
TC	12.74	16.27	2.68	2.20	27.95	13.17

5.2.4 Conclusion

We observed that the talking face was more attractive than other faces in the frame for participants with AV condition, after stimulus of speech. Then, to select talking face automatically, we fused hand-labeled face map with dynamic/motion saliency map, which is the output of dynamic/motion pathways in Marat's *et al.* visual saliency model. For eye positions of participants with AV condition, this talking face saliency map performed better compared to hand-labeled face map and dynamic/motion saliency map both evaluated by NSS and TC. Also, compared to condition V, this talking face saliency map increased accuracy of prediction for condition AV after stimulus of speech.

5.3 Preliminary audio-visual saliency model of musical instrument class

Compared to speech class, which is well discussed in recent decades, music class is less explored. To better understand the influence of audio-visual interaction, when soundtrack is

music, we propose a deeper investigation of eye movement behavior of musical instrument class.

In musical instrument class, sound source in the visual scene is a potential salient region, and according to knowledge presented in previous section that faces in the scene are preferred by the visual system compared to other object categories, faces are also potential salient regions. If there is no face in the scene, sound source will gain attention rather than other objects. In the films, situation may be more complex. For example, in audio aspect, soundtrack is music; in visual aspect, there is more than one face in the scene, moreover, there is only one musical instrument in the frame, which is in conjunction with soundtrack of music. In this section, we focus on this complex situation. We extract three clip snippets from music class (seven clip snippets) in database of audio-visual experiment II, which satisfy the conditions above, to create musical instrument class in the following analysis.

Database

Fig. 5.10 shows a frame example of each clip snippet with soundtrack of musical instrument. Three clip snippets are selected from music class with complex situation, which satisfy conditions above. There is only one musical instrument carrying out music sound, and more than one face are in the scene. However, we do not limit the number of other objects in the scene. For example, clip snippet I contains four musical instruments. Because clip snippet I and IV have both speech and musical instrument periods separately over time, they also appeared in speech class.

5.3.1 Eye movement behavior of musical instrument class

For musical instrument class, we consider the sound source of musical instrument class is musical instrument, which is in conjunction with visual motion and carries out music sound. In order to find out whether player's face attract more attention than other faces, we first present an investigation of eye movement behavior of the participants on one example – clip snippet I over time. Clip snippet I contains four faces in the scene, and one of them is player's face. Four musical instruments are presented in the scene, and one of them is sound source. The soundtrack of clip snippet I is manually labeled: from frame 12 to 30 is speech period (discussed in previous section); from frame 58 to the end is musical instrument period.

In clip snippet I, from left to right in the frame, four faces are indexed from 1 to 4, and correlated hand-labeled face map are M_{f1}, M_{f2}, M_{f3} and M_{f4} (shown in Fig. 5.3). The stimulus of musical instrument in soundtrack starts from frame 58, and face 4 is the

(a) Clip snippet I

(b) Clip snippet IV

(c) Clip snippet VI

Figure 5.10: Frame examples extracted from three clip snippets in musical instrument class: clip snippets I, IV and VI.

player's face. Fig. 5.4 shows the NSS values of groups with AV and V conditions over time for M_{f1}, M_{f2}, M_{f3} and M_{f4} separately. In (d), visually, NSS_{AV} increases sharply after frame 60, and this increasing does not appear in NSS_V. Compared to NSS_{AV} in (a), (b) and (c), after frame 60, M_{f4} gets the highest value, suggesting that player's face is more attractive than other faces in the frame. Similar performance of another criterion TC is shown in Fig. 5.5 to confirm the results.

5.3.2 Proposal of an audio-visual saliency model

From investigation of clip snippet I above, we found that after stimulus of musical instrument, participants tend to move their eyes first to the player's face rather than to the sound source (musical instrument). Moreover, compared to player's face, other faces in the scene have less attractability. Hence, for a saliency model of musical instrument class, we propose to detect the player's face as salient region after the musical instrument sound stimulus. Fusion strategy II based on three *hypotheses* below, is proposed to detect player's face automatically:

- Visual motion is in conjunction with associated sound source (musical instrument).

- At the beginning of musical instrument period, visual motion is highest on sound source (musical instrument) region.

- The Euclidean distance between musical instrument and the player's face is shortest, compared to distance between musical instrument and non-player's face in the frame.

Before fusion strategy II, which fused motion and face information to detect player's face, two main procedures should be done:

- Soundtrack should be classified to speech and non-speech periods. Each period has its corresponding starting frame and ending frame in visual over time. Original soundtrack in clip snippets contains different types of sound, such as speech, music, animal, etc.. Moreover, most of the time, the soundtrack is multiple audio bands. It means that at a certain time, the audio signal is mixed by more than one type of sound. It increases the difficulty to classify the sound to speech and non-speech periods. To avoid the error from misclassification, speech and non-speech periods are labeled manually.

- Locations of all the faces in the scene should be given. For the reason that the clip snippets have complex background and turning faces (not frontal face, which is hard for face detector, like the well-known face detector proposed by Viola *et al.* [Viola 2004]) in the frame. We hand-labeled the positions of all the faces in each frame to avoid the error from mislabel faces, before the selection of talking face.

The flow chart of the audio-visual saliency model of musical instrument is shown in Fig. A.11, and the algorithm of fusion strategy II is described briefly below:

1. If it is the starting frame of music, fusion strategy II begins to work to locate sound source (musical instrument). The localization of the sound source comes from the dynamic/motion saliency map. More precisely, the sound source is the region characterized by 20% pixels with highest values in the dynamic/motion saliency map. If the region of 20% pixels with highest values is not connex, region with higher spatial mean value is considered as sound source.

2. Then, the Euclidean distance between the center of sound source and each individual face in the frame is calculated. This distance is recorded as $D_{mf_i}(k)$ (i is the index number of the face, and k is the number of frame). k=1 is the starting frame of this music period.

3. After, we calculate the temporal mean of $D_{mf_i}(k)$ (k=1,2...7) (recorded as $\overline{D}_{mf_i}$) for each individual face. Duration of seven frames is chosen for the reason that it is equal to the mean fixation duration value.

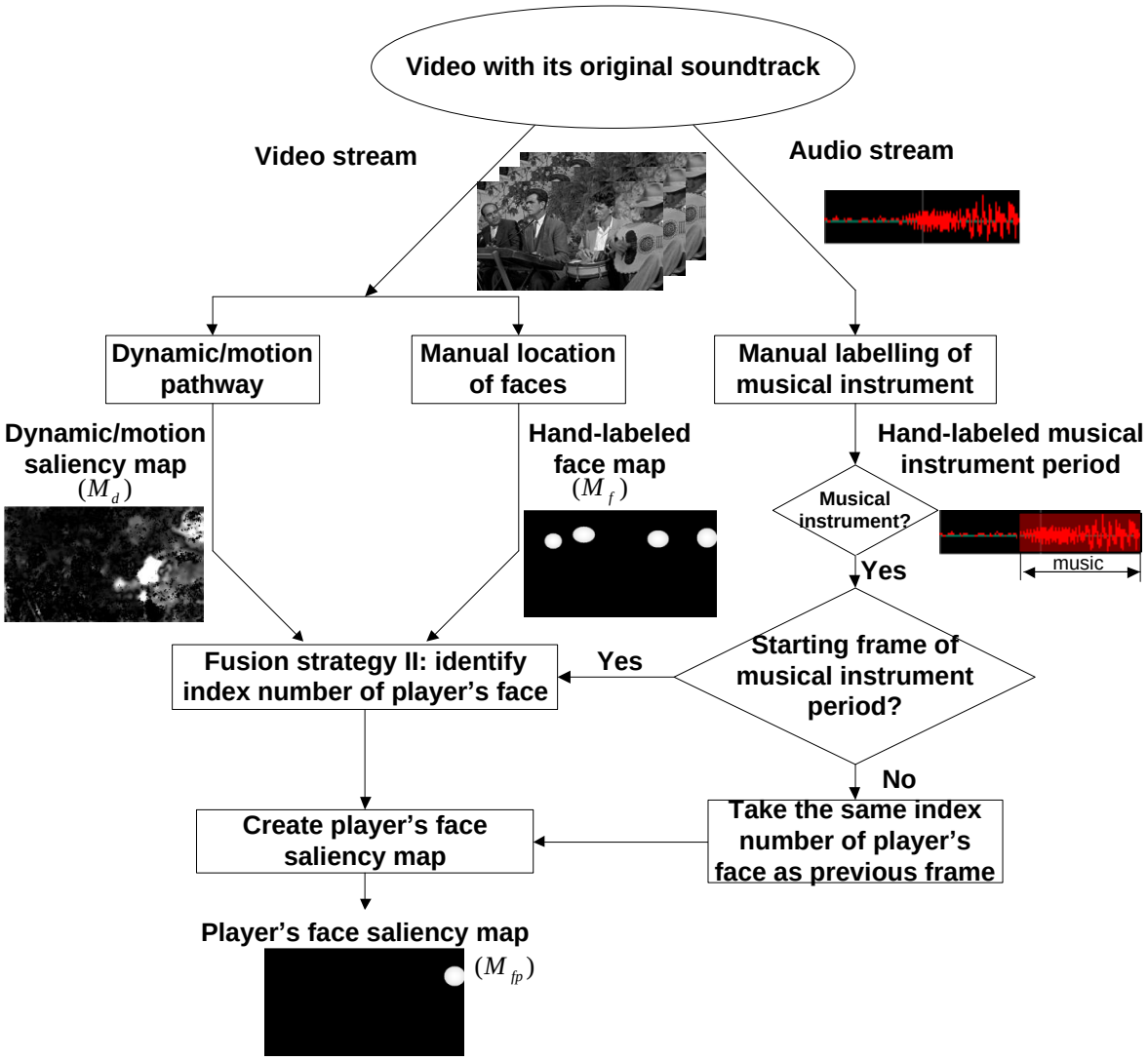


Figure 5.11: Flow chart of proposed saliency model for musical instrument class.

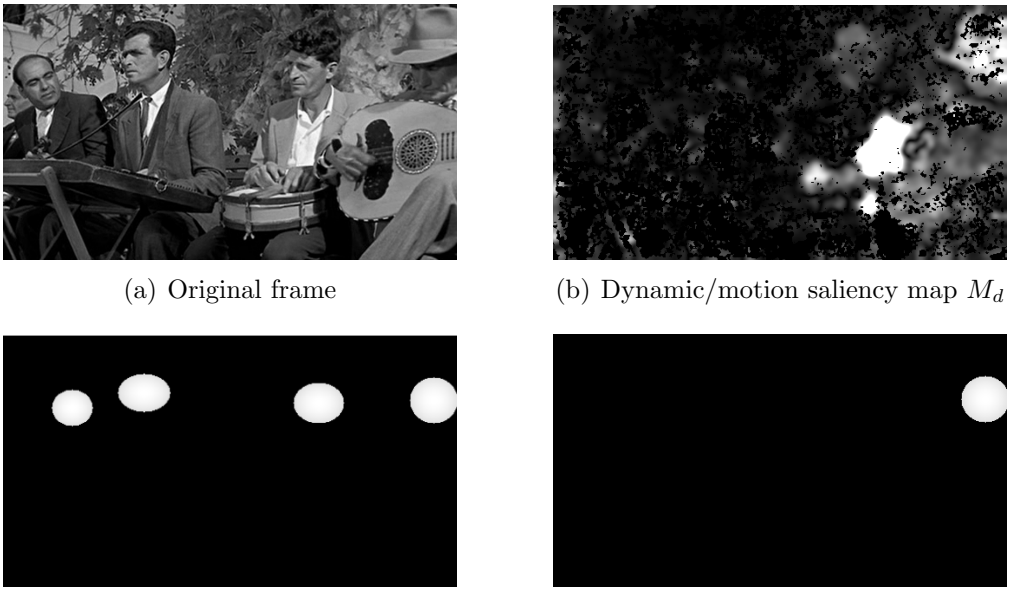
4. We considered the face region, which has lowest value of $\overline{D}_{mf}$ as player's face region. If there are more than one face having lowest $\overline{D}_{mf}$ value, face with higher motion value will be considered as player's face. Once the player's face is selected at the beginning of music period, this selection will be kept over all this music period.

5. At last, we added a 2-D Gaussian to the center of selected player's face region to create *player's face saliency map* M_{fp}.

This fusion strategy II to select player's face from all the faces, is tested on the three clip snippets in musical instrument class. The player's faces in these three clip snippets are 100% correctly selected.

5.3.3 Comparison with visual saliency model

In order to test the accuracy of prediction of proposed player's face saliency map M_{fp}, we compare M_{fp} with other two visual saliency maps: dynamic/motion saliency map M_d and hand-labeled face map M_f. To evaluate the prediction accuracy of each saliency map, the same criteria are chosen: NSS and TC. Fig. 5.12 shows a frame example of M_d, M_f and M_{fp}.



(a) Original frame

(b) Dynamic/motion saliency map M_d

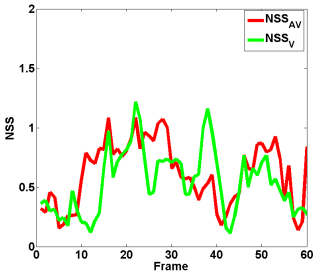
(c) Hand-labeled face map M_f

(d) Player's face saliency map M_{fp}

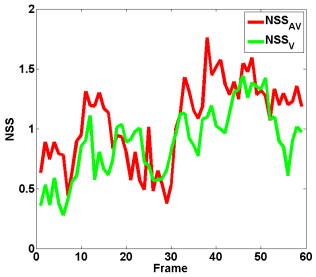
Figure 5.12: A frame example of dynamic/motion saliency map M_d, hand-labeled face map M_f and player's face saliency map M_{fp}.

First, M_d, M_f and M_{fp} are evaluated by criterion NSS. Fig. 5.13 shows the average value of NSS of testing data of three clip snippets with group of participants with AV condition (respectively with V condition). All the clip snippets are synchronized with the starting frame of music. In (a) and (b), the performance of NSS_{AV} and NSS_V are similar: stable over time and lower NSS value. In (c), NSS_{AV} increases sharply after the stimulus of music sound, suggesting that player's face attracts attention for participants with AV condition. However, NSS_V is stable over time, suggesting that player's face has no particular attraction for participants with V condition.

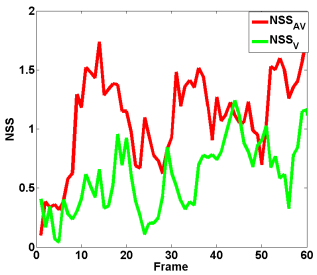
Then, to confirm the result, M_d, M_f and M_{fp} are evaluated by criterion TC. Fig. 5.14 shows the average value of TC of testing data of three clip snippets. All the clip snippets are synchronized with the starting frame of music. In (c), both TC_{AV} and TC_V perform similar to NSS_{AV} and NSS_V (in Fig. 5.13). Hand-labeled face map M_f (in (b)) gets highest difference between NSS and TC. For TC, it only takes into account 20% of most salient regions. Hence, in M_f, only small region of the center of each face is considered.



(a) Dynamic/motion saliency map M_d evaluated by NSS

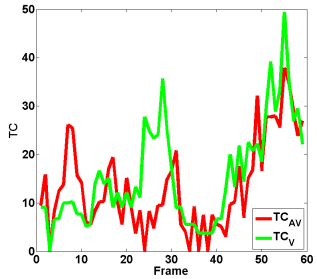


(b) Hand-labeled face map M_f evaluated by NSS

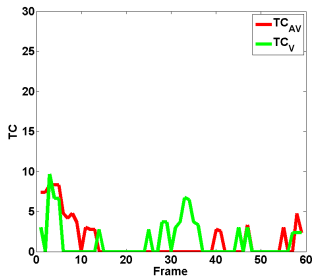


(c) Player's face saliency map M_{fp} evaluated by NSS

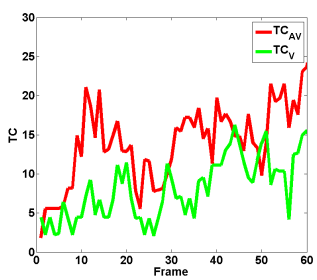
Figure 5.13: Results of prediction accuracy of mean of three clip snippets for M_d, M_f and M_{fp}, evaluated by NSS. When maps are compared with group with AV condition (respectively with V condition), results are called NSS_{AV} (NSS_V). Frame 1 is the starting frame of music.



(a) Dynamic/motion saliency map M_d evaluated by TC



(b) Hand-labeled face map M_f evaluated by TC



(c) Player's face saliency map M_{fp} evaluated by TC

Figure 5.14: Results of prediction accuracy of mean of three clip snippets for M_d, M_f and M_{fp}, evaluated by TC. When maps are compared with group with AV condition (respectively with V condition), results are called TC_{AV} (TC_V. Frame 1 is the starting frame of music.

At last, we observe in detail the same period from frame 6 to 30 (1 second after speech stimuli). Table 5.2 shows the mean value of NSS and TC from frame 6 to 30 after the starting frame of music of group of participants with AV condition (respectively with V condition). Compared to dynamic/motion saliency map and hand-labeled face map, player's face saliency map performs best in group of participants with AV condition both evaluated in NSS and TC. However, when compared to group with participants with V condition, this player's face saliency map does not perform better than the other two maps.

Table 5.2: The mean value (from frame 6 to 30 after the starting frame of music) of NSS and TC when compared with dynamic/motion saliency map M_d, hand-labeled face map M_f, and player's face saliency map M_{fp} presented in Fig. 5.13 and 5.14

Saliency map	M_d		M_f		M_{fp}	
Condition	AV	V	AV	V	AV	V
NSS	0.76	0.57	0.82	0.74	1.05	0.46
TC	9.51	13.75	1.03	0.51	12.16	6.06

5.3.4 Conclusion

We observed that the player's face was more attractive than sound source (musical instrument) and also than other faces in the frame for participants with AV condition in musical instrument class. We proposed the fusion strategy II, which fused hand-labeled face map with dynamic/motion saliency map, which is the output of dynamic/motion pathways in Marat's *et al.* visual saliency model. Sound source (musical instrument) is the region characterized by the 20% pixels with highest values in the dynamic/motion saliency map. After, the face, which was closer to the musical instrument, was considered as player's face. Compared to condition V, this player's face saliency map increased accuracy of prediction for participants with AV condition after stimulus of music. Also, this player's face saliency map performed better than hand-labeled face map and dynamic/motion saliency map both evaluated by NSS and TC.

5.4 General conclusion

This chapter presented a preliminary audio-visual saliency model with two different fusion strategies of speech (fusion strategy I) and musical instrument (fusion strategy II) sound. Based on knowledge that sound source is tightly linked to motion, and faces attract more attention than other objects, we fused dynamic/motion saliency map (from a visual saliency model proposed by Marat's *et al.*) and hand-labeled face map.

Fusion strategy I of speech class fused dynamic/motion saliency map with hand-labeled face map to select talking face as salient region automatically. This talking face saliency map performed better compared to other maps separately both evaluated by criteria of NSS and TC in the database of speech class. Also, this talking face saliency map increased accuracy of prediction for participants with AV condition compared to V condition.

Fusion strategy II of musical instrument class selects player's face as salient region automatically, by fusing dynamic/motion saliency map with hand-labeled face map. This player's face saliency map performed better compared to other maps separately in the database of musical instrument class. Also, it increased accuracy of prediction for partic-

ipants with AV condition compared to V condition.

Chapter 6

Conclusions and perspectives

In this thesis, we focused on better understanding the influence of sound in videos on eye movement. We began our study through an eye tracking experiment to answer the questions: "Is there an influence of sound on eye movement in videos? If yes, does different types of sound influence eye movement in videos differently?". Two audio-visual experiments were designed to explore the influence of audio-visual interaction on eye movement. To observe this influence, experimental eye positions of participants were recorded and analyzed. Finally, from observations during the experiments, a preliminary audio-visual saliency model was proposed to predict salient regions in videos with soundtrack. In the following, the main contributions and several perspectives are summarized.

6.1 Conclusions

In this study, two audio-visual experiments were designed to explore the influence of audio-visual interaction in human behavior. Short video excerpts with their original soundtrack were selected as experimental video data in the experiments. The video data were presented with two conditions: with their original soundtrack (audio-visual (AV) condition); the same video data without any sound (visual (V) condition). Through the analysis of the difference between the eye positions of participants with AV and V conditions, several results were observed.

Audio-visual experiment I: is there an influence of sound on eye movement in videos?

- *Sound influences eye movement in videos.* In audio-visual experiment I, a group of participants watched video data with AV condition and another group of participants watched the same video data with V condition. We analyzed the difference of eye positions between group of participants with AV and V conditions, and observed

the existence of sound influence on gaze when looking at videos.

Moreover, *different types of sound influenced eye movement differently.* We defined three classes of sound manually: on-screen speech class, non-speech class and non-sound class (intensity below 40 dB). Difference of eye positions between the group of participants with AV and V conditions was highest in on-screen speech class and obviously lowest in non-sound class.

- *Sound affects the prediction accuracy of a visual saliency model.* We compared the experimental eye positions of group of participants with AV condition (respectively with V condition) with Marat's visual saliency model. The results showed that prediction accuracy of visual saliency model decreased in group of participants with AV condition. Especially, when the sound was on-screen speech, the decreasing of prediction accuracy was significant at a level of 1%.

- *Sound source attracts attention.* We investigated the frames with only one sound source in the screen and located the coordinates of the sound source manually to create the sound saliency maps. By Comparing the experimental data of the eye positions (groups with AV and V conditions) and the sound saliency maps, the prediction accuracy increases significantly, when soundtrack is on-screen speech.

Audio-visual experiment II: which type of sound influences eye movement in videos?

Experiment I showed that sound influenced eye movements differently depending on the sound type. To enrich our study of this influence of audio-visual interaction on eye movement, the second audio-visual experiment (experiment II) was designed and considered.

- *Human voice affects visual gaze in videos.* Video data in experiment II contained thirteen more refined sound classes. We investigated the influence of thirteen types of sound on gaze separately, through the analysis of the difference of eye positions between groups with AV and V conditions. The results confirmed that the effect of sound was different depending on the kind of sound, and the classes with human voice cluster (*i.e.* speech, singer, human noise and singers) had the greatest effect.

- *Participants move their eyes to the sound source in human voice cluster.* We assumed that the sound source in the frame attracted attention and therefore calculated the distance between sound source and eye positions of the group of participants with AV condition. The results showed that sound source significantly attracted human eye position only when the sound was human voice cluster.

- *Participants with AV condition had a shorter average duration of fixation than with V condition.* A typical model (paired t-test) and a more recent model (mixed-effect model) were adopted in the calculation. Both results showed that participants with AV condition had a shorter average duration of fixation than V condition, suggesting that participants with AV condition move their eyes more frequently than participants with V condition.

- *Sound reduces the prediction accuracy of visual saliency models.* Comparisons of the experimental eye positions with Marat's and Itti's visual saliency models were calculated. The prediction accuracy of both saliency models decreased more in group with AV condition than with V condition during frames 6 to 56 after the appearance of the second sound.

Preliminary audio-visual saliency model

In previous analysis, we found that the sound source saliency maps increased prediction accuracy of group of participants with AV condition. Because in our video data, there is no spatial information contained in the sound signal, it is a difficult task to locate sound source in the visual scene automatically. We propose another approach. We assume that visual motion is in conjunction and synchronized with associated soundtrack. In order to improve the prediction accuracy of a visual saliency model with AV condition, by adding sound information, a preliminary audio-visual saliency model is proposed. More precisely, two fusion strategies for speech and musical instrument sound classes are proposed in this model:

Because prediction accuracy of visual saliency model decreased with AV condition, based on the knowledge acquired from the above experiments, we proposed a preliminary audio-visual saliency model. In this model, by fusing audio and visual information, two fusion strategies for speech and musical instrument sound classes were proposed to improve the prediction accuracy.

- *For speech class, identify the talking face.* We focused on clip snippets with speech soundtrack and multiple faces in the scene, but only one face is talking. Based on the *hypothesis* that talking face is moving, we proposed to fuse face map (hand-labeled) and dynamic saliency map (from Marat's visual saliency model), to locate the talking face as salient region automatically. This talking face saliency map performs better than face map and dynamic saliency map separately.

- *For musical instrument class, identify the player's face.* We focused on clip snippets with music soundtrack and multiple faces in the scene, but only one face is player's

face. An additional hypothesis was that the Euclidean distance between musical instrument and the player's face is shortest, compared to other faces in the frame. Then, a fusion strategy of face map and dynamic saliency map was proposed to detect player's face automatically. This player's face saliency map performs better than face map and dynamic saliency map separately.

6.2 Perspectives

This work has several perspectives focused on three aspects: eye tracking experiment (to deeper analyze sound influence on gaze), more efficient audio-visual saliency model and possibilities for using in other applications.

Test with non-original soundtrack in audio-visual experiment

In the audio-visual experiment I and II, influence of different sound types are not compared directly, but through the comparison between participants with AV and V conditions of each sound type. In future experiments, it will be interesting to compare the effect of different sound types directly. New conditions of A and V can be created with non-original soundtracks. As a paradigm, different types of sound can be added to the same video excerpt to create different AV conditions with non-original soundtracks. This non-original soundtrack method is already adopted by researches [Vilaró 2012] to investigate how the soundtrack influences perception and comprehension of the scene. With these non-original soundtracks, we can investigate how the participants manage the conflicts of audio and visual contents. However, the influence of non-original soundtrack maybe different according to whether the soundtrack could correspond or not to a visual event in the video [Hidaka 2010, Gordon 2011]. Moreover, if the soundtrack and its corresponding visual event in the scene is unsynchronized, it would be interesting to investigate whether the influence of sound on eye movement would be different.

Improvement of audio-visual saliency model

- *Face detector:* In the preliminary audio-visual saliency model, all faces are hand-labeled. Before hand-labeling all faces in the frames, we tested the well-known face detector proposed in [Viola 2004]. However, faces in films usually are not front faces and the background is complex. In this difficult context, this face detector is not reliable enough. To solve this problem and increase the performance in videos of Viola-Jones algorithm, some researchers introduced a face tracking process after the detection of faces in the algorithm [Cao 2009]. Adding AdaBoost (Adaptive

Boosting) learning to the Viola-Jones algorithm is also a possible way to increase the detection of faces [Zhang 2010]. It is a possible way to improve face detection.

- *Classification of sound type:* We concluded that different sound class affects human gaze differently. But for the moment, soundtrack is manually labeled. For an automatic audio-visual saliency model, it is possible to introduce a sound classification process. For this process, different audio features can be extracted as for example Mel Frequency Cepstral Coefficients (MFCC) as proposed in [Feng 2011, Feki 2011]. For the classification, different algorithms exist, like Hidden Markov Model (HMM), Support Vector Machine (SVM), neural network [Tkac 2011]. The sound classification is particularly complex when there are mixtures of different sounds (for example, speech and background music). Moreover, *sound transition detector* between two successive sounds is also a possible process to be added in the model.

- *Robust fusion strategy:* The algorithms of face detection and sound classification above will bring misclassifications. In this case, to reduce the influence of the misclassification, we can introduce weighted fusion strategy by using confidence values of the detection and the classification. Also, if there are more than one sound source in the screen, the weight between audio and visual saliency map could be different.

- *Sound source localization in the image:* Audio signals in experiments I and II contain no spatial information. In a future work, by using data containing spatial audio information, it would be interesting to locate one (or more) sound source. In this case, it would be possible to investigate audio-visual saliency models based on the fusion of sound source location and visual saliency maps, as already proposed in a robotic application [Ruesch 2008].

Appendix A

Résumé en Français

Dans la vie quotidienne, nous recevons une grande quantité d'informations provenant de l'environnement en utilisant nos cinq sens: la vision, l'audition, le goût, l'odorat et le toucher. Parmi ces cinq sens, nous dépendons davantage du sens de la vision. En effet, environ 80% de l'information sur l'environnement est acquise par la vision [Begbie 1996]. Par exemple, lors d'une tâche de navigation, la vue nous permet d'éviter les obstacles. La perception visuelle est une tâche complexe, qui est constituée d'un grand nombre de mécanismes. Dans le cerveau, le cortex visuel est responsable du traitement de cette entrée visuelle. Le cortex visuel primaire transmet des informations à deux voies principales: l'une appelée voie dorsale, qui est associée au mouvement, la représentation de l'emplacement des objets, et le contrôle des yeux; l'autre est la voie ventrale, qui est associé à la reconnaissance des formes et à la représentation des objets.

Des quantités importantes d'informations visuelles atteignent nos yeux à chaque instant, mais notre capacité visuelle n'est pas infinie. Afin de pouvoir réagir rapidement et correctement dès la réception des informations de l'environnement, il existe des mécanismes dans notre cerveau pour identifier un sous-ensemble d'information sensorielle essentielle d'une scène avant de poursuivre son traitement. Ces mécanismes guident l'attention vers des régions particulières. Les yeux vont s'orientés vers des régions particulières appelées *régions saillantes* qui attirent l'attention. Les capteurs pour la vision sont les yeux, qui permettent l'entrée de la lumière et sa conversion en impulsions électro-chimiques dans les neurones. Dans l'œil, des images haute résolution sont fournies par le centre de la rétine appelée macula, qui est responsable de la vision centrale. Un champ visuel plus large avec une résolution inférieure est fourni par la partie restante de la rétine.

L'étude du mouvement des yeux permet une meilleure compréhension du système visuel et des mécanismes dans le cerveau pour sélectionner les régions saillantes. La modélisation de l'attention visuelle permet de prédire les régions saillantes. Il y a beaucoup d'applications à ce type de modèles. Par exemple la sélection des régions saillantes peut

être utilisé pour régler le niveau de compression dans des vidéos, ou pour guider le mouvement d'un robot mobile.

L'*audition* est également un sens important pour recueillir des informations dans l'environnement. Par exemple, lors d'une navigation, des alarmes sonores peuvent aussi nous aider à éviter les obstacles. Dans le cerveau, le cortex auditif est une région qui traite le son et contribue ainsi à la capacité à entendre. Les neurones du cortex auditif primaire peuvent être considérés comme ayant des champs récepteurs couvrant une gamme de fréquences sonores de telle manière que les neurones d'un coté du cortex auditif répondent à des fréquences basses, et ceux de l'autre coté répondent à des fréquences élevées. la partie restante du cortex auditif intervient dans les traitements suivants et distingue les types de sons: la parole, la musique ou le bruit. Afin de pouvoir réagir rapidement après avoir entendu le bruit de l'environnement, il existe aussi des mécanismes d'attention dans le cerveau pour orienter l'attention vers les événements saillants particuliers dans le domaine audio. La modélisation du système d'attention auditive doit être capable de prévoir ces événements saillants, et peut être également appliquée à la détection d'événements, tels que la parole ou la musique.

A.1 Problèmes

Nos différents sens reçoivent des informations corrélées correspondant aux mêmes objets ou événements. Ces informations sont combinées dans notre cerveau. Par conséquent, le comportement de l'homme n'est pas influencé par un seul sens, mais par l'interaction de plusieurs sens. Il est donc important notamment d'étudier comment la vision interagit avec l'audition. Les premières recherches ont considéré qu'un sens est séparé des autres modalités sensorielles. Ainsi l'intégration de caractéristiques provenant d'une seule modalité (visuelle ou auditive) a été beaucoup étudiée.

Des études récentes d'intégration multimodale ont été réalisées. Dans [Quigley 2008], l'influence de l'interaction audiovisuelle sur le mouvement des yeux a été abordée. Cette étude était dédiée au mécanisme de l'influence de la position de la source sonore sur le mouvement des yeux. Le son était diffusé par des haut-parleurs placés aux quatre coins d'un écran et les stimuli visuels étaient des images statiques. Cependant, l'effet du son sur le regard dans le cas de vidéos est peu étudié:

- Est-ce que le son a une influence sur le mouvement des yeux, quand on regarde des vidéos (stimuli dynamiques et complexes) avec la bande son originale?
- Est-ce que cette influence est différente selon le type de son?

Quelques modèles de saillance audiovisuelle qui simulent le comportement sous l'influence des interactions audiovisuelles, ont été utilisées dans certaines applications, par exemple pour sélectionner des images clés dans des vidéos [Lee 2011, Wang 2012]. Dans ces exemples, un modèle de saillance visuelle est utilisé pour prédire les régions saillantes dans des frames (ou images). Un modèle de saillance audio est utilisé séparément pour prédire les événements audio saillants. Chaque modèle fournit une valeur de saillance unidimensionnelle pour chaque frame (donc des informations spatiales fournies par le modèle de saillance visuelle sont perdues). Les images clés sont sélectionnées selon la combinaison des courbes de saillances visuelle et audio.

A.2 Objectifs

Le premier objectif de cette thèse est de fournir une meilleure compréhension de l'influence de l'interaction audiovisuelle sur le regard humain. Pour cet effet, nous avons conçu des expériences audiovisuelles pour étudier l'influence du son sur le regard de l'homme dans des vidéos. Plus précisément, nous cherchons des réponses aux deux questions décrites dans la section précédente.

De plus, à l'aide des connaissances acquises dans l'expérience, nous cherchons à améliorer un modèle de saillance visuelle existant en ajoutant une voie audio supplémentaire. L'objectif est de proposer un modèle de saillance audiovisuelle qui prédit plus précisément les régions saillantes pour les vidéos avec leur bande son originale.

A.3 Contributions

Dans cette thèse, nous nous intéressons à une meilleure compréhension de l'influence du son sur le mouvement des yeux dans des vidéos. Nous débutons notre étude par une expérience de suivi oculaire pour répondre aux questions suivantes : « Est-ce que le son dans les vidéos influence le mouvement des yeux? », « Si oui, quels types de son ont une influence sur le mouvement des yeux dans les vidéos? ». Deux expériences audiovisuelles sont conçues pour étudier l'influence de l'interaction audiovisuelle sur le mouvement des yeux. Pour observer cette influence, les positions expérimentales des yeux des participants ont été enregistrées et analysées. Enfin, à partir de ces observations, un modèle préliminaire de saillance audiovisuelle est proposé pour prédire des régions saillantes dans les vidéos avec la bandes son originale.

Expérience audiovisuelle I : Est-ce que le son dans les vidéos influence le mouvement des yeux?

- *Description de l'expérience audiovisuelle I.*

  Dans cet expérience, une base de données composée de soixante morceaux de vidéos de durée de 5 à 8 secondes, morceaux appelés « extraits (ou clip snippets) », ont été sélectionnés à partir de sources de films hétérogènes. Chaque clip est composé de 6 extraits (ou clip snippets), qui proviennent de différents films. La Fig. A.1 montre le contenu de chaque extrait dans le « clip 1 ».

  Cette base de données a été créée avec deux conditions : condition audiovisuelle AV, les données vidéo avec des bandes son originales; Condition visuelle V, les mêmes données video en éliminant les sons. La Fig. A.2 illustre l'évolution temporelle de ces essais expérimentaux. Les participants ont été invités à regarder les dix clips sans aucune tâche particulière. Un groupe des participants a regardé les données vidéo avec la condition AV, l'autre groupe a regardé les mêmes données vidéo avec la condition V. Les dix séquences ont été présentées à chaque participant dans un ordre aléatoire. Les positions oculaires ont été enregistrées par un oculomètre Eyelink II.

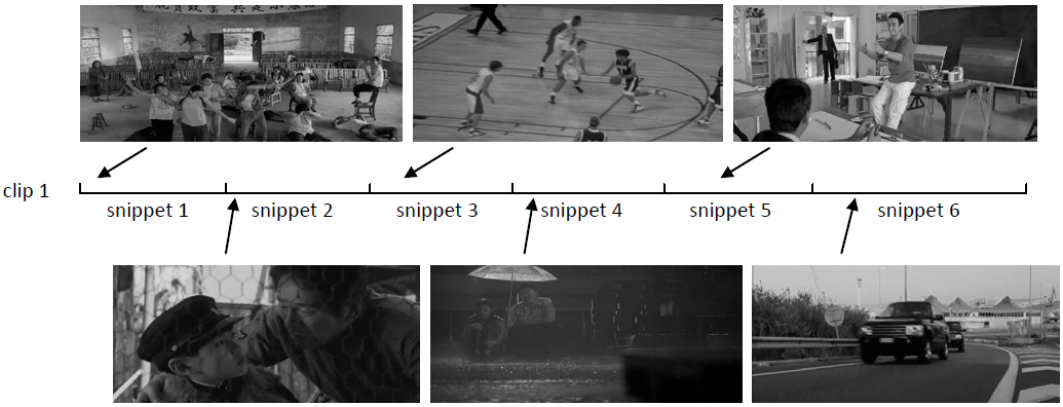


Figure A.1: Le contenu de chaque extrait dans le « clip 1 ». Chaque extrait (ou clip snippet) provient de différents films.

  Voici un exemple de positions des yeux dans la Fig. A.3.

- *Le son a une influence sur le mouvement des yeux dans des vidéos.*

  En observant qualitativement les positions des yeux des participants, nous concluons que les différents types de son ont une influence différente sur les positions des yeux. Par conséquent, nous faisons une classification manuelle de toutes les positions des yeux dans trois classes selon le type de sons : parole à l'écran (le son provient d'un locuteur visible sur l'écran), non-parole (signal audio de type quelconque sauf la parole) et non-son (sons d'intensité inférieure à 40 dB).

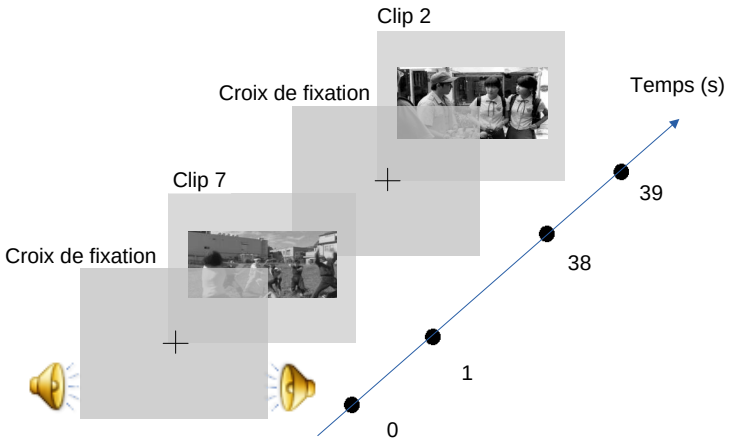


Figure A.2: Évolution temporelle de deux clips de la condition AV. Pour contrôler le regard du participant, une croix de fixation est placée au centre de l'écran avant chaque clip. Cette séquence est répétée avec les dix clips présentés dans un ordre aléatoire pour chaque participant.

Figure A.3: Un exemple de positions des yeux expérimentales de deux groupes de participants. Les points rouges représentent la position des yeux de participants dans le groupe avec la condition AV, et les points verts représentent les positions des yeux de participants dans le groupe avec la condition V.

Afin de mesurer les différences de position des yeux entre le groupe avec la condition AV et celui avec la condition V, deux mesures différentes ont été envisagées: la Distance Médiane md et le Coefficient de Corrélation linéaire cc.

La Fig. A.4 montre les résultats du test ANOVA des trois classes [1]. Le résultat $F(2, 742) = 9.24$ et $p < 10^{-4}$ indique que, parmi les trois classes: parole à l'écran, non-parole et non-son, au moins la valeur moyenne d'une classe est significativement différente de celles des deux autres classes. De plus, la valeur moyenne de md a tendance à décroitre à partir du groupe avec de la parole à l'écran, vers le groupe non-son. La valeur moyenne du groupe avec de la parole à l'écran est très différente de celle des deux autres classes: parole à l'écran et non-parole ($F(1, 673) = 12.27$,

[1] Dans la figure, '*' indique que la valeur de p est $< 0,05$, '**' indique que la valeur de p est $< 0,01$, '***' indique que la valeur de p est $< 0,001$.

$p < 10^{-3}$), parole à l'écran et non-son ($F(1, 420) = 10.44$, $p < 10^{-2}$). Elle obtient la valeur la plus élevée entre ces trois classes avec les données obtenues par la distance médiane *md* et elle correspond à la différence la plus grande entre les groupes avec AV et V conditions. Entre les groupes non-parole et non-son ($F(1, 391) = 1.99$, $p = 0.16$), la différence n'est pas significative avec le test ANOVA.

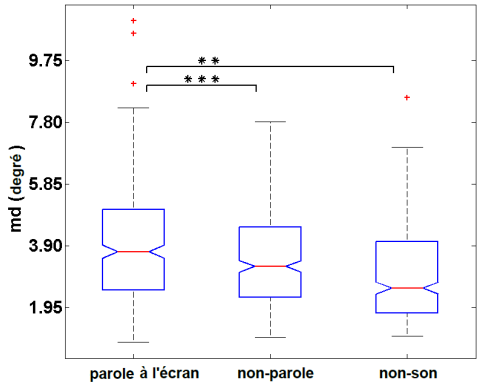


Figure A.4: Comparaison (test ANOVA) sur les distances médianes *md* entre positions oculaires calculées entre les deux groupes de participants (avec conditions AV et V) pour trois classes de sons: parole à l'écran, non-parole et non-son.

De plus, le résultat de *cc* confirme que la classe avec de la parole à l'écran a la différence la plus élevée des positions des yeux entre le groupe des participants avec la condition AV et celui avec la condition V . La classe non-son a la différence la plus petite.

- *Le son a un effet sur la précision de la prédiction d'un modèle de saillance visuelle.*

  Pour compléter l'analyse, nous avons étudié dans ce paragraphe s'il existe un effet sonore sur un modèle de saillance visuelle. Afin d'évaluer si la précision de prédiction de modèle de saillance visuelle décroit, nous comparons les positions expérimentales des yeux du groupe avec la condition AV (également avec la condition V) avec un modèle de saillance visuelle.

  Pour le modèle de saillance visuelle, nous avons choisi le modèle de saillance spatio-temporelle développé par Marat *et al.* [Marat 2009]. Il est inspiré par la biologie des premières étapes du système visuel humain, et est composé de deux cartes de saillance: carte statique (sortie de la voie statique) et carte dynamique (sortie de la voie dynamique). La voie statique du modèle de saillance visuelle est constituée de deux types d'interaction basée sur le rayon des champs de réception. La carte de saillance statique représente principalement le bord des objets, qui ont un grand contraste spatial. La voie dynamique est étroitement liée au mouvement et

en particulier au déplacement d'une région par rapport au fond. La carte de saillance dynamique est sensible à l'amplitude de mouvement sur le fond et non pas à l'orientation du mouvement. Pour l'évaluation, nous avons choisi le critère de "Normalized Scanpath Saliency" (NSS) qui a été proposé par Peters et Itti [Peters 2005]. Il est conçu particulièrement pour comparer les positions des yeux avec les zones saillantes extraites par un modèle de saillance. Nous avons comparé les positions des yeux expérimentales des groupes des participants avec la condition AV (également avec la condition V) avec le modèle de saillance visuelle proposé par Marat *et al.*. Les résultats ont montré que la précision de la prédiction du modèle de saillance proposé par Marat *et al.* diminue lorsqu'il est testé sur les vidéos avec des bandes son originales. La diminution de la précision de la prédiction est apparue à la fois dans la voie statique et la voie dynamique. De plus, la diminution de la précision de prédiction était différente pour les différents types de son. Pour la classe avec de la parole à l'écran, la diminution de la précision de prédiction était significative à un niveau de 1%. Toutefois, pour les deux autres classes: celle non-parole et celle non-son, la précision de prédiction n'était pas significativement différente entre les conditions AV et V.

- *La source du son peut attirer le regard.*

  Comme nous avons conclu que la précision de la prédiction d'un modèle de saillance visuelle diminuait quand il a été appliqué sur les données vidéo avec des bandes son originales, nous avons essayé de trouver une méthode pour compléter ce modèle de saillance visuelle pour améliorer la précision quand il est utilisé avec des vidéos avec des bandes son originales.

  Selon nos observations, la source du son dans la vidéo attire le regard. Pour simplifier le problème, nous ne considérons que les extraits de clip avec une seule source de son à chaque frame. Par conséquent, nous avons marqué manuellement les coordonnées de la source de son et l'avons appelé « voie de la localisation des sons ». Ensuite, nous appliquons une fonction gaussienne bi-dimensionnelle centrée sur la position de la source sonore afin d'obtenir une carte de saillance son. Enfin, nous comparons les données expérimentales de la position des yeux (groupe avec la condition AV et celui avec la condition V) et les cartes de saillance sonores en utilisant la méthode NSS .

  Les résultats ont montré que la précision de prédiction du modèle de saillance peut être augmentée par l'ajout d'une voie son en localisant la source sonore. La carte de saillance sonore, créé par une fonction Gaussienne bi-dimensionnelle de taille appropriée appliquée sur la source sonore, a augmenté la précision de la prédiction

dans la condition AV . Cette augmentation a été significative lorsqu'il y avait une parole à l'écran. Cependant, pour la classe non-parole, la précision de prédiction de cette carte de saillance n'était pas significativement différente entre les conditions AV et V.

Expérience audiovisuelle II : Quel type de son influence le mouvement des yeux dans les vidéos?

Grâce à l'analyse de la première expérience audiovisuelle, nous avons observé que le son influence différemment le regard de l'homme dans les vidéos en fonction du type de son, et cet effet est plus important pour la classe avec des paroles à l'écran. Nous ne considérons que trois classes de sons sans aucun contrôle strict de l'événement sonore au fil du temps.

Pour étudier plus profondément l'influence du son dans les vidéos sur le regard, une deuxième expérience audiovisuelle est présentée dans cette section pour répondre à la question : quel type de son influence le regard de l'homme ? Nous comparons le comportement du regard humain avec treize classes de sons définies plus finement. Les vidéos extraits sont choisis de sorte que le début d'un son se produit au milieu d'une scène visuelle. De cette façon, nous évitons un changement simultané du contenu de la scène visuelle et de la bande son. L'objectif est d'analyser l'effet du son en comparant les positions des yeux avec les conditions AV et V.

Par conséquent, nous avons conçu une nouvelle expérience audiovisuelle de deux groupes de participants avec les conditions audiovisuelle (AV) et visuelles (V). Ensuite, nous avons comparé la différence des positions des yeux du groupe avec la condition AV et le groupe avec la condition V des treize classes de son séparément. Pour étudier l'endroit où les humains regardent après l'apparition des stimuli auditifs, nous avons analysé la distance entre la source sonore et les positions des yeux. Ensuite, les durées de fixation entre les groupes avec des conditions AV et V sont analysées. Enfin, les positions expérimentales des yeux sont comparées avec les régions prédites par deux modèles de saillance visuelle (Marat *et al.* et Itti *et al.*). Une partie des résultats a été publié dans [Song 2012].

- *Description de l'expérience audiovisuelle II.*

  Cette expérience audiovisuelle a été conçue pour étudier quel type de son influence le regard dans les vidéos en observant les positions des yeux des participants. Le principe de la conception de l'expérience est que chaque participant regarde la moitié des extraits vidéo en version originale (condition AV) et l'autre moitié des extraits vidéo sans bande sonore (condition V). Ensuite, nous étudions l'effet sonore à travers l'analyse de la différence des positions des yeux entre ces deux groupes de participants.

Par rapport à l'expérience I, ces données vidéo ont contenu treize classes plus raffinées de sons (description de chaque classe de son est présentée sur la Fig. A.5). Cette classification est vérifiée par une pré-expérience.

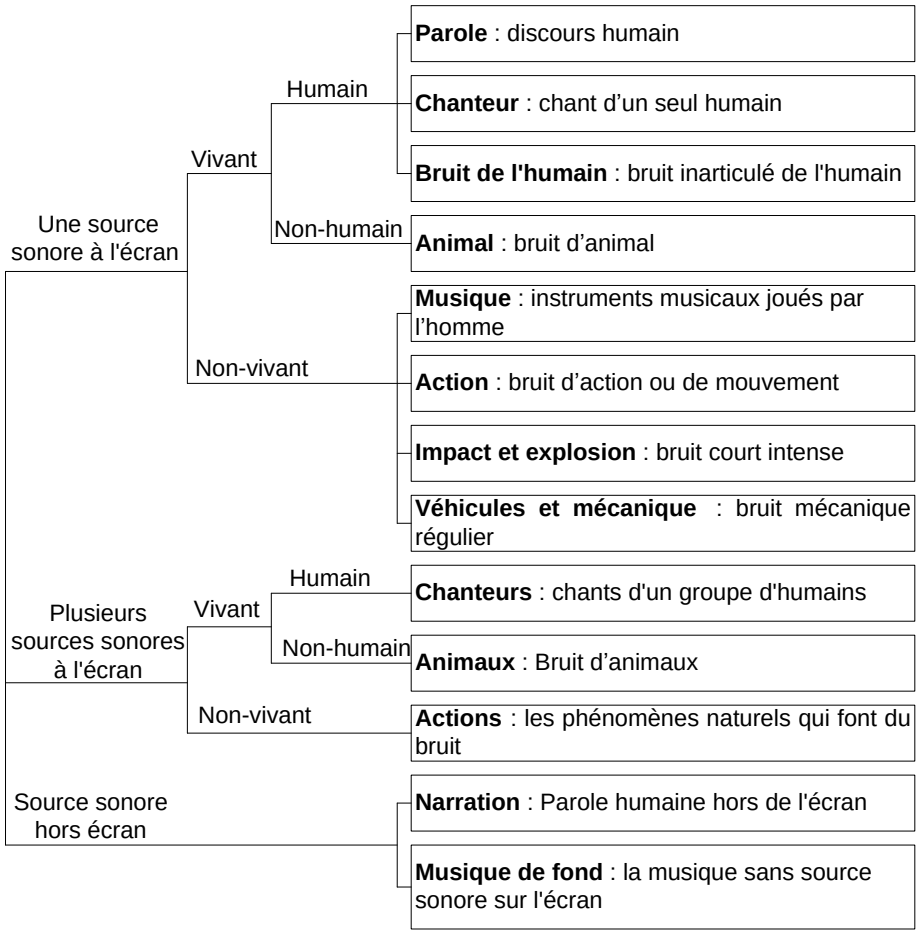


Figure A.5: Classification hiérarchique du deuxième son

Une autre différence était que les extraits vidéo ont été choisis avec la présence du son qui correspondait à la scène visuelle et un changement de son s'est produit au milieu d'extraits (appelé deuxième son) pour éviter le changement simultané des contenus visuel et audio. Cette conception évite aussi l'effet du biais central [Dorr 2010]. Un exemple d'un extrait (ou clip snippet) est présenté dans la Fig. A.6. Dans cette expérience, chaque participant a regardé la moitié des extraits vidéo avec la condition AV et l'autre moitié avec la condition V.

- *La voix humaine dans les vidéos influence le regard.*

  Afin d'étudier l'effet du son sur le regard visuel, nous avons à souvean comparé les positions des yeux des participants avec la condition AV et ceux avec la condition

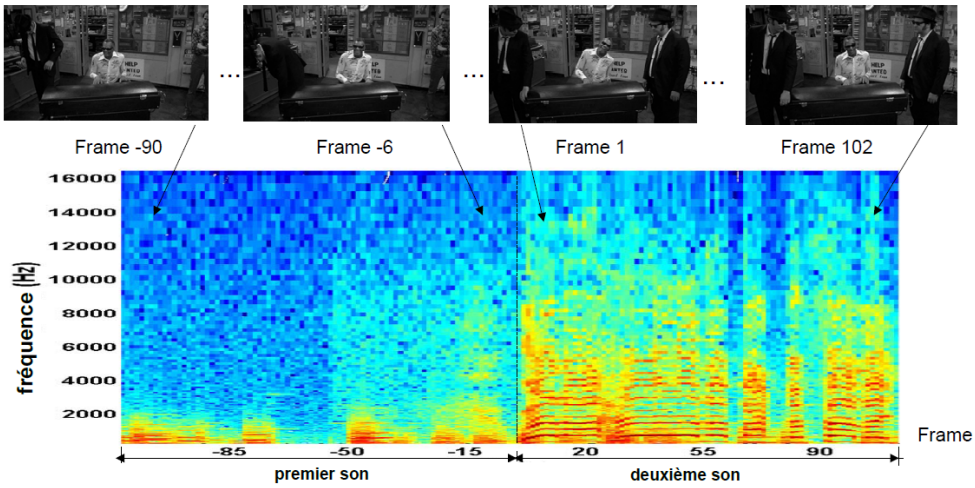


Figure A.6: Un exemple de quelques frames d'un extrait (ou clip snippet) avec la bande son associée. La bande son est une succession de deux types de son. Dans cet exemple, le premier son provient de l'homme au centre qui joue au piano, et le deuxième son provient de l'homme qui chante au centre.

V.

D'abord, nous avons évalué la distance moyenne d entre les positions des yeux des participants dans les deux groupes avec des conditions AV et V et dans trois groupes de classes (voir Fig. A.5): " une source sonore à l'écran", " plusieurs sources sonores à l'écran" et " source sonore hors écran". Pour chaque extrait (ou clip snippet), nous avons choisi 25 images (de la frame 6 à 30 pour éliminer le temps de réaction d'environ 5 frames) après le début du deuxième son. Nous avons utilisé le test ANOVA pour comparer la distance d entre les différents groupes de classes. Dans la Fig. A.7, un test ANOVA indique que le groupe " source sonore hors écran" présente la distance d la plus faible parmi les trois groupes de classes. La différence est significative entre les groupes "une source sonore à l'écran" et " source sonore hors écran" ($F(1, 175) = 7.94$, $p < 10^{-2}$), et également significative entre "plusieurs sources sonores à l'écran" et " source sonore hors écran" ($F(1, 73) = 8.69$, $p < 10^{-3}$). Celle entre les groupes "une source sonore à l'écran" et "plusieurs sources sonores à l'écran" n'est pas significativement différente ($F(1, 184) = 0.12$, $p = 0.73$). Ces résultats sont confirmés par deux autres critères: la divergence de Kullback-Leibler KLD et le coefficient de corrélation linéaire cc.

Ensuite, nous avons analysé les treize classes sonores séparément. Nous n'avons pas directement analysé l'effet sonore à partie de l'information audio, mais à l'aide des positions des yeux des participants; positions qui sont également basées sur l'information visuelle. Afin de réduire l'influence de l'information visuelle, nous avons créé une ligne de référence d_R (moyenne de 5000 réalisations aléatoires) pour

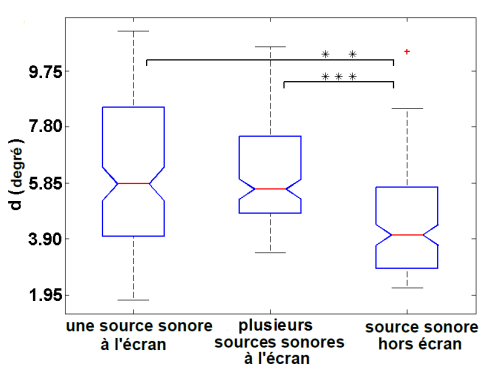
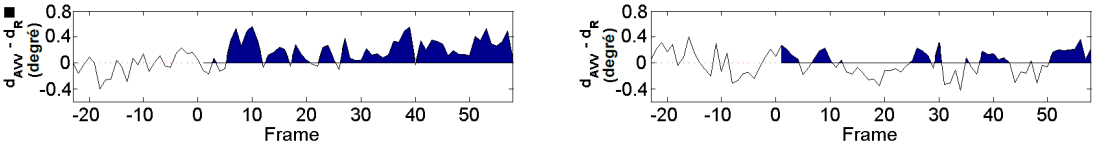


Figure A.7: Critères de distance moyenne d entre les participants avec les conditions AV et V dans trois groupes de classes: “une source sonore à l’écran”, “plusieurs sources sonores à l’écran” et “source sonore hors écran”. Une distance d plus grande représente une différence plus élevée entre les groupes avec les conditions AV et V.

la comparaison statistique en effectuant une “randomisation” [Edgington 2007].



(a) parole (b) impact et explosion

Figure A.8: Différence moyenne $(d_{AVV} - d_R)$ au cours du temps pour la classe “parole” (11 extraits ou clip snippets) et “impact et explosion” (8 extraits ou clip snippets). La frame 1 correspond au début du deuxième son. Les régions foncées représentent une $(d_{AVV} - d_R)$ positive, ce qui indique que la différence entre les groupes AV et V est supérieure à celle entre les deux groupes aléatoires.

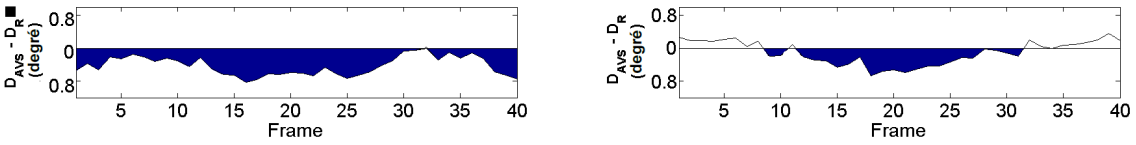
La Fig. A.8 montre la différence au cours du temps entre d_{AVV} et d_R pour deux classes: “parole ” (humain) et “impact et explosion” (non-humain). Si $(d_{AVV} - d_R)$ est positive, la différence entre les groupes AV et V est plus grande que celle entre deux groupes créés aléatoirement. Le comportement au fil du temps est différent pour les deux classes sonores présentées.

Nous avons fait l’étude pour savoir quelles classes ont la différence la plus grande entre d_{AVV} et d_R pour les frames successives. Pour quantifier l’effet sonore, il est préférable de mesurer l’effet du son pour chaque classe de son sur une certaine durée, plutôt que sur chaque image seule. Nous avons étudié sur une période suffisamment longue d’une seconde (25 frames) de la frame 6 à 30 après le début du deuxième son. Nous avons comparé $\overline{d}_{AVV}$ (la moyenne temporelle de d_{AVV} sur ces 25 frames) avec la distribution de $\overline{d}_i$, où $\overline{d}_i$ est la moyenne temporelle d_i des groupes aléatoires pour les 25 frames pour l’essai aléatoire i. Pour estimer la probabilité d’obtenir une

$\overline{d}_i$ supérieur à $\overline{d}_{AVV}$, nous avons calculé $p = n/5000$, où n est le nombre de $\overline{d}_i$ qui sont supérieurs à $\overline{d}_{AVV}$. Les résultats ont montré qu'à partir de la frame de 6 à 30 après le début du deuxième son, les classes de sons « parole », «chanteur», «bruit de l'humain», et «chanteurs», ont des valeurs élevées $\overline{d}_{AVV}$ (donc des valeurs p faibles), indiquant que la voix humaine affecte significativement le regard visuel ($p < 0.05$).

- *Les participants déplacent leurs yeux vers la source sonore lorsqu'il s'agit de voix humaine.*

  Nous voulons vérifier l'hypothèse que les participants avec la condition AV déplacent leurs yeux vers la source sonore dès le début du deuxième son. Nous avons seulement analysé la classe de son "une source de son à l'écran". Nous avons d'abord manuellement localisé les coordonnées approximatives du centre de la source du son. Ensuite, nous avons calculé la distance Euclidienne entre la position des yeux de chaque participant avec la condition AV et la source sonore. La moyenne de ces distances Euclidiennes donne la valeur D_{AVS}, qui est affectée à la fois par l'image et des informations sonores. De la même façon, afin de réduire l'influence de l'information visuelle, nous avons créé une *ligne de référence* en effectuant une "randomisation". Nous avons considéré la distance Euclidienne moyenne entre la position des yeux des participants du G1 (composé de 18 participants, qui sont sélectionnés aléatoirement dans l'ensemble des participants des groupes avec les conditions AV et V) et la source de son ($D_i, i = 1, 2...5000$). Nous avons utilisé la moyenne des 5000 valeurs de distance pour calculer la ligne de base (D_R), qui n'a été affectée que par les informations de l'image. Ensuite, pour chaque frame, nous avons calculé $D_{AVS} - D_R$ pour toutes les classes qui ont une source de son. Cette différence met en lumière l'influence de l'information sonore.



(a) parole (b) impact et explosion

Figure A.9: Différence moyenne ($D_{AVS} - D_R$) au cours du temps pour les classes "parole" et "impact et explosion". Les régions foncées représentent une ($D_{AVS} - D_R$) négative. Une telle valeur indique que le groupe à la condition AV est plus proche de la source de son que le groupe aléatoire.

La Fig. A.9 montre la différence entre D_{AVS} et D_R dans le temps et pour les classes "parole" et "impact et explosion". Lorsque les valeurs sont négatives, le groupe avec la condition AV est plus proche de la source sonore que le groupe aléatoire. En

conclusion, les différentes classes sonores se comportent différemment.

Pour savoir quelles classes ont la plus grande différence entre D_{AVS} et D_R et pour quantifier l'effet sonore, nous analysons sur une même durée d'une seconde (25 images) comme dans l'analyse précédente, de la frame 6 à 30 après le début du deuxième son. Nous avons comparé $\overline{D}_{AVS}$(la moyenne de D_{AVS} à travers de 25 frames) avec la distribution de $\overline{D}_i(i = 1, 2, ...5000)$, où $\overline{D}_i$ est la moyenne de D_i entre G1 et la source de son dans les 25 frames pour l'essai aléatoire i. Pour estimer la probabilité d'obtenir une $\overline{D}_i$ inférieure à $\overline{D}_{AVS}$, nous calculons $p = n/5000$, où n est le nombre $\overline{D}_i$ qui est plus petit que $\overline{D}_{AVS}$ ($p < 0.05$) entre la frame 6 à 30 après le début du deuxième son. Cet effet apparaît seulement dans les classes sonores suivantes: parole, chanteur(s) et bruit de l'humain. Cette effet indique que les participants ont tendance à déplacer leurs yeux vers la source du son uniquement quand ils entendent de la voix humaine.

- *Les participants avec la condition AV ont une durée moyenne de fixation plus courte que ceux avec la condition V.*

  Nous avons aussi étudié l'effet du son pour toute la base de données sur les distributions de durée de fixation. Il est classique d'étudier ces paramètres [Tatler 2011]. Pour chaque participant, nous avons calculé la durée moyenne de fixation pour chaque clip. Une méthode traditionnelle - le test t apparié est employé. Pour chaque clip, la condition AV a une durée moyenne de fixation plus courte (6.17 frames, 247 ms) que la condition V (6.82 frames, 273 ms). La différence correspondante est significative ($t(9) = 2.479, p = 0.035$). Par participant, la condition AV a toujours une durée moyenne de fixation plus courte (6.19 frames, 248 ms) que la condition V (6.75 frames, 270 ms). La différence correspondante est aussi significative ($t(35) = 2.697, p = 0.011$). Cela signifie que les participants avec la condition AV ont tendance à déplacer leurs yeux plus souvent par rapport aux participants avec la condition V. De plus ce résultat est confirmé par une méthode plus récente - modèle à effets mixtes [Baayen 2008].

- *Le son réduit la précision de la prédiction des modèles de saillance visuelle.*

  Dans l'analyse de l'expérience précédente, la comparaison des positions des yeux par le modèle de saillance visuelle proposé par Marat *et al.* a montré que la précision de la prédiction diminue lorsque les données vidéo sont associées avec des bandes son originales. Est-ce que les performances du modèle de saillance visuelle sont adaptées à des données vidéo avec bande sonore? Par le savoir, nous avons répété la comparaison de la base de données dans l'expérience II en synchronisant tous

les extraits (ou clip snippets) au début du deuxième son de chaque extrait (ou clip snippet).

Nous effectuous la comparaison des positions des expérimentales des yeux d'une part avec les voies statique et dynamique dans le modèle de saillance proposé par Marat *et al.*, et d'autre part avec le modèle de saillance proposé par Itti *et al.* en 1998, associé à la voie supplémentaire de mouvement (en 2003). Les résultats des NSS et TC indiquent tous les deux que la précision de la prédiction a tendance à diminuer dans le groupe avec la condition AV plutôt que dans le groupe avec la condition V, au cours des frames 6-56 après le début du deuxième son. Cette diminution de la précision de prévision est significative à un niveau 5% du modèle de saillance visuelle proposé par Marat's *et al.* et de la voie de mouvement dans le modèle de saillance visuelle proposé par Itti's *et al.*.

Modèle préliminaire de saillance audiovisuelle

Dans l'analyse précédente, nous avons montré que les "cartes de saillance de la source de son" ont augmenté la précision de la prédiction du groupe des participants avec la condition AV. Parce qu'il n'y a aucune information spatiale portée dans le signal sonore dans nos données vidéo, il est difficile de localiser la source du son dans la scène visuelle automatiquement. Nous proposons une autre approche. Nous supposons que le mouvement visuel est en conjonction et est synchronisé avec la bande son associée. Afin d'améliorer la précision de la prédiction d'un modèle de saillance visuelle avec la condition AV, un modèle préliminaire de saillance audiovisuel est proposé en ajoutant des informations sonores. Plus précisément, deux stratégies de fusion pour les classes sonores "parole" et "instrument de musique" sont proposés dans ce modèle:

- *Détecter le visage parlant pour la classe de parole.*

  Nous nous sommes concentrés sur des extraits de clip avec des bandes son de parole et avec plusieurs visages à l'écran, mais un seul visage correspond à la seule personne qui parle. Basé sur l'*hypothèse* que le visage parlant est en mouvement, nous avons proposé de fusionner la carte de visages (marqués manuellement) et la carte de saillance dynamique (dans le modèle de saillance visuelle de Marat) pour localiser automatiquement le visage parlant en tant que région saillante.

  L'organigramme du modèle de saillance audiovisuelle est représentée dans la Fig. A.10. L'algorithme de la stratégie de fusion I est brièvement décrite ci-dessous:

  1. Si c'est la première frame de la parole (la borne de l'événement audio), nous

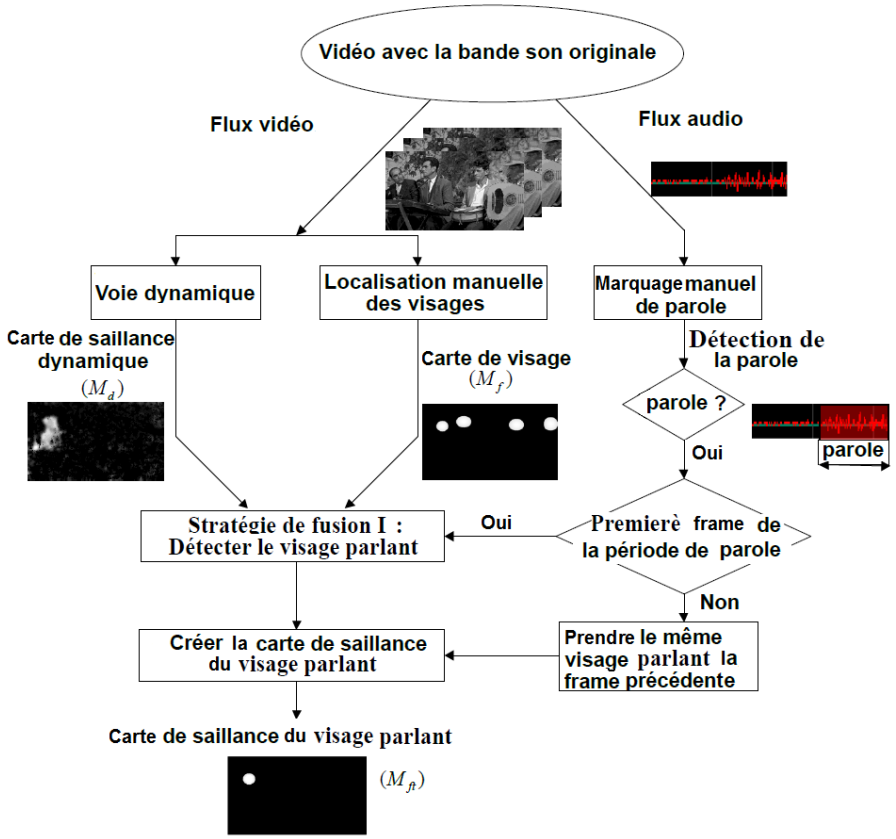


Figure A.10: Organigramme du modèle de saillance proposé pour la classe parole.

calculons la moyenne spatiale de la valeur de saillance dynamique (ou de mouvement) de chaque région du visage marqué manuellement.

2. Nous avons considéré la région du visage, qui a la plus haute valeur de mouvement, comme étant la région du visage parlant. Une fois le visage parlant sélectionné au début de la période de la parole, cette sélection sera conservée sur la durée entière de la parole.

4. Enfin, nous avons ajouté une fonction Gaussienne bi-dimensionnelle autour du visage parlant choisi pour créer une *carte de saillance du visage parlant* M_m.

Cette stratégie de fusion I de sélection du visage parlant parmi tous les visages se comporte bien sur la base de cinq extraits de clips choisis. Les visages parlants dans ces cinq extraits de clips sont correctement sélectionnés. En outre, la carte de saillance du visage parlant donne de meilleurs résultats par rapport à la carte de visage seule et la carte de saillance dynamique seule évaluées par NSS et TC [Torralba 2006]. En outre, la carte de saillance de visage parlant augmente la précision de la prédiction pour la condition AV après détection de la parole.

- *Détecter le visage du joueur pour la classe instrument de musique.*

  Nous nous sommes concentrés sur des extraits (ou clip snippets) avec des bandes son musicales et avec plusieurs visages à l'écran, mais un seul visage de joueur d'un instrument. L'hypothèse supplémentaire est que la distance Euclidienne entre un instrument musical et le visage du joueur est la plus courte, par rapport à la distance à d'autres visages dans l'image. Ensuite, une stratégie de fusion de la carte de visage et de la carte de saillance dynamique a été proposée pour détecter automatiquement le visage du joueur.

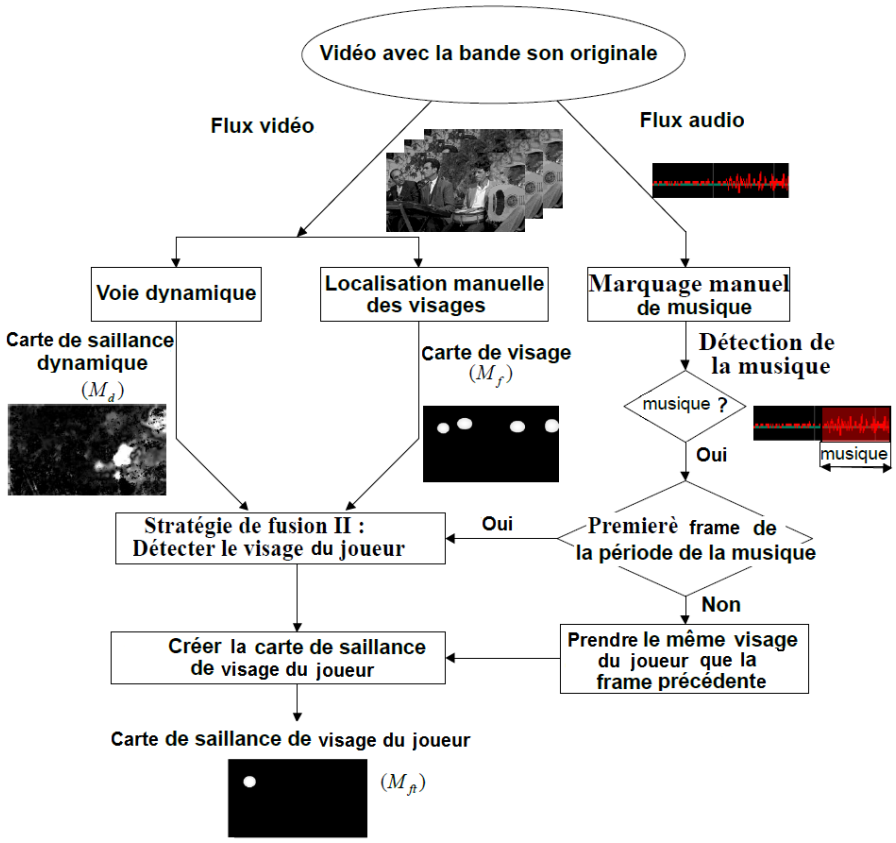


Figure A.11: Organigramme du modèle de saillance audiovisuelle proposé pour la classe instrument de musique.

L'organigramme du modèle de saillance audiovisuelle proposé pour la classe instrument de musique est présenté dans la Fig. A.11. L'algorithme de stratégie de fusion II est assez similaire à la stratégie de fusion I. Nous considérons la région avec la plus haute valeur de mouvement comme source sonore - instrument de musique. Ensuite, le visage, qui est le plus proche de la source sonore, est le visage du joueur. Après, on a crée la carte de saillance de visage du joueur M_{fp}.

Cette stratégie de fusion II, qui a pour objectif de sélectionner le visage de joueur

parmi tous les visages, est testé sur les trois extraits (ou clip snippets) dans la classe instrument de musique. Le visage du joueur dans ces trois extraits (ou clip snippets) est sélectionné correctement. Par rapport à la condition V, la carte de saillance de visage du joueur augmente la précision de la prédiction pour les participants avec la condition AV après détection de la musique. Aussi, la carte de saillance de visage de joueur a de meilleure performance que la carte de visage seule et la carte de saillance dynamique seule qui sont toutes évaluées par les critères NSS et TC.

A.4 Conclusions générales et perspectives

Dans cette thèse, nous nous sommes intéressés à une meilleure compréhension de l'influence du son sur le mouvement des yeux dans des vidéos. Nous débutons notre étude par les expériences de suivi oculaire entre deux groups de participants avec la condition audio-visuelle AV (avec bande son originale) et avec la condition visuelle V. Ces expériences audiovisuelles sont conçues pour étudier l'influence de l'interaction audiovisuelle sur le mouvement des yeux. Pour observer cette influence, les positions expérimentales des yeux des participants ont été enregistrées et analysées. Avec notre base de vidéos, il apparaît que seules les classes de son "parole", "chanteur(s)" et "bruit humain" influencent de manière significative le mouvement des yeux. Enfin, à partir de ces observations, un modèle préliminaire de saillance audiovisuelle est proposé pour prédire des régions saillantes dans les vidéos avec des bandes son: parole et instrument de musique.

Dans les travaux futurs, nous avons différents aspects à améliorer dans notre modèle préliminaire de saillance audiovisuelle. Tout d'abord, nous pouvons remplacer le marquage manuel des visages par un détecteur de visage, qui réalise une détection automatique de visage. Deuxièmement, nous pouvons remplacer l'étiquetage manuel des bandes-son par un classifieur sonore. Les algorithmes de détection de visage et de classification de son apporteront des erreurs de classification. Dans ce cas, afin de réduire l'influence de l'erreur de classification, nous pouvons introduire une stratégie de fusion pondérée en utilisant des valeurs de confiance pour la détection et la classification.

Bibliography

[Ahveninen 2012] J. Ahveninen, I. P. Jääskeläinen, J. W. Belliveau, M. Hämäläinen, F. H. Lin and T. Raij. *Dissociable influences of auditory object vs. spatial attention on visual system oscillatory activity.* Plos One, vol. 7, no. 6, pages 1–10, 2012.

[Alho 2012] K. Alho, J. Salonen, T. Rinne, S. V. Medvedev, K. Hugdahl and H. Hämäläinen. *Attention-related modulation of auditory-cortex responses to speech sounds during dichotic listening.* Brain Research, vol. 1442, no. 9, pages 47–54, March 2012.

[Alpert 2008] G. F. Alpert, G. Hein, N. Tsai, M. J. Naumer and R. T. Knight. *Temporal characteristics of audiovisual information processing.* The Journal of Neuroscience, vol. 28, no. 20, pages 5344–5349, May 2008.

[Astroweb] `http://www.astro.virginia.edu/class/oconnell/astr1230/human-eye.html`.

[Atrey 2010] P. K. Atrey, M. A. Hossain, A. E. Saddik and M. S. Kankanhalli. *Multimodal fusion for multimedia analysis: a survey.* Multimedia Systems, vol. 16, pages 345–379, 2010.

[Awh 2006] E. Awh, K. M. Armstrong and T. Moore. *Visual and oculomotor selection: Links, causes and implications for spatial attention.* Trends in Cognive Sciences, vol. 10, no. 3, pages 124–30, 2006.

[Baayen 2008] R. H. Baayen, D. J. Davidson and D. M. Bates. *Mixed-effects modeling with crossed random effects for subjects and items.* Journal of Memory and Language, vol. 59, pages 390–412, 2008.

[Baluch 2011] F. Baluch and L. Itti. *Mechanisms of top-down attention.* Trends in Neurosciences, vol. 34, pages 210–224, March 2011.

[Bates a] `https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html`.

[Bates b] D. M. Bates and D. Sarkar. *lme4: Linear mixed-effects models using S4 classes.* R package version 0.99875-6.

[Begbie 1996] G. H. Begbie. Seeing and the eye; an introduction to vision. National History Press, 1996.

[Borji 2011] A. Borji, D. N. Sihite and L. Itti. *Computational modeling of top-down cisual attention in interactive environments.* British Machine Vision Conference (BMVC 2011), vol. 85, pages 1–12, Sep 2011.

[Borji 2012] A. Borji, D. N. Sihite and L. Itti. *Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study.* IEEE Transactions on Image Process., vol. 22, no. 1, pages 55–69, January 2012.

[Bredin 2007] H. Bredin and G. Chollet. *Audio-visual speech synchrony measure for talking-face identity verification.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, pages 233–236, 2007.

[Bregman 1990] A. S. Bregman. Auditory scene analysis: the perceptual organization of sound. Cambridge, Massachusetts: The MIT Press, 1990.

[Buhr 2009] H. J. Buhr. *The effect of sound for visual scene exploration: Computational and psychophysical approaches.* Master thesis, GIPSA-lab, 2009.

[Burton 2006] H. Burton and D. G. McLaren. *Visual cortex activation in late-onset, Braille naive blind individuals: An fMRI study during semantic and phonological tasks with heard words.* Neuroscience Letters, vol. 392, pages 38–42, 2006.

[Cao 2009] Y. Cao, X. Wei, L. Zhao and R. Di Federico. *Robust multi-clue face tracking system.* Global Congress on Intelligent Systems (GCIS), vol. 4, pages 513–517, May 2009.

[Carlson 2009] N. R. Carlson and C. D. Heth. Psychology the science of behaviour. Pearson Education Canada, 4th edition, 2009.

[Carmi 2006] R. Carmi and L. Itti. *Visual causes versus correlates of attentional selection in dynamic scenes.* Vision Research, vol. 46, pages 4333–4345, Oct. 2006.

[Carriere 2008] B. N. Carriere, D. W. Royal and M. T. Wallace. *Spatial heterogeneity of cortical receptive fields and its impact on multisensory interactions.* Journal of Neurophysiology, vol. 99, no. 5, pages 2357–2368, May 2008.

[Cohen 1994] M. M. Cohen and D. W. Massaro. *Development and experimentation with synthetic visible speech.* Behavior Research Methods, Instruments, and Computers, vol. 26, pages 260–265, 1994.

[Connor 2004] C. E. Connor, H. E. Egeth and S. Yantis. *Visual attention: bottom-up versus top-down.* Current Biology, vol. 14, pages 850–852, Oct. 2004.

[Corder 2009] G. W. Corder and D. I. Foreman. Nonparametric statistics for non-statisticians. Wiley, 2009.

[Corneil 2002] B. D. Corneil, M. van Wanrooij, D. P. Munoz and A. J. Van Opstal. *Auditory-visual interactions subserving goal-directed saccades in a complex scene.* Journal of Neurophysiology, vol. 88, no. 1, pages 438–454, 2002.

[Coutrot 2012a] A. Coutrot, N. Guyader, G. Ionescu and A. Caplier. *Influence of soundtrack on eye movements during video exploration.* Journal of Eye Movement Research, vol. 5, no. 4, pages 1–10, 2012.

[Coutrot 2012b] A. Coutrot, G. Ionescu, N. Guyader and A. Caplier. *Exploration libre de vidéos : influence du son sur les mouvements oculaires consécutifs à un événement sonore saillant.* 15ème journées COmpression et Représentation des Signaux Audiovisuels (CORESA), May 2012.

[Coutrot 2013] A. Coutrot, N. Guyader, G. Ionescu and A. Caplier. *Video viewing: do auditory salient events capture visual attention?* Annals of Telecommunication, pages 1–9, 2013.

[Dalton 2007] P. Dalton and C. Spence. *Attentional capture in serial audiovisual search tasks.* Perception & Psychophysics, vol. 69, pages 422–438, 2007.

[Demidenko 2004] E. Demidenko. Mixed models: Theory and applications. Wiley, 1 edition, August 2004.

[Desimone 1995] R. Desimone and J. Duncan. *Neural mechanisms of selective visual-attention.* Annual Review of Neuroscience, vol. 18, pages 193–222, 1995.

[Dorr 2010] M. Dorr, T. Martinetz, K. R. Gegenfurtner and E. Barth. *Variability of eye movements when viewing dynamic natural scenes.* Journal of Vision, vol. 10, no. 28, pages 1–17, August 2010.

[Driver 2008] J. Driver and T. Noesselt. *Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments.* The Journal of Neuroscience, vol. 57, pages 11–23, 2008.

[Duangudom 2007] V. Duangudom and D. V. Anderson. *Using auditory saliency to understand complex auditory scenes.* 15th European Signal Processing Conference (EUSIPCO), pages 1206–1210, September 2007.

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

[EprimeSite] `http://www.pstnet.com/eprime.cfm`.

[Erkelens 2011] C. J. Erkelens. *A dual visual-local feedback model of the vergence eye movement system.* Journal of Vision, vol. 11, no. 10, 2011.

[Escoffier 2010] N. Escoffier, D. Y. J. Sheng and A. Schirmer. *Unattended musical beats enhance visual processing.* Acta Psychologica, vol. 135, no. 1, pages 12–16, 2010.

[Evangelopoulos 2008a] G. Evangelopoulos, K. Rapantzikos, P. Maragos, T. Avrithis and A. Potamianos. Audiovisual attention modeling and salient event detection. Springer, 2008.

[Evangelopoulos 2008b] G. Evangelopoulos, K. Rapantzikos, A. Potamianos, P. A. Maragos, A. Zlatintsi and Y. S. Avrithis. *Movie summarization based on audiovisual saliency detection.* IEEE International Conference on Image Processing, vol. 33, pages 2528–2531, Oct. 2008.

[EyelinkWeb] `http://sr-research.com/EL_II.html`.

[Feki 2011] I. Feki, A. B. Ammar and A. M. Alimi. *Environmental sound extraction and incremental learning approach for real time concepts identification.* IEEE Symposium on Computational Intelligence for Multimedia, Signal and Vision Processing (CIMSIVP), pages 33 – 38, April 2011.

[Feng 2011] H. Feng, C. Jiang and X. Yang. *An audio classification and speech recognition system for video content analysis.* International Conference on Multimedia Technology (ICMT), pages 5272 – 5276, July 2011.

[Findlay 2009] J. M. Findlay. *Saccadic eye movement programming: sensory and attentional factors.* Psychological Research, vol. 73, no. 2, pages 127–135, March 2009.

[Frens 1995] M. A. Frens, A. J. Vanopstal and R. F. Van der Willigen. *Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements.* Perception & Psychophysics, vol. 6, pages 802–816, 1995.

[Gelman 2005] A. Gelman. *Analysis of variance? Why it is more important than ever.* The Annals of Statistics, vol. 33, pages 1–53, 2005.

[Ghazanfar 2006] A. A. Ghazanfar and C. E. Schroeder. *Is neocortex essentially multisensory?* Trends Cognition Science, vol. 10, pages 278–285, 2006.

[Gibbons 2003] J. D. Gibbons. Nonparametric statistical inference. CRC Press, 4 edition, 2003.

[Glotin 2001] H. Glotin, D. Vergyr, C. Neti, G. Potamianos and J. Luettin. *Weighting schemes for audio-visual fusion in speech recognition.* IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pages 173–176, 2001.

[Gordon 2011] M. S. Gordon and M. Hibberts. *Audiovisual speech from emotionally expressive and lateralized faces.* The Quarterly Journal of Experimental Psychology, vol. 64, no. 4, pages 730–750, 2011.

[Groh 2002] J. M. Groh and U. W. Reiss. *Visual and auditory integration.* In Encyclopedia of the Human Brain, volume 4, pages 739–752. Elsevier Science, 2002.

[Henderson 1999] J. M. Henderson and A. Hollingworth. *High-level scene perception.* Annu. Rev. Psychol., vol. 50, pages 243–271, 1999.

[Heron 2004] J. Heron, D. Whitaker and P. V. McGraw. *Sensory uncertainty governs the extent of audio-visual interaction.* Vision Research, vol. 44, pages 2875–2884, 2004.

[Hidaka 2010] S. Hidaka, W. Teramoto, J. Gyoba and Y. Suzuki. *Sound can prolong the visible persistence of moving visual objects.* Vision Research, vol. 50, pages 2093–2099, 2010.

[Hoffman 1995] J. E. Hoffman and B. Subramaniam. *The role of visual attention in saccadic eye movements.* Perception & Psychophysics, vol. 57, no. 6, pages 787–795, 1995.

[Hollander 1999] M. Hollander and D. A. Wolfe. Nonparametric statistical methods. John Wiley & Sons, 1999.

[Hooge 2012] I. Hooge, E. Over and C. Erkelens. *Ineffective visual search: Search performance deteriorates near borders due to inappropriate fixation durations and saccade amplitudes.* Journal of Vision, vol. 12, no. 9, page 263, August 2012.

[Howell 1987] D. C. Howell. Statistical methods for psychology. PWS Publishers, 1987.

[IfdWeb] `http://www.ifd.mavt.ethz.ch/research/group_lk/projects/cochlear_mechanics`.

[InriaWeb] `http://perception.inrialpes.fr/~Horaud/POP/TutorialsFEB06/JonBarker.pdf`.

[Ionescu 2009] G. Ionescu, N. Guyader and A. Guérin-Dugué. *SoftEye software.* IDDN.FR.001.200017.000.S.P.2010.003.31235, 2009.

[Itti 1998] L. Itti, C. Koch and E. Niebur. *A model of saliency-based visual attention for rapid scene analysis.* Trans. on pattern analysis and machine intelligence, vol. 20, no. 11, pages 1254–1259, Sept. 1998.

[Itti 2003] L. Itti, N. Dhavale and F. Pighin. *Realistic avatar eye and head animation using a neurobiological model of visual attention.* Proc. SPIE 48th Annual International Symposium on Optical Science and Technology, vol. 5200, pages 64–78, August 2003.

[Itti 2005] L. Itti. *Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes.* Visual Cognition, vol. 12, no. 6, pages 1093–1123, 2005.

[Iyengar 2003] G. Iyengar, H. J. Nock and C. Neti. *Audio-visual synchrony for detection of monologue in video archives.* IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pages 329–332, July 2003.

[Jaffré 2006] G. Jaffré and J. Pinquier. *Audio/video fusion: a preprocessing step for multimodal person identification.* International Workshop on Multimodal User Authentification, May 2006.

[Jeffers 1971] J. Jeffers and M. Barley. Speechreading (lipreading). Springfield, 1971.

[Kalinli 2007] O. Kalinli and S. Narayanan. *A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech.* Proceedings of InterSpeech, pages 1941–1944, August 2007.

[Kayser 2005] C. Kayser, C. I. Petkov, M. Lippert and N. K. Logothetis. *Mechanisms for allocating auditory attention: an auditory saliency map.* Current Biology, vol. 15, no. 8, pages 1943–1947, November 2005.

[Keller 2009] Y. Keller, R. R. Coifman, S. Lafon and S. W. Zucker. *Audio-visual group recognition using diffusion maps.* IEEE Trans. Signal Process, vol. 58, no. 1, pages 403–413, Jan. 2009.

[Kühn 2012] B. Kühn, B. Schauerte, R. Stiefelhagen and K. Kroschel. *A modular audio-visual scene analysis and attention system for humanoid robots.* The 43rd Intl. Symp. on Robotics (ISR), pages 29–31, August 2012.

[Kidron 2005] E. Kidron, Y. Y. Schechner and M. Elad. *Pixels that sound.* Proc. IEEE Computer Vision & Pattern Recognition, vol. 1, pages 88–96, 2005.

[Kim 2004] J. Kim, C. Davis and P. Krins. *Amodal processing of visual speech as revealed by priming.* Cognition, vol. 93, pages B39–B47, 2004.

[King 2009] A. J. King. *Visual influences on auditory spatial learning.* Philosophical Trans. of the Royal Society, vol. 364, pages 331–339, 2009.

[Koch 1985] C. Koch and S. Ullman. *Shifts in selective visual attention: towards the underlying neural circuitry.* Human Neurobiology, vol. 4, pages 219–227, 1985.

[Körding 2007] K. P. Körding, U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum and L. Shams. *Causal inference in multisensory perception.* Plos One, vol. 26, no. 2, pages 1–10, 2007.

[Kuling 2012] I. A. Kuling, R. L. J. van Eijk, J. F. Juola and A. Kohlrausch. *Effects of stimulus duration on audio-visual synchrony perception.* Experimental Brain Research, vol. 221, pages 403–412, 2012.

[Langton 2008] S. R. H. Langton, A. S. Law, A. M. Burton and S. R. Schweinberger. *Attention capture by faces.* Cognition, vol. 107, pages 330–342, 2008.

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

[Lewald 2001] J. Lewald, W. H. Ehrenstein and R. Guski. *Spatio-temporal constraints for auditory–visual integration.* Behavioral Brain Research, vol. 121, no. 1, pages 69–79, June 2001.

[Liang 2008] M. Liang, R. M. van Leeuwen and M. J. Proulx. *Propagation of brain activity during audiovisual integration.* The Journal of Neuroscience, vol. 28, no. 36, pages 8861–8862, September 2008.

[Lilliefors 1969] H. W. Lilliefors. *On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown.* Journal of the American Statistical Association, vol. 64, pages 387–389, 1969.

[Litvin 2010] Y. Litvin, I. Cohen and D. Chazan. *Monaural speech/music source separation using discrete energy separation algorithm.* Signal Processing, no. 90, pages 3147–3163, 2010.

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

[Marat 2009] S. Marat, T. Ho. Phuoc, L. Granjon, N. Guyader, D. Pellerin and A. Guérin-Dugué. *Modelling Spatio-temporal saliency to predict gaze direction for short video.* International Journal of Computer Vision, vol. 82, pages 231–243, May 2009.

[Marat 2010] S. Marat. *Modèles de saillance visuelle par fusion d'informations sur la luminance, le mouvement et les visages pour la prédiction de mouvements oculaires lors de l'exploration de vidéos.* PhD thesis, Université de Grenoble, 2010.

[Marat 2012] S. Marat, A. Rahman, D. Pellerin, N. Guyader and D. Houzet. *Improving visual saliency by adding 'face feature map' and 'center bias'.* Cognitive Computation, vol. 5, no. 1, pages 63–75, March 2012.

[Marchant 2012] J. L. Marchant, C. C. Ruff and J. Driver. *Audiovisual synchrony enhances BOLD responses in a brain network including multisensory STS while also enhancing target-detection performance for both modalities.* Human Brain Mapping, vol. 33, no. 5, pages 1212–1224, May 2012.

[McGurk 1976] H. McGurk and J. MacDonald. *Hearing lips and seeing voices.* Nature, vol. 264, pages 746–748, December 1976.

[Mercier 2012] M. R. Mercier, J. J. Foxe, I. C. Fiebelkorn, J. S. Butler, T. H. Schwartz and S. Molholm. *Auditory modulation of oscillatory activity in extra-striate visual cortex and its contribution to audio-visual multisensory integration: A human intracranial EEG study.* Seeing and Perceiving, vol. 25, no. 1, pages 198–198(1), 2012.

[Molholm 2002] S. Molholm, W. Ritter, M. M. Murray, D. C. Javitt, C. E. Schroeder and J. J. Foxe. *Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study.* Cognitive Brain Research, vol. 14, no. 1, pages 115–128, June 2002.

[Niessen 2008] M. E. Niessen, L. van Maanen and T. C. Andringa. *Disambiguating sounds through context.* IEEE International Conference on Semantic Computing, pages 88 – 95, Aug. 2008.

[Nordfang 2010] M. Nordfang and C. Bundesen. *Is initial visual selection completely stimulus-driven?* Acta Psychologica, vol. 135, pages 106–108, 2010.

[Ntoumanis 2005] N. Ntoumanis. A step-by-step guide to spss for sport and exercise studies. Taylor and Francis e-Library, 2005.

[O'Brien 2010] J. H. O'Brien, M. D. Luca and M. Ernst. *Audiovisual integration: the duration of uncertain times.* Journal of Vision, vol. 10, no. 7, page 1408, August 2010.

[O'Donovan 2007] A. O'Donovan and R. Duraiswami. *Microphone arrays as generalized cameras for integrated audio visual processing.* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–8, June 2007.

[Osterberg 1935] G. Osterberg. *Topography of the layer of rods and cones in the human retina.* Acta Ophthalmol, pages 1–103, 1935.

[Oyster 1999] C. W. Oyster. The human eye structure and function. Sinauer Associates Incorporated, 1999.

[Parkhurst 2002] D. Parkhurst, K. Law and E. Niebur. *Modeling the role of salience in the allocation of overt visual attention.* Vision research, vol. 42, pages 107–123, 2002.

[Perreira Da Silva 2010] M. Perreira Da Silva. *Modèle computationnel d'attention pour la vision adaptative.* PhD thesis, Université de la Rochelle, 2010.

[Perrott 1990] D. R. Perrott, K. Saberi, K. Brown and T. Z. Strybel. *Auditory psychomotor coordination and visual search performance.* Perception & Psychophysics, vol. 48, pages 214–226, 1990.

[Peters 2005] R. J. Peters, A. Iyer, L. Itti and C. Koch. *Components of bottom up gaze allocation in natural images.* Vision Research, vol. 45, pages 2397–2416, August 2005.

[Pickles 2012] J. O. Pickles. An introduction to the physiology of hearing. Emerald Group, 4 edition, August 2012.

[PraatSite] `http://www.fon.hum.uva.nl/praat/`.

[Quigley 2008] C. Quigley, S. Onat, S. Harding, M. Cooke and P. König. *Audio-visual integration during overt visual attention.* J. Eye Movement Research, vol. 1, pages 1–17, 2008.

[R-project Web] `http://www.r-project.org/`.

[Rahman 2013] A. Rahman. *Face perception in videos: Contributions to a visual saliency model and its implementation on GPUs.* PhD thesis, Université de Grenoble, 2013.

[Ramenahalli 2013] R. Ramenahalli, D. Mendat, S. Dura-Bernal, E. Culurciello, E. Niebur and A. G. Andreou. *Audio-visual saliency map: overview, basic models and hardware implementation.* Proceedings of the Information Sciences and Systems conference (CISS), 2013.

[Ro 2001] T. Ro, C. Russell and N. Lavie. *Changing faces: a detection advantage in the flicker paradigm.* Psychological Science, vol. 12, no. 1, pages 94–99, 2001.

[Romanski 2012] L. M. Romanski and J. Hwang. *Timing of audiovisual inputs to the prefrontal cortex and multisensory integration.* Neuroscience, vol. 214, no. 12, pages 36–48, July 2012.

[Rosenblum 1997] L. D. Rosenblum, M. A. Schmuckler and J. Q. Johnson. *The McGurk effect in infants.* Perception & Psychophysics, vol. 59, no. 3, pages 347–357, 1997.

[Rossion 2000] B. Rossion, I. Gauthier, M. J. Tarr, P. Despland, R. Bruyer, S. Linotte and M. Crommelinck. *The N170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: an electrophysiological account of face-specific processes in the human brain.* Neuroreport, vol. 11, no. 1, pages 69–74, Jan. 2000.

[Ruesch 2008] J. Ruesch, M. Lopes, A. Bernardino, J. Hörnstein, J. Santos-Victor and R. Pfeifer. *Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub.* IEEE International Conference on Robotics and Automation, pages 19–23, May 2008.

[Schwartz 2004] J-L. Schwartz, F. Berthommier and C. Savariaux. *Seeing to hear better: evidence for early audio-visual interactions in speech identification.* Cognition, vol. 93, pages B69–B78, 2004.

[Shamma 2001] S. Shamma. *On the role of space and time in auditory processing.* Trends in cognitive sciences, vol. 5, no. 8, pages 340–348, August 2001.

[Shivappa 2008] S. T. Shivappa, B. D. Rao and M. M. Trivedi. *Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition.* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2241–2244, 2008.

[Shivappa 2010] S. T. Shivappa, M. M. Trivedi and B. D. Rao. *Audiovisual information fusion in human - computer interfaces and intelligent environments: a survey.* Proceedings of the IEEE, vol. 98, no. 10, pages 1692–1715, Oct. 2010.

[Siagian 2007] C. Siagian and L. Itti. *Biologically-inspired robotics vision monte-carlo localization in the outdoor environment.* IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 1723–1730, October 2007.

[Sinnett 2008] S. Sinnett, S. S. Faraco and C. Spence. *The co-occurrence of multisensory competition and facilitation.* Acta Psychologica, vol. 128, no. 1, pages 153–161, 2008.

[Song 2011a] G. Song, D. Pellerin and L. Granjon. *Influence of sound on visual gaze when looking at videos.* 16th European Conference on Eye Movements (ECEM), August 2011.

[Song 2011b] G. Song, D. Pellerin and L. Granjon. *Sound effect on visual gaze when looking at videos.* 19th European Signal Processing Conference (EUSIPCO), pages 2034–2038, September 2011.

[Song 2012] G. Song, D. Pellerin and L. Granjon. *How different kinds of sound in videos can influence gaze.* International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), pages 1–4, 2012.

[Spence 1997] C. Spence and J.Driver. *Audiovisual links in exogenous covert spatial orienting.* Percept & Psychophys, vol. 59, pages 1–12, 1997.

[StanfordSite] `http://scien.stanford.edu/pages/labsite/2006/psych221/projects/06/cukur/intro.htmll`.

[StanfordWeb] `http://www.stanford.edu/class/me220/data/lectures/lect01/auditory.html`.

[Stein 1993] B. E. Stein and M. A. Meredith. The merging of the senses. The MIT Press, January 1993.

[Stoper 1973] A. E. Stoper. *Apparent motion of stimuli presented stroboscopically during pursuit movement of the eye.* Perception & Psychophysics, vol. 13, no. 2, pages 201–211, 1973.

[Summerfield 1987] Q. Summerfield. Some preliminaries to a comprehensive account of audio-visual speech perception. Hillsdale, 1987.

[Talantzis 2006] F. Talantzis, F. Pnevmatikakis and L. C. Polymenakos. *Real time audio-visual person tracking.* IEEE 8th Workshop on Multimedia Signal Processing, no. 243-247, 2006.

[Tatler 2005] B. W. Tatler, R. J. Baddeley and I. D. Gilchrist. *Visual correlates of fixation selection: effects of scale and time.* Vision Research, vol. 45, pages 643–659, 2005.

[Tatler 2011] B. W. Tatler, M. M. Hayhoe, M. F. Land and D. H. Ballard. *Eye guidance in natural vision: Reinterpreting salience.* Journal of Vision, vol. 11, no. 5, pages 1–23, 2011.

[Tkac 2011] J. Tkac and J. Macalak. *Sound analysis and classification of signals.* International Symposium MECHATRONIKA, pages 55 – 58, June 2011.

[Torralba 2006] A. Torralba, A. Oliva, M. S. Castelhano and J. M. Henderson. *Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search.* Psychol Rev 113, vol. 4, pages 766–786, 2006.

[Treisman 1980] A. M. Treisman. *A feature-integration theory of attention.* Cognitive Psychology, vol. 12, pages 97–136, 1980.

[Treisman 1982] A. Treisman and H. Schmidt. *lllusory conjunctions in the perception of objects.* Cognitive Psychology, vol. 14, pages 107–141, 1982.

[Treisman 1985] A. Treisman. *Preattentive processing in vision.* Computer Vision, Graphics, and Image Processing, vol. 31, pages 156–177, 1985.

[Tseng 2009] P. H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz and L. Itti. *Quantifying center bias of observers in free viewing of dynamic natural scenes.* Journal of Vision, vol. 9, no. 7, pages 1–16, 2009.

[Tsuchida 2012] T. Tsuchida and G. W. Cottrell. *Auditory saliency using natural statistics.* CogSci 2012 Proceedings, pages 1048–1053, 2012.

[Tuomainen 2005] J. Tuomainen, T. S. Andersen, K. Tiippana and M. Sams. *Audio-visual speech perception is special.* Cognition, vol. 96, no. 1, pages B13–B22, May 2005.

[Võ 2012] M. L. H. Võ, T. J. Smith, P. K. Mital and J. M. Henderson. *Do the eyes really have it? Dynamic allocation of attention when viewing moving faces.* Journal of Vision, vol. 12, no. 13, pages 1–14, 2012.

[Van der Burg 2008] E. Van der Burg, C. N. L. Olivers and A. W. Bronkhorst. *Pip and pop: nonspatial auditory signals improve spatial visual search.* Journal of Experimental Psychology: Human Perception and Performance, vol. 34, no. 5, pages 1053–1065, 2008.

[Van der Burg 2010] E. Van der Burg, G. Brederoo, M. R. Nieuwenstein, J. Theeuwes and C. N. Olivers. *Audiovisual semantic interference and attention: Evidence from the attentional blink paradigm.* Acta Psychologica, vol. 134, no. 2, pages 198–205, 2010.

[Vilaró 2012] A. Vilaró, A. T. Duchowski, P. Orero, T. Grindinger, S. Tetreault and E. di Giovanni. *How sound is the Pear Tree Story? Testing the effect of varying audio stimuli on visual attention distribution.* Perspectives: Studies in Translatology, vol. 20, no. 1, pages 55–65, March 2012.

[Viola 2004] P. Viola and M. J. Jones. *Robust real-time face detection.* International Journal of Computer Vision, vol. 57, no. 2, pages 137–154, 2004.

[Vroomen 2000] J. Vroomen and B. De Gelder. *Sound enhances visual perception: cross-modal effects of auditory organization on vision.* Journal of Experimental Psychology: Human Perception and Performance, vol. 26, pages 1583–1590, 2000.

[Vroomen 2011] J. Vroomen and J. J. Stekelenburg. *Perception of intersensory synchrony in audiovisual speech: not that special.* Cognition, vol. 118, pages 75–83, 2011.

[Wang 2000] Y. Wang, Z. Liu and J. Huang. *Multimedia content analysis: using both audio and visual clues.* IEEE Signal Processing Magazine, no. 12-36, 2000.

[Wang 2012] F. Wang and C. W. Ngo. *Summarizing rushes videos by motion, object, and event understanding.* IEEE Trans. on Multimedia, vol. 14, no. 1, pages 76–87, Feb. 2012.

[Wilcoxon 1945] F. Wilcoxon. *Individual comparisons by ranking methods.* Biometrics Bulletin, vol. 1, no. 6, pages 80–83, Dec. 1945.

[Wolf 2000] J. M. Wolf, G. A. Alvarez and T. S. Horowitz. *Attention is fast but volition is slow.* Nature, vol. 406, page 691, 2000.

[Wolfe 1994] J. M. Wolfe. *Guided search 2.0 A revised model of visual search.* Psychonomic Bulletin and Review, vol. 1, no. 2, pages 202–238, 1994.

[Xu 2008] C. Xu, J. Wang, H. Lu and Y. Zhang. *A novel framework for semantic annotation and personalized retrieval of sports video.* IEEE Transactions on Multimedia, vol. 10, no. 3, pages 421–436, April 2008.

[Zacks 2001] J. Zacks and B. Tversky. *Event structure in perception and conception.* Psychological Bulletin, no. 127, pages 3–21, 2001.

[Zhang 2008] L. Zhang, T. K. Marks, M. H. Tong, H. Shan and G. W. Cottrell. *SUN: A Bayesian framework for saliency using natural statistics.* Journal of Vision, vol. 8, no. 7, 2008.

[Zhang 2010] C. Zhang and Z. Zhang. *A survey of recent advances in face detection.* Technical report, Microsoft Research, 2010.

[Zimmerman 1997] D. W. Zimmerman. *A note on interpretation of the paired-samples t test.* Journal of Educational and Behavioral Statistics, vol. 22, no. 3, pages 349–360, 1997.

[Zlatintsi 2012] A. Zlatintsi, P. Maragos, A. Potamianos and G. Evangelopoulos. *A saliency-based approach to audio event detection and summarization.* 20th European Signal Processing Conference (EUSIPCO), pages 27–31, August 2012.

[Zou 2012] H. Zou, H. J. Müller and Z. Shi. *Non-spatial sounds regulate eye movements and enhance visual search.* Journal of Vision, vol. 12, no. 5, pages 1–18, 2012.

List of Related Publications

1. G. Song, D. Pellerin and L. Granjon, Influence of sound on visual gaze when looking at videos, *16th European Conference on Eye Movements (ECEM 2011)*, Marseille, France, August 2011.

2. G. Song, D. Pellerin and L. Granjon, Sound effect on visual gaze when looking at videos, *19th European Signal Processing Conference (EUSIPCO 2011)*, Barcelona, Spain, September 2011.

3. A. Rahman, G. Song, D. Pellerin and D. Houzet, Spatio-temporal fusion of visual attention model, *19th European Signal Processing Conference (EUSIPCO 2011)*, Barcelona, Spain, September 2011.

4. G. Song, D. Pellerin and L. Granjon, How different kinds of sound in videos can influence gaze, *13th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2012)*, Dublin, Ireland, May 2012.

5. G. Song, D. Pellerin and L. Granjon, Different types of sounds influence gaze differently in videos, *Journal of Eye Movement Research*, under review, 2013.

Abstract — There exist mechanisms in the brain to bias attention towards particular regions, namely the salient regions. According to existing literature, the visual attention can be studied towards eye movements, however the sound effect on eye movement in videos is little known. The aim of this thesis is to investigate the influence of sound in videos on eye movement and to propose an audio-visual saliency model to predict salient regions in videos more accurately. For this purpose, we have designed two audio-visual experiments of eye tracking. In the experiments, participants watched video excerpts either with original soundtracks (AV condition), or without soundtrack (V condition). The results show that the effect of sound is different according to the types of sound and that the classes "speech", "singer", "human noise" and "singers" have the greatest effect. Finally, we proposed a preliminary audio-visual saliency model for speech and musical instrument sound classes. The audio-visual fusion strategies defined in the model improves its predictability with AV condition.

Keywords: Eye movement, Attention, Video, Sound, Audio-visual experiment, Audio-visual saliency model.

Résumé — Il existe des mécanismes dans le cerveau qui portent notre attention sur des régions particulières de notre environnement appelées régions saillantes. Alors que l'attention visuelle a fait l'objet de nombreuses études, l'effet du son sur les mouvements oculaires a encore peu été exploré. L'objectif de cette thèse est d'étudier l'influence du son dans les vidéos sur le mouvement des yeux et de proposer un modèle de saillance audiovisuelle pour prédire plus précisément les régions saillantes dans les vidéos. Nous avons conçu dans ce but deux expériences audiovisuelles de suivi du regard. Dans ces expériences, les participants ont regardé des extraits de vidéos soit avec la bande originale (condition audiovisuelle AV), soit sans bande son (condition visuelle V). Les résultats montrent que l'effet du son est différent selon les types de son et que les classes contenant de la voix humaine (classes « parole », « chanteur(s) », et « bruit humain») ont le plus grand effet. Enfin, nous avons proposé un modèle préliminaire de saillance audiovisuelle avec deux stratégies de fusion d'informations audiovisuelles : l'une pour la classe « parole », l'autre pour la classe « instrument de musique ». Ces stratégies de fusion dans le modèle améliorent la précision de prédiction des régions saillantes pour la condition AV.

Mots clés: Mouvement oculaire, Attention, Vidéo, Son, Expérience audiovisuelle, Modèle de saillance audiovisuelle.