Audiovisual spatial congruence, and applications to 3D sound and stereoscopic video.

Cédric R. André

Department of Electrical Engineering and Computer Science

Faculty of Applied Sciences

University of Liège, Belgium

Thesis submitted in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy (PhD) in Engineering Sciences

December 2013

This page intentionally left blank.

This page intentionally left blank.

Abstract

While 3D cinema is becoming increasingly established, little effort has focused on the general problem of producing a 3D sound scene spatially coherent with the visual content of a stereoscopic-3D (s-3D) movie. The perceptual relevance of such spatial audiovisual coherence is of significant interest.

In this thesis, we investigate the possibility of adding spatially accurate sound rendering to regular s-3D cinema. Our goal is to provide a perceptually matched sound source at the position of every object producing sound in the visual scene. We examine and contribute to the understanding of the usefulness and the feasibility of this combination.

By usefulness, we mean that the technology should positively contribute to the experience, and in particular to the storytelling. In order to carry out experiments proving the usefulness, it is necessary to have an appropriate s-3D movie and its corresponding 3D audio soundtrack. We first present the procedure followed to obtain this joint 3D video and audio content from an existing animated s-3D movie, problems encountered, and some of the solutions employed. Second, as s-3D cinema aims at providing the spectator with a strong impression of being part of the movie (sense of presence), we investigate the impact of the spatial rendering quality of the soundtrack on the reported sense of presence. The short 3D audiovisual content is presented with three different soundtracks. These soundtracks differ by their spatial rendering quality, from stereo (low spatial coherence) to Wave Field Synthesis (WFS, high spatial coherence). The original stereo version serves as a reference. Results show that the sound condition does not impact on the sense of presence of all participants. However, participants can be classified according to three different levels of presence sensitivity with the sound condition impacting only on the highest level (12 out of 33 participants). Within this group, the spatially coherent soundtrack provides a lower reported sense of presence than the other custom soundtrack. The analysis of the participants' heart rate variability (HRV) shows that the frequency-domain parameters correlate to the reported presence scores.

By feasibility, we mean that a large portion of the spectators in the audience should benefit from this new technology. In this thesis, we explain why the combination of accurate sound positioning and stereoscopic-3D images can lead to an incongruence between the sound and the image for multiple spectators. Then, we adapt to s-3D viewing a method originally proposed for 2D images in the literature to reduce this error. Finally, a subjective experiment is carried out to prove the efficiency of the method. In this experiment, an angular error between an s-3D video and a spatially accurate sound reproduced through WFS is simulated. The psychometric curve is measured with the method of constant stimuli, and the threshold for bimodal integration is estimated. The impact of the presence of background noise is also investigated. A comparison is made between the case without any background noise and the case with an SNR of 4 dBA. Estimates of the thresholds and the slopes, as well as their confidence intervals, are obtained for each level of background noise. When background noise is present, the point of subjective equality (PSE) is higher (19.4° instead of 18.3°) and the slope is steeper (-0.077 instead of -0.062 per degree). Because of the overlap between the confidence intervals, however, it is not possible to statistically differentiate between the two levels of noise. The implications for the sound reproduction in a cinema theater are discussed.

Acknowledgements

I am truly thankful to my two supervisors Jacques G. Verly and Jean-Jacques Embrechts at the University of Liège for letting me pursue this research on such a timely topic.

I would like to show my deepest gratitude to Brian F.G. Katz, at LIMSI-CNRS, for welcoming me in his lab and provide me with his constant support. Credit for many aspects of the results in this work goes to him as well as Marc Rébillat, formerly at LIMSI-CNRS, now at Université Paris Descartes, and Étienne Corteel, at sonic emotion labs, for their fruitful collaboration on the experiments in this thesis and their work on the SMART-I^2 platform. I would also like to thank Matthieu Courgeon, at LIMSI-CNRS, for his work on the software tool **MARC** and its integration in the SMART-I^2.

The open source movie "Elephants Dream" was a very important part of this doctoral work. I would like to acknowledge Wolfgang Draxinger who provided me with the sources of the stereoscopic version of the movie and Marc Évrard for his work on the creation of a 3D soundtrack for the movie.

This dissertation would not have been possible without the financial support of the University of Liège, under the form of a salary for a teaching assistant position and two financial aids for my research stays at the LIMSI-CNRS, in France.

Finally, I also wish to thank all the colleagues, the friends and my family who supported me along the way. In particular, I thank Pierre T. and Géraldine with whom I shared an office, the whole Audio & Acoustics team at LIMSI-CNRS for the very good time I spent there, Grégoire, Pierre P., François, Thérèse, and all the others with whom I enjoyed spending numerous lunchtimes, and all the PhD students I met during the various events of the PhD Student Network I had the chance to organize and manage. A special thanks goes to my mother whose countless sacrifices led me to where I am today.

This page intentionally left blank.

Contents

This page intentionally left blank.

Nomenclature

Roman Symbols

c	Speed of sound (340 m/s)
A, B, X, $\dots$	A number in $\mathbb{R}$ (possibly a coordinate)
f	Frequency (temporal) (Hz)
k	Wavenumber ($k = \omega/c$) (m^{-1})
M	Order of a spherical (or cylindrical) wave expansion
a, b, x, $\dots$	A point in $\mathbb{R}^3$
${}^{A}x$	The point x expressed in the coordinate system (A)
${}^{B}_{A}\mathbf{R}$	The rotation matrix describing the rotation from the coordinate system (A) to the coordinate system (B)
$\mathbf{u}$, $\mathbf{v}$, $\dots$	A vector in $\mathbb{R}^3$, or a point or a vector in homogeneous coordinates
$\overrightarrow{ab}$, $\mathbf{ab}$, or $b - a$	The vector linking a and b in $\mathbb{R}^3$

Greek Symbols

ω	Angular frequency ($\omega = 2\pi f$) (rad/s)

Acronyms

2-AFC	Two-alternative forced choice
CGI	Computer-generated images
DAW	Digital audio workstation
DCP	Digital Cinema Package
GUI	Graphical user interface

HDTV	High-definition television
HMD	Head-mounted display
HOA	Higher-Order Ambisonics
HRIR	Head Related Impulse Response
HRTF	Head Related Transfer Function
ILD	Interaural level difference
ITD	Interaural time difference
LFE	Low-frequency effects
MAA	Minimum audible angle
NFC-HOA	Near-Field Compensated Higher-Order Ambisonics
SNR	Signal-to-noise ratio
s-3D	stereoscopic-3D
SVUP	Swedish Viewer-User Presence
TPI	Temple Presence Inventory
VBAP	Vector Base Amplitude Panning
WFS	Wave Field Synthesis

Chapter
1

Introduction

Highlights

✓ The objective of the thesis is to investigate the addition of spatially accurate sound rendering, or “3D sound”, to regular stereoscopic-3D video.

✓ This thesis revolves around two main objectives, respectively concerned with usefulness and feasibility of this new technology.

✓ Some considerations on hardware and software used in this thesis are given.

✓ The contributions of the thesis are given.

✓ The organization of the manuscript is described.

Contents

1.1 Motivation for the thesis

The motivation for this thesis is the understanding and improvement of the personal experience of people watching stereoscopic-3D (s-3D) movies, in particular in a movie theater setting. The objective of the thesis is to investigate the addition of spatially accurate sound rendering, or "3D sound", to regular s-3D video.

While 3D cinema is becoming more established, little effort has focused on the general problem of producing a 3D sound scene (or *soundscape*[*]) spatially coherent with the visual scene of an s-3D movie.

3D sound aims at reproducing the correct sensation of direction and distance to the sound source as well as the room effect. The acoustic sweet-spot, where sound reproduction is controlled, can be either a portion of the space or the listener's ears.

Today's cinema sound systems are designed to immerse the spectators in sound sources that are confined to the horizontal plane at the height of the spectators' ears [Dager, 2013]. 3D audio techniques are able to produce such a rendering, but can also do much more in terms of localization. It is thus interesting to investigate the use of 3D audio in combination with s-3D video, in a movie theater environment.

The last major improvement in the spatialization of sound in movie theaters was the addition of sound channels fed uniformly to loudspeakers on the back and the sides of the theaters. These channels were introduced to immerse the audience in the movie they are watching, by producing a sound that has no particular direction of arrival. These channels have always served as a way to provide ambience as well as special, non-localized sound effects. Allen [1991] notes that, because a movie director wants to avoid taking the audience attention away from the screen, the rear channels rarely carry a sound effect which would make the spectators turn their heads. In addition, when Dolby Labs introduced Dolby Surround 5.1, attention was mostly given to the sound events in the horizontal plane, because they are far more likely than events above or under the audience [Allen, 1991].

3D audio is capable of precise localization, which makes its use in conjunction with an s-3D movie seems contrary to these principles that have guided movie directors for decades. Therefore, there is a need for psychological studies to show whether or not a realistic soundscape improves the perceptual immersion, i.e. the feeling of "being part" of the movie, by comparison to a general immersion such

[*] A soundscape is understood as the auditory equivalent of a visual landscape.

as that provided by Dolby's fourth channel.

1.2 Objectives of the thesis

We investigate here the possibility of adding spatially accurate sound rendering to regular s-3D cinema. Our goal is to provide a perceptually matched sound source at the position of every object producing sound in the scene. Note that the object need not be present in the visual field. For example, one could hear a character's steps well after the character has left the screen. Still, the sound would originate from a likely position, say to the left or right of the screen.

For the market to consider adopting this new technology, its usefulness and feasibility should be proven. By usefulness, we mean that the technology should positively contribute to the experience, and in particular to the storytelling. By feasibility, we mean that a large portion of the spectators in the audience should benefit from this new technology.

This thesis examines and contributes to the understanding of these two aspects. Therefore, the thesis revolves around two main objectives, respectively concerned with usefulness and feasibility. With respect to the "usefulness" issue, the objective is:

> to study, in the cinema context, the cognitive differences between a traditional sound rendering (stereophony), and a highly precise spatial sound rendering (Wave Field Synthesis or WFS). In particular, we will examine whether a higher spatial coherence between sound and image leads to an increased sense of presence for the audience.

In order to achieve this first objective, we investigated different ways to obtain a content combining 3D audio with s-3D video. We found that the easiest and most cost effective way to carry out a first investigation was to recreate the soundtrack of an existing animation movie. Indeed, when the 3D scenes used to render one such movie are known, it is possible to extract the coordinates of the visual objects emitting sound in an automatic fashion. Then, the 3D audio soundtrack is obtained by assigning these coordinates a sound.

With respect to the "feasibility" issue, the objective is:

> to evaluate the possible angular error between the sound and the image when presenting precise spatial sound through Wave Field Synthesis (WFS) in combination with s-3D video to spectators seated at different locations.

As we will see later on, the visual perception of two spectators looking at an s-3D image depends on their respective location. To put it simply, if an object in the scene is supposed to appear halfway between the screen plane and the spectator, it will appear so for every spectator. Therefore, each spectator perceives the visual object at a different location in a room, depending on his own position. In addition, unless sound is reproduced via headphones, there can only be one sound source corresponding to one visual object. As a result, the position of the sound source will be chosen with respect to one "ideal" seating position, leading to a potential error between the sound and the image for spectators seated at another position. The perception of this error will be quantitatively evaluated in order to verify that a large portion of the audience can benefit from spatially accurate sound rendering in a movie theater.

1.3 Hardware and software considerations

1.3.1 Hardware

When working with s-3D imaging and 3D sound, the minimum equipment needed would be a computer equipped with an s-3D capable display and headphones. These peripherals can be bought off-the-shelf at a price around 300€ in early 2013 (that is, not counting the computer itself). This is arguably low cost. However this does not emulate "cinema" conditions. The large scale equivalent equipment is an s-3D digital projector and a loudspeaker rig (linear array, hemisphere, sphere, . . .), with a price that can easily reach 30 000€. Such an equipment is still rarely encountered. Systems that combine 3D sound with s-3D imaging in laboratories are often virtual reality systems [Amatriain et al., 2009; Lentz et al., 2007; Rébillat et al., 2008]. This is why we collaborated with Brian Katz, PhD, from the LIMSI-CNRS, who could provide us, in addition to his large experience in 3D sound, with the necessary equipment.

1.3.2 Open-source software

In the course of this work, open-source software was used wherever possible. The reasons for this are:

- Easy access, usually through a simple download from the internet.
- Low cost. Open-source software is free.

- Portability. Open-source software often runs on multiple platforms, including commercial ones.
- Possibility of customization. If a software tool does not have a specific feature, the missing feature can be coded and added to the software.

Free software has tremendously evolved in the recent years. It makes it possible for a scientist to perform his work on a stable platform, with the guarantee that the work remains usable for quite some time. The list of open-source software tools used within the course of this thesis is given in Appendix A.1.

1.3.3 Proprietary software

A few proprietary software tools were used in this work. The reason these proprietary tools were used is that they were already being used by colleagues of the author. Because these were commercial off-the-shelf products, their use does not impair the repeatability of the results. The list of the proprietary software tools used for this thesis is given in Appendix A.2.

1.4 Review of the key concepts

Given the highly multidisciplinary topic, we begin by providing a review of the key topics involved, this for the benefit of the reader. This thesis mainly builds on auditory perception and visual perception, sound engineering, stereoscopic imaging, geometry, psychophysics, and statistics. Chapters 2 to 5 provide a review of the concepts that are necessary for the understanding of this thesis.

1.5 Contributions of the thesis

The novel contributions of this thesis are as follows.

First, we collected from various sources the scientific and technical knowledge necessary to reach the objectives of the thesis. To the best of our knowledge, this knowledge has never been put in writing in a single document.

Second, we defined and implemented a strategy for producing a true, experimental 3D audiovisual content [Évrard et al., 2011].

Third, we conducted an experiment, using this newly created content, to measure the impact of spatial sound rendering on the reported sense of presence of the spectators [André et al., 2012].

Fourth, we adapted an existing method to s-3D to reduce the angular disparity between spatially accurate sound and s-3D video [André et al., 2013].

Fifth, we conducted an experiment evaluating the impact of an off-axis seating on the perceived audio-visual congruence [André et al., 2014].

Sixth, we evaluated the angular disparity range between the auditory and visual stimuli that provides the same feeling of congruence as compared to no angular disparity [André et al., 2013].

1.6 Organization of the manuscript

Chapter 2 introduces the reader to auditory perception, visual perception, and auditory-visual perception.

Chapter 3 focuses on the reproduction of spatial sound in a movie theater. First, stereo sound is discussed and a nomenclature of one and two-amplifying channel audio systems is given. Then, a short historical review describes the technological evolution of sound systems in movie theaters. Finally, different 3D audio technologies are discussed.

Chapter 4 describes in more detail the principles of s-3D visualization. The whole processing chain is considered, from capture through transmission to reproduction. Then, the differences between natural human viewing and s-3D viewing are discussed, as well as the potential consequences of these differences.

Chapter 5 lists the new challenges related to the combination of 3D sound with s-3D imagery. It then describes the SMART-I^2, the 3D audiovisual platform used in this work. Chapter 5 concludes the review of the key concepts.

Chapter 6 marks the beginning of the novel contribution of the thesis. In this chapter, we describe the process we used to obtain a first 3D audiovisual content. We implemented a true 3D soundtrack for an existing s-3D animation movie using object-based sound mixing. We designed the new audio track to be as similar as possible to, and inspired by, the original soundtrack.

Chapter 7 presents the results of an experiment measuring the impact of spatial sound rendering on the sense of presence, as reported by the spectators. We compared the cognitive differences between a traditional sound rendering (stereo), and a highly precise spatial sound rendering (Wave Field Synthesis or WFS). The experiment used the 3D audiovisual content introduced in Chapter 6. Using a post-stimuli questionnaire based on previous reports regarding the sense of presence, various cognitive effects were extracted and compared. These results were also compared to measures extracted from the spectators' electrocardiographic (ECG) recordings.

Chapter 8 presents the results of an experiment evaluating the impact of an off-axis seating position on the perceived audio-visual congruence. In the study of Chapter 7, all the subjects were seated at the ideal position, to guarantee an accurate sound reproduction. In the study of Chapter 8, we consider the potential angular error between the sound and the image when presenting precise spatial sound through WFS in combination with s-3D video to spectators seated at different locations. The spectators evaluated the spatial coherence between a displayed virtual character and a reproduced speech sound. The sensibility of the perceived coherence was also tested in the presence or absence of additional background noise.

Chapter 9 summarizes the work performed in this thesis. A general discussion of the results is given as well as some pointers for future work.

Bibliography

Allen, I., Feb. 1991. Matching the sound to the picture. In: Audio Eng. Soc. 9th Int. Conf.: Television Sound Today and Tomorrow.
http://www.aes.org/e-lib/browse.cfm?elib=5352 2

Amatriain, X., Kuchera-Morin, J., Hollerer, T., Pope, S. T., 2009. The AlloSphere: immersive multimedia for scientific discovery and artistic exploration. IEEE MultiMedia 16 (2), 64–75.
http://dx.doi.org/10.1109/MMUL.2009.35 4

André, C. R., Corteel, É., Embrechts, J.-J., Verly, J. G., Katz, B. F., Jan. 2014. Subjective evaluation of the audiovisual spatial congruence in the case of stereoscopic-3D video and Wave Field Synthesis. International Journal of Human-Computer Studies 72 (1), 23–32.
http://dx.doi.org/10.1016/j.ijhcs.2013.09.004 6

André, C. R., Corteel, É., Embrechts, J.-J., Verly, J. G., Katz, B. F. G., 2013. A new valited method for improving the audiovisual spatial congruence in the case of stereoscopic-3D video and Wave Field Synthesis. In: 2013 International Conference on 3D Imaging (IC3D). pp. 1–8. 6

André, C. R., Embrechts, J.-J., Verly, J. G., Dec. 2009. Adding sound to movies:

historical trends and new challenges of 3D. In: 3D Stereo MEDIA Conf. Liège, Belgium.
http://hdl.handle.net/2268/32679

André, C. R., Embrechts, J.-J., Verly, J. G., Nov. 2010a. Adding 3D sound to 3D cinema: Identification and evaluation of different reproduction techniques. In: Proc. 2nd Int. Conf. on Audio Language and Image Processing (ICALIP 2010). Shanghai, China, pp. 130–137.
http://hdl.handle.net/2268/73310

André, C. R., Embrechts, J.-J., Verly, J. G., May 2010b. Adding 3D sound to 3D cinema: new challenges and perspectives. In: Proc. of the URSI Forum 2010. Bruxelles, Belgium.
http://hdl.handle.net/2268/62686

André, C. R., Rébillat, M., Embrechts, J.-J., Verly, J. G., Katz, B. F. G., 2012. Sound for 3D cinema and the sense of presence. In: Proc. of the 18th Int. Conf. on Auditory Display (ICAD 2012). Atlanta, GA, pp. 14–21.
http://hdl.handle.net/2268/127803 5

Dager, N., Oct. 2013. A short history of cinema sound. Online (Accessed 14-November-2013).
http://www.digitalcinemareport.com/article/short-history-cinema-sound 2

Évrard, M., André, C. R., Verly, J. G., Embrechts, J.-J., Katz, B. F. G., 2011. Object-based sound re-mix for spatially coherent audio rendering of an existing stereoscopic-3D animation movie. In: Audio Eng. Soc. Conv. 131. New York, NY.
http://hdl.handle.net/2268/108048 5

Lentz, T., Schröder, D., Vorländer, M., Assenmacher, I., 2007. Virtual reality system with integrated sound field simulation and reproduction. EURASIP Journal on Advances in Signal Processing 2007 (1), 187–187.
http://dx.doi.org/10.1155/2007/70540 4

Rébillat, M., Corteel, É., Katz, B. F. G., Oct. 2008. SMART-I^2: Spatial Multi-User Audio-Visual Real Time Interactive Interface. In: Audio Eng. Soc. Conv. 125.
http://www.aes.org/e-lib/browse.cfm?elib=14760 4

Part I

State of the art

Chapter
2

Overview of auditory, visual, and auditory-visual perception

Highlights

✓ A review on auditory spatial perception is given.

✓ A review on visual spatial perception is given.

✓ A review on auditory-visual perception is given.

Contents

This chapter presents a review of literature on the subjects of auditory perception, visual perception, and auditory-visual perception.

Concerning the use of the words "auditory" and "audio", one should note that "auditory" is generally found in the literature on perception, while "audio" is generally used in the literature to reference electrical or other representations of signals containing audible information. We will usually keep this distinction and talk, for example, about an auditory-visual stimulus, and an audio-visual equipment.

2.1 Overview of auditory spatial perception

The human (peripheral) auditory system is made up of two fixed ears on each side of the head (Figure 2.1). They are the primary point of entrance of sound waves in our sensory system. We often associate a location to a sound source. The location is extracted by the brain from the information contained in the two sound-wave signals entering the auditory canals. The direction of origin of the sound source is determined solely from the wave pressures in the auditory canals [Middlebrooks et al., 1989; Wiener and Ross, 1946].

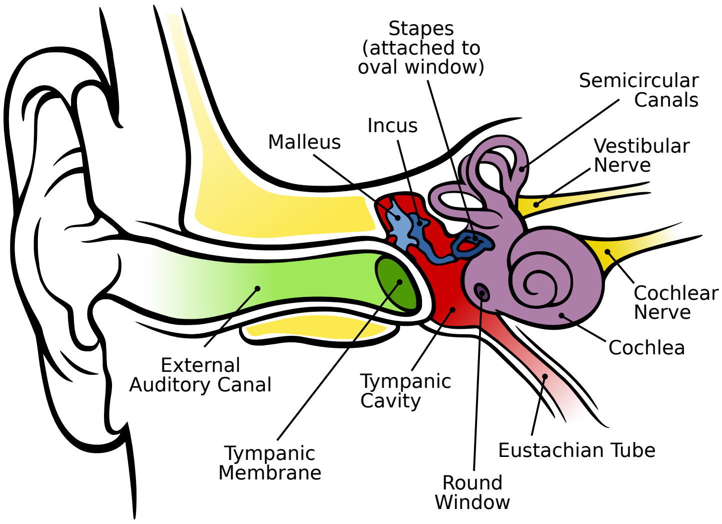


Figure 2.1: A diagram of the anatomy of the human ear. From [Chittka and Brockmann, 2005; Wikipedia, 2013a]. License: Creative Commons Attribution 2.5 Generic.

We are only sensitive to a part of the spectrum of the sound signal called the auditory range. For a young, healthy listener, this range is approximately 20 Hz to 20 kHz. We can hear sound coming from literally everywhere around us, provided it reaches us with a sufficient intensity. At 1 kHz, the minimal air pressure that can be detected by the auditory system is 20 µPa. This is the reference level for the dB SPL (Sound Pressure Level) units.

Because the mechanisms that determine the location of a sound are different in the horizontal plane and the vertical plane, it is customary to express the auditory space in spherical coordinates, with the center of the head at the origin, the X-axis pointing forward (i.e. towards the nose), and the Z-axis pointing upward. The Y-axis is defined in such a way that the X, Y, and Z-axes form a right-handed coordinate system. These axes are shown in Figure 2.2, which also shows the azimuth angle θ and the elevation angle ϕ. The *median plane* is the vertical plane XZ passing through the nose, or the plane $\theta = 0$. When a sound comes from the side of head, the ear located on the same side of the head as the sound source is the *ipsilateral* ear, and the ear on the opposite side of the head is the *contralateral* ear.

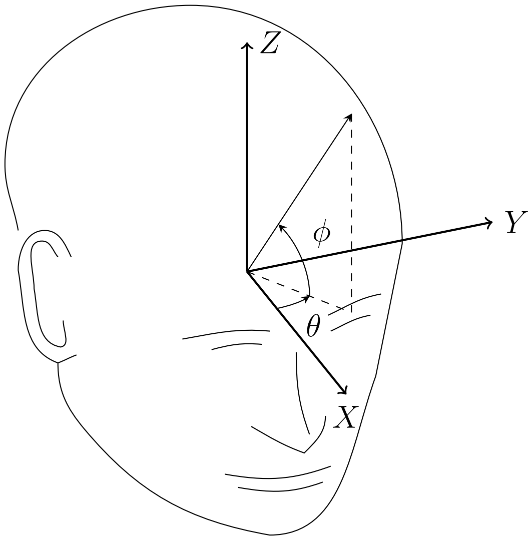


Figure 2.2: Geometry for auditory localization: X, Y, Z coordinate frame, angles θ and ϕ, and median plane XZ.

2.1.1 Sound localization in the horizontal plane

Because of the placement of our two ears, a sound coming towards our head must travel further to reach the contralateral ear than it has to reach the ipsilateral ear.

This results in an *interaural time difference*[*] (ITD) [Stewart, 1920a] between the signals at the two ears. The ITD is illustrated in Figure 2.3 for the ideal case of a spherical head of radius a.

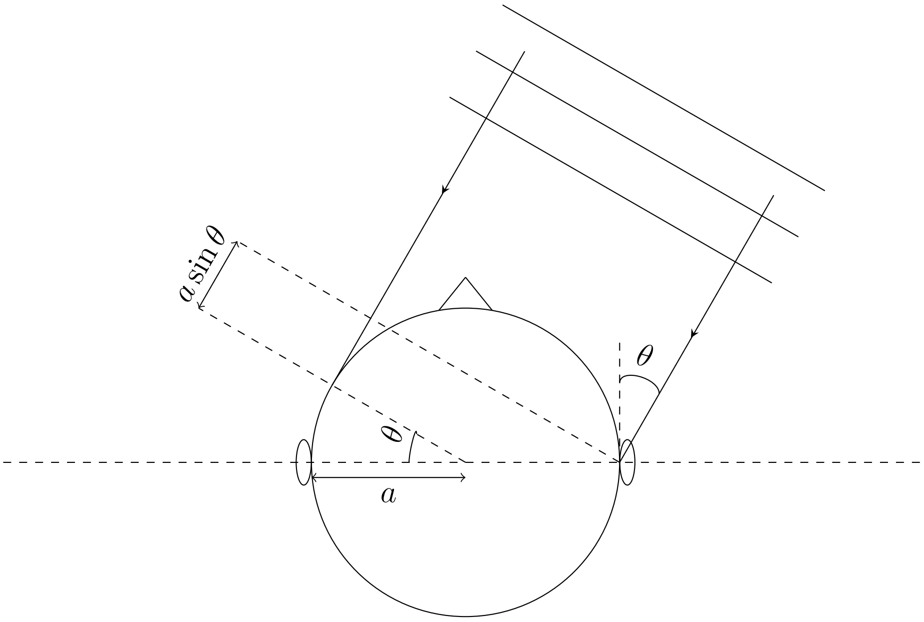


Figure 2.3: A (planar) sound wave impinging on an ideal spherical head from a distant source. An interaural time delay (ITD) is produced because it takes longer for the signal to reach the contralateral ear (here the left ear). As a first approximation, the ITD is equal to $a(\sin\theta + \theta)/c$. After Stern et al. [2006].

Physically, the usefulness of the ITD resulting from simple signals, such as sinusoids, is limited to longer wavelengths, or, equivalently, to low frequencies. At a lower wavelength, when the wavelength becomes on the order of the head size, the sound wave diffracts around the head. In practice, the ITD is an important cue when the signal contains frequencies below around 1.5 kHz [Blauert, 1997]. Indeed, at frequencies above around 1.5 kHz, it is not possible to extract the azimuth θ from the phase difference between the sinusoidal signals at the left and right ears (the phase difference is larger than 2π).

The auditory system can also find an ITD in a frequency rich signal. McFadden and Pasanen [1976], for example, distinguish between time delays at onset, the only cue for sounds shorter than about 1 ms, and ongoing time delays. These ongoing delays are further divided into two categories: (1) the delays resulting from fine structure analysis of the signal, which correspond to the ITD of [Stewart, 1920a], and (2) the delays found in the envelope of the signal. The ongoing interaural

[*] When the signal is purely sinusoidal, one can also talk about *interaural phase difference*.

differences have been shown to be much more important than the onset difference for sufficiently long stimuli (longer than about 100 ms) [Buell et al., 2008; Tobias and Schubert, 1959]. Still, the sensitivity to ITDs at high frequencies is lower than the sensitivity to ITDs at low frequencies.

When the wavelength is smaller than the diameter of the head, the head becomes an obstacle to the propagation of the wave and the pressure level at the contralateral ear is lower than that at the ipsilateral ear [Stewart, 1920b]. The resulting *interaural level difference* (ILD) is an important cue for signals containing frequencies above around 1.5 kHz. The ILD cue becomes negligible at frequencies below 1 kHz [Begault, 1994].

The ITD and ILD are binaural cues, and together are referred to as the *duplex theory* of sound localization [Rayleigh, 1907]. Despite their importance in horizontal localization, the binaural cues are ambiguous for elevation in the median plane, where the ITD and ILD are constant, and on the *cone of confusion*, the imaginary locus of source positions, centered on the interaural axis, which result in constant ITD and ILD cues (see Figure 2.4). The cone of confusion results in front/back ambiguities and elevation ambiguities. A listener can easily resolve the front/back ambiguities by rotating his/her head [Thurlow and Runge, 1967; Wightman and Kistler, 1999]. This requires that the stimulus be long enough, but Perrett and Noble [1997a] already obtained a significant reduction in front/back errors with a 0.5 s-long stimulus.

It has also been shown that it is possible to localize sounds in the horizontal plane with just one ear. The visible part of the ear, or *pinna*, acts as a filter when it collects sound and directs it into the ear canal. Because of the cavities and the convolutions of the pinna, the amplitude of some frequencies are amplified while that of other frequencies are reduced, depending on the source position. Although this effect is mainly useful in elevation estimation, as we will see later on, it also plays a role in resolving front/back ambiguities and a marginal role in horizontal localization, when the stimulus contains high frequencies [Flannery and Butler, 1981; Musicant and Butler, 1984].

2.1.2 Sound localization in the median plane

An early experiment by Blauert [1969] suggested that participants judged the direction of origin of a one-third octave noise signal in the median plane on the basis of its center frequency rather than of its physical direction of origin, i.e. that of the loudspeaker. The spectrum of a sound signal is largely modified by

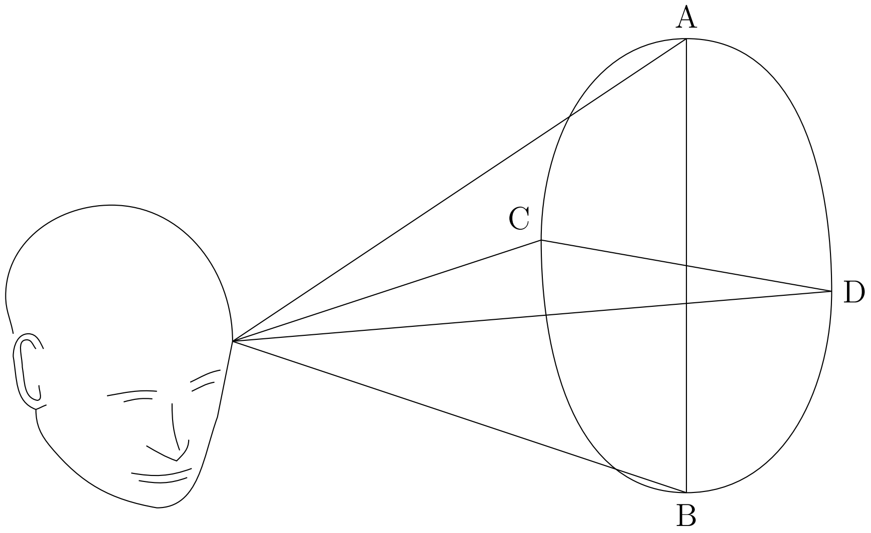


Figure 2.4: The cone of confusion has its apex at one ear and its axis is the line crossing the two ears. Its basis is circular. Any two sound sources diametrically opposed on a circular cross-section of the cone (say A and B, or C and D) produce the same binaural cues to the listener.

the presence of the pinna, the head, and the torso, which all cause absorption and reflections [Gardner, 1973; Roffler and Butler, 1968]. These spectral cues are generally considered to be the most important cues to localization in the median plane. However, just as spectral cues are secondary cues to horizontal localization, binaural cues serve as secondary cues to vertical localization.

Since localization in the median plane relies primarily on frequency effects, the presence or absence of certain frequency bands has an impact on the performance of the subjects in elevation localization tasks. King and Oldfield [1997], for example, have shown that progressively reducing the bandwidth of a signal reduces the ability to perceive the elevation of sound and increases the chances of front/back confusion. This result was further confirmed by Carlile et al. [1999], who showed that front/back confusions happen when the frequencies above 2 kHz were removed from the signal. Indeed, at low frequencies, the phase difference is the only remaining cue and this cue is ambiguous for elevation.

According to Asano et al. [1990], the perception of elevation of white noise is concentrated in the two regions between about $4-5$ kHz and $8-10$ kHz. Using 1-octave signals, Langendijk and Bronkhorst [2002] have shown that up/down cues are located mainly in the $6-12$ kHz band, and front/back cues in the $8-16$ kHz band.

These frequency effects are summarized by the Head Related Impulse Responses (HRIRs), or, in the frequency domain, by the Head Related Transfer Functions (HRTFs), which are the Fourier transforms of the HRIRs. HRTFs consist in a set of two functions, one for each ear, each representing the interaction of sound with the listener's head and torso for that ear. HRTFs depend on the position of the source and vary significantly from one subject to another [Møller et al., 1995].

Just as in the case of horizontal localization, head movements improve the localization in the median plane [Thurlow and Runge, 1967]. Head rotation also helps for up/down judgment [Perrett and Noble, 1997b].

2.1.3 Auditory distance perception

Here, we limit our analysis to the far-field, where the curvature of the wavefront need not be taken into account. A review on auditory distance perception can be found in [Zahorik et al., 2005].

It is generally accepted that the perceived estimated distance D' to the source can be related to the actual distance D by the compressive power function

$$D' = kD^a \tag{2.1}$$

where k and a are parameters of the fit. By fitting 84 datasets from several studies, Zahorik et al. [2005] found that k is close to one (mean $\{k\} = 1.32$) and that a is consistently less than one (mean $\{a\} = 0.54$). This results in overestimation of close distances and underestimation of far distances, as seen in Figure 2.5.

A variety of cues serve to the perception of the auditory distance to a sound source. A distinction must be made between a relative cue and an absolute cue. A relative cue allows the listener to compare the differences in distance between sources and the absolute cue allows him or her to give a distance measure for each source. Most often, familiarity with a sound source makes a relative cue become absolute for this particular sound source.

We now turn to the description of auditory distance cues.

Intensity The acoustic intensity of a sound source gives information on the distance to this source because the intensity decreases with increasing distance to the source [Coleman, 1963; Gamble, 1909]. The precise relationship between intensity and distance depends on properties of the environment and of the source. In the free field*, the intensity of a point-source of fixed power

*The free field is an environment free from any reflective surface in any direction.

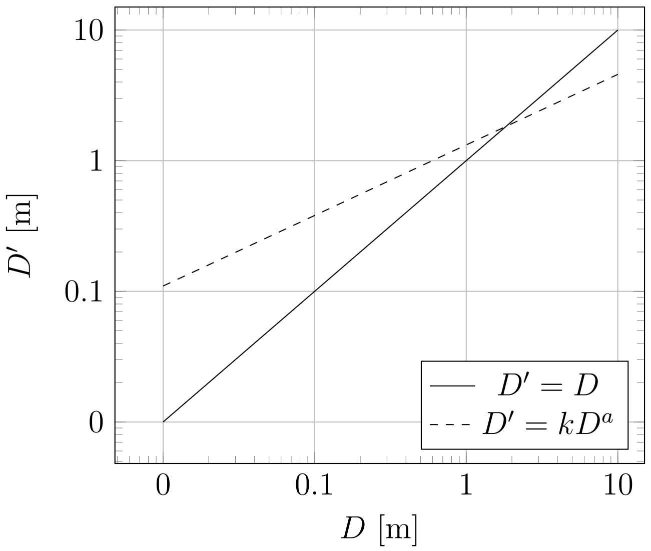


Figure 2.5: The perceived distance D' as a function of the actual distance D to the source on a log-log scale. The dashed line corresponds to the ideal case where $D' = D$. The continuous line corresponds to the relation $D' = kD^a$ with $k = 1.32$ and $a = 0.54$, which characterize human perception.

decreases by 6 dB for every doubling of distance. The intensity is a relative cue. It requires familiarity with the source to become absolute [Mershon and King, 1975].

Air absorption At large distances, the properties of the air modify the sound spectrum, mainly by attenuating the high frequencies [Coleman, 1963, 1968]. It is a relative cue [Little et al., 1992].

Motion parallax and variation of loudness with time When both the observer and the sound source are in translation relative to one another, the motion parallax, that is, the angular movement of the source perceived by the listener, gives a small cue about the distance to the source [Speigle and Loomis, 1993]. Also for a translation movement between the observer and the sound source, the ratio between the intensity and its time-derivative is linked to the time-to-contact, and, assuming constant velocity, gives information about the distance to the source [Shaw et al., 1991].

Direct-to-reverberant ratio In a reverberant environment*, part of the sound from a source reaches the listener directly (the primary source), and the rest arrives after interacting with reflecting surfaces (the secondary sources). The

*A reverberant environment, contrary to the free field, contains reflective surfaces.

ratio of energies of the direct part and the reflected part gives information on the distance to the source. As a first approximation, the (late) reverberant part can be approximated by a diffuse field of constant energy with respect to the position of the secondary source. Therefore, a close source induces a larger direct-to-reverberant ratio. Since this cue varies from room to room, it is generally a relative cue. When the room is familiar, the direct-to-reverberant ratio becomes an absolute cue [Mershon and King, 1975].

2.1.4 Non-acoustic cues

Our perception of the surrounding world is by essence multimodal, that is, we integrate information across our senses to build an accurate and reliable percept of the world. This holds true of course for our ability to localize sound sources in space. Powerful cues about the location of a sound source are derived from non-acoustic sources of information.

Vision Auditory localization can be strongly influenced by the presence of a potential visual target. The impact on azimuth evaluation is called *ventriloquism* [Thurlow and Jack, 1973]. This perceptual phenomenon, where a visual source position captures the position of a discrepant sound source, will be more thoroughly reviewed in Section 2.3.2. A visual stimulus can also influence the evaluation of the auditory distance. Zahorik [2001] found that the addition of a visual stimulus improved the distance judgment accuracy and lowered judgment variability compared to the auditory-only stimulus.

Prior knowledge and familiarity In [Coleman, 1962], listeners presented with an array of loudspeakers at different distances had to judge which of them was playing an unfamiliar stimulus. At first, the listeners were unable to correctly identify the source and consistently associated the source to the closest loudspeaker, but gradually improved their judgment at each trial. This indicates that increasing familiarity can lead to better localization accuracy. Familiarity also turns a relative cue into an absolute one.

Expectation Speech signals are familiar to all listeners. These signals have particular features that one uses, for example, to distinguish between shouting and whispering. In anechoic conditions, Gardner [1969] showed that the distance to a source of whispered live speech was underestimated while the distance to its shouted equivalent was overestimated. Gardner argued that listeners probably associated whispering with being close to the interlocutor, and shouting with being away from him or her.

2.2 Overview of visual spatial perception

As humans, we see the world with our two eyes located side by side on our face. Figure 2.6 shows a cross-section of the eye with the name of some of its key parts. We are only sensitive to a part of the electromagnetic spectrum called the visible spectrum. For an average viewer, this range includes wavelengths from roughly 400 to 700 nm. The monocular visual field of a normal eye extends horizontally to approximately 60° towards the nose and 100° away from the nose [Spector, 1990]. This field extends vertically to about 60° upwards and 75° upwards. The binocular region, where the two monocular visual fields overlap, extends horizontally to about 120°.

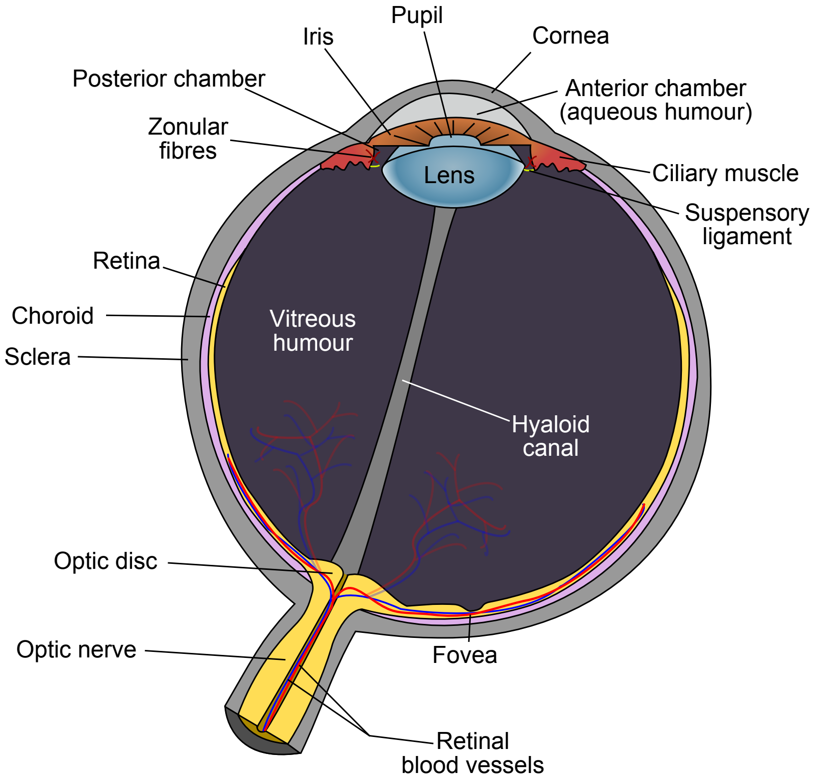


Figure 2.6: A diagram of the anatomy of the human eye. From Wikipedia [2013b]. License: public domain.

Our eyes work as an optical system, which captures the light that originates from our surroundings. The iris dynamically controls the quantity of light that enters the system. The iris thus plays the role of an aperture stop, and the pupil constitutes the aperture. The light is focused by the cornea and the (crystalline) lens on the retina. In optical terms, the cornea has a fixed focal length. This means that the object is in focus (the image of a point is a point) only at a given distance. To allow us to see sharply, the focus of the eye is corrected by the lens, which is dynamically controlled through a process called *accommodation* [Atchison, 1995]. The retina, where the light is focused, contains photoreceptor cells, i.e. cells that are sensitive to the incident light. These cells transmit their information to neurons that form the optical nerve, linked directly to the brain. The presence of the optical nerve results in a blind spot, called the optical disc.

In optical terms, the optical axis of the eye is the imaginary line that passes through each curvature center of the lens, or more accurately, its closest fit, because the centers do not lie on a single line. The visual axis is the line that passes through the object of attention and the fovea, which is the region of the retina with the highest acuity. In order to track objects in the visual field, the eyes can move in their orbits. The opposite movements of the visual axes that allows the projection of the fixated object to remain on each fovea is called the *vergence*. It is coupled to the process of accommodation [Fincham and Walton, 1957].

In the retina, one finds two types of photoreceptors, namely rods and cones. Their spatial distributions in the retina are highly non-uniform, as can be seen in Figure 2.7. Rods are by far the most numerous: a typical retina contains about 100 million of these cells [Roorda, 2002]. They are monochromatic and their high sensitivity makes them responsible for our vision at low light intensities (*scotopic* vision). Several rods are connected to one neuron in order to amplify the information signal. This grouping results in a lower spatial resolution. Because the rods are not useful for the perception of the visual stimuli we are most interested in in this thesis[*], they are not further discussed here.

Cones are less numerous than rods. There are about 5 million cones and they are responsible for our vision at daylight (*photopic* vision). They come in three types, L (long), M (middle) and S (short) depending on the wavelength range they are sensitive to. The spectral sensitivity of the three types of cones is shown in Figure 2.8. The spatial distribution of cones is also non-uniform. L and M cones

[*] To be precise, the extreme case of a screen presenting a completely black image would fall in the range of *mesopic* vision, where our percept is a mixture of signals from rods and cones [Kennel, 2012].

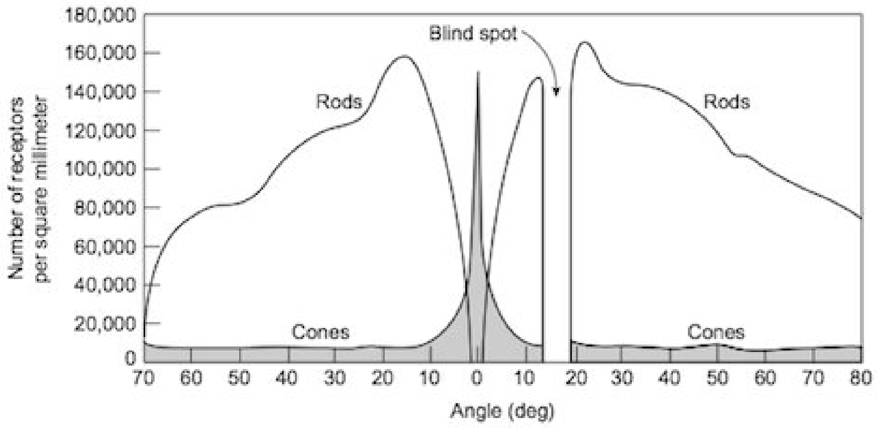


Figure 2.7: The spatial density of rods and cones in the human retina. Only cones are present in the fovea and rods dominate outside this region. After Osterberg [1935]. Image downloaded from www.uxmatters.com.

are concentrated in the fovea while the S cones are spread around the rest of the retina. The eye does not need a high spatial resolution at the shorter wavelengths because of the chromatic aberration of the eye optical system [Wandell, 1995]. The refractive index of a lens is dependent on the wavelength of the incoming light. Therefore, different wavelengths focus at different distances from the lens.

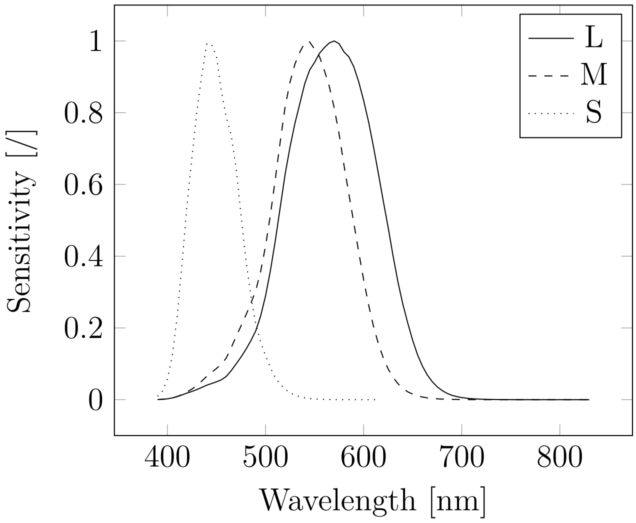


Figure 2.8: Spectral sensitivities of the L, M and S cones in the human eye. Data from Stockman and Sharpe [2000], also available on the Internet.

2.2.1 Visual perception of direction

We look at our surroundings with two spatially separate eyes. Our brain therefore catches two different views of the same scene. Still, under most circumstances, single objects appear to us. Our brain assigns to each object a single visual direction from the two disparate visual directions of each eye. That is, one must distinguish between the egocentric direction, which specifies the location of the object with respect to the head, and the oculocentric direction, which specifies the location with respect to the retinal projection of each eye. These may or may not be identical directions depending on the viewing condition. The brain must therefore deduce the egocentric direction from the two oculocentric directions and the rotation of both eyes.

The principles of visual direction were defined by Hering [Ono, 1979]. They explain how we form this unique perception of direction from the two different directions suggested by our eyes. He called the *visual line* the line passing through any object and the aperture of one eye, where all the light rays cross. The principal visual line, or *visual axis*, is the visual line of the fixated object, which goes through the fovea. Hering suggested we see the world as if we had a third *cyclopean* eye located midway between our two eyes (see Figure 2.9). An object which stimulates one or both foveæ will be seen on the line passing through the cyclopean eye and the intersection of the two visual axes (the fixation point). An object which stimulates any other position of either retina will be seen from the cyclopean eye with an angular deviation from the line passing through the cyclopean eye and the fixation point.

Sometimes, however, the information from the two eyes is not fused and double vision occurs. Only certain combinations of points on the retinæ give rise to a unique perception of direction. A point on the retina is said to be *corresponding* to a point in the other retina if their simultaneous stimulations give rise to a unique perception of direction [von Noorden and Campos, 2002]. As a first approximation, these points are on, or equidistant to, the two foveæ. On the one hand, the excitation of corresponding points gives rise to a unified, single percept, through a process called *stereopsis*, provided the images on the retinæ are similar in size, brightness, and sharpness. On the other hand, the excitation of non-corresponding points gives rise to double vision or *diplopia*.

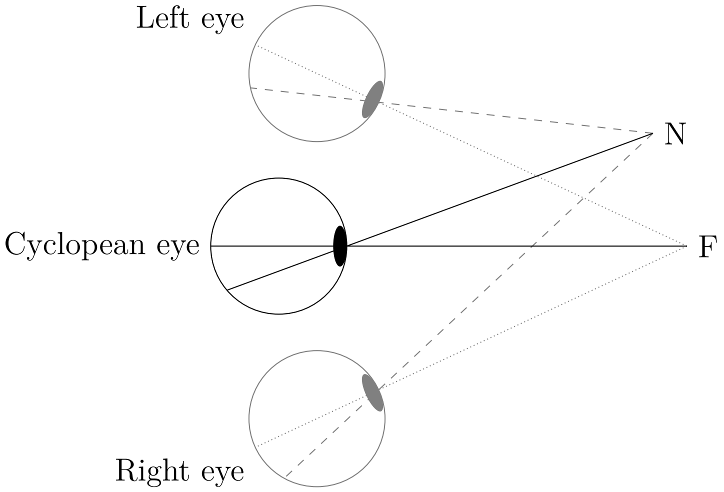


Figure 2.9: The determination of the direction of a point N when the eyes fixate the point F, using the cyclopean eye model. The images of F fall on the fovea in each eye, and the dotted lines from F to the retinæ are the visual axes. The images of a nearer point, N, fall on the temporal retina of the right eye and the nasal retina of the left eye, and the lines from N to the retinæ are the visual lines. Because the images of N are equidistant from the foveæ, the cyclopean eye is used to determine its corresponding unique perceived direction.

2.2.2 Visual perception of depth

Thanks to the perception of visual depth, humans can evaluate visual distances quite accurately. The information we obtain about the layout of the visual scene surrounding us is not always expressed in absolute units, like meters. It is therefore useful to differentiate several notions relating to the characterization of this layout.

> Egocentric or absolute distance is the distance from the observer to a point in the scene. Relative distance (absolute depth) is the depth separation between two points in egocentrically scaled units. Relative depth is the non-metric depth ordering of two or more points, or the ratios of depth separations between more than two points (e.g., slant). (Vishwanath and Blaser [2010])

As in the case of the auditory distance evaluation, the compressive power function $D' = kD^a$ is also used to link the real distance to the visual egocentric distance. The exponent corresponding to the visual perception is usually slightly less than 1 [Cook, 1978]. In a review, Da Silva [1985] showed that, although there were large inter-individual differences, a mean exponent of 0.9 approximates fairly well all the reviewed studies.

Recent experiments using a head-mounted display (HMD) revealed that neither binocular presentation (as opposed to monocular presentation) or a fixed interocular distance (as opposed to an interocular distance fitted to the viewer's) are the cause of the observed compression in distance [Willemsen et al., 2008].

Still, experiments involving a visually directed action rather than direct judgment of distance show that the underestimation of close distances is due to the employed experimental methods rather than an incorrect mental representation of space [Loomis et al., 1992], at least up to distances of 15 m [Fukusima et al., 1997].

Recent evidence also suggests an influence of the depicted environment [Interrante et al., 2006]. The compression of distance disappeared when participants were presented with a virtual environment which replicated perfectly a real environment that they had previously experienced. However, the hypothesis that participants formed a metrically accurate mental model of the real environment and used it to calibrate the virtual environment can be discarded [Interrante et al., 2008].

Another limitation of the compressive power model appears at large distances (more than 30 m). Not all distances are underestimated. Recent evidence suggests that distances up to about 100 m are underestimated and distances beyond that are overestimated [Daum and Hecht, 2009].

Different reviews give different lists of cues associated to our perception of depth. We report here the list proposed by Cutting and Vishton [1995], who also indicated the relative strength of these cues relative to the others as a function of the distance between the observer and the object of interest. This is shown in Figure 2.10. Here, we split the cues into two categories, the monocular cues, which are available to us even when one of our eyes is closed, and the binocular cues, available only when our two eyes catch a different image of the same scene.

Occlusion Occlusion occurs when an opaque objects hides, at least partly, another object behind it. This cue gives information on the order of the objects in the scene. However, this cue is purely ordinal: it does not give any information on the distance separating the two objects. It is a reliable cue at all distances (Figure 2.11(b)).

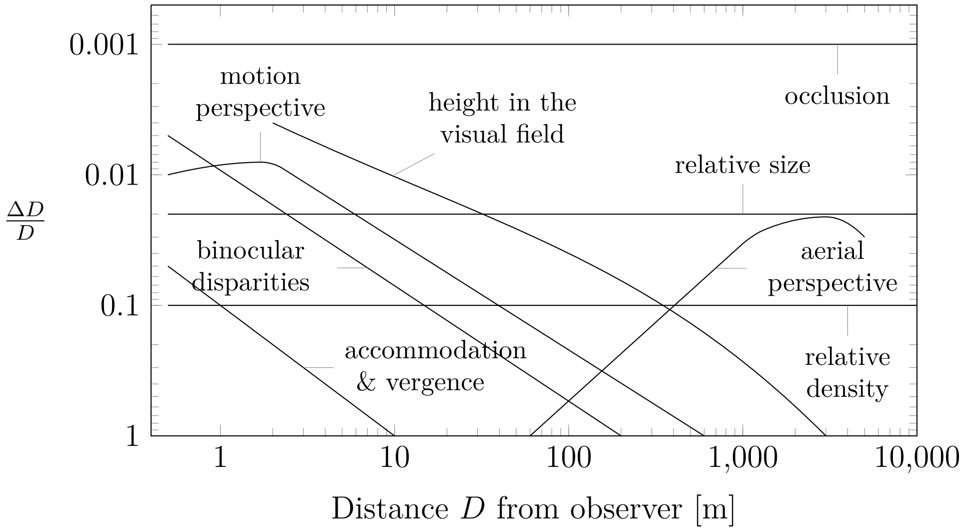


Figure 2.10: Just-discriminable depth thresholds as a function of the log of distance from the observer, from 0.5 to 5000 m, for nine different sources of information about scene layout. After Cutting and Vishton [1995].

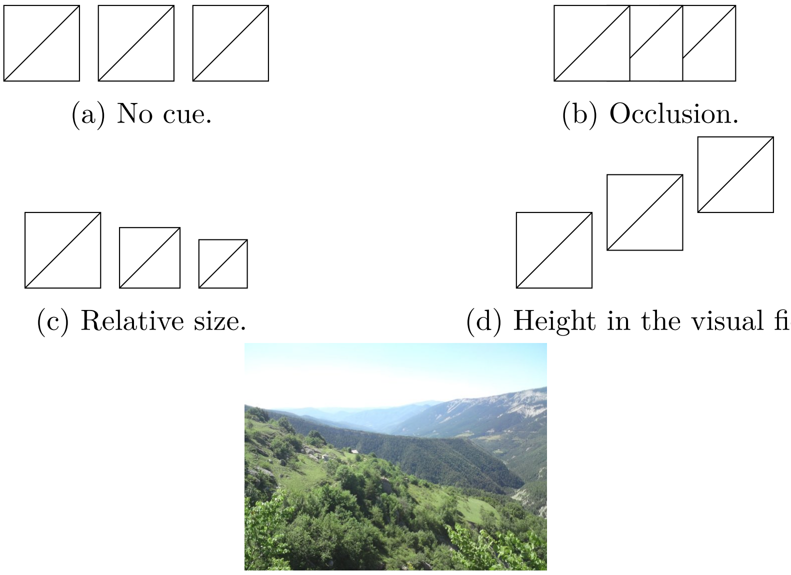
Relative size and relative density When two objects are known to be of similar size, but their absolute size is unknown, then the difference between their projected size on the retina gives information on their relative positions. The object with the largest size on the retina is closer than the other. When the absolute size of one object is known to the observer, then the cue becomes absolute. Like occlusion, this cue is fairly reliable at all distances (Figure 2.11(c)).

Height in the visual field When looking at the scene from bottom to top, the order of the basis of objects lying on the opaque ground gives ordinal information on the layout of the scene. This information becomes absolute when one knows of the height of one's eyes and the orientation of the ground plane. Contrary to the previous cues, the reliability of the height in the visual field decreases with distance (Figure 2.11(d)).

Aerial perspective Aerial perspective refers to the filtering produced by the atmosphere when distance increases. Objects in the distance get bluer and decrease in contrast with the distance (Figure 2.11(e)). This is mainly effective at large distances.

Motion parallax and motion perspective When the observer and the fixated object are in movement relative to one another, the motion parallax, that is the relative movement of the projection of the object on the observer's retina, gives a cue to the distance to the object. When the projection of a whole scene moves on the retina, one talks about motion perspective. Dynamic cues are powerful as they give the relative position of all objects as well as one's own velocity.

Amongst the monocular cues, the static cues correspond to the cues available in a regular 2D picture. They are illustrated in Figure 2.11.



(a) No cue.

(b) Occlusion.

(c) Relative size.

(d) Height in the visual field.

(e) Aerial perspective.

Figure 2.11: Static monocular depth cues. (a) The basic case with no depth cue. (b), (c), (d) Illustration of 3 monocular cues added to the case of figure (a). (e) © C. R. André, 2009. This picture was taken by the author in the surroundings of the village of Peyresq, in the south of France. It illustrates the cue of aerial perspective: the hills get bluer and bluer as the egocentric distance increases.

It might be surprising that linear perspective or texture gradients are not part of this list, as they seem to give important information about the visual layout. Cutting and Vishton [1995] argue that these elements are just a combination of continuous variations in relative size and density, which are described in the list.

We now turn to the binocular cues, which are available to us when each of our two eyes catch a different view of the same scene.

Accommodation and vergence Strictly speaking, accommodation is a monocular process. Still, we consider it with vergence because the two processes are physiologically coupled. The state of the eye muscles that produce vergence gives a non-optical indication of the distance to the fixated object. The combined action of these cues rapidly decreases with increasing distance. They are mostly useful below 2 m [Leibowitz et al., 1972].

Binocular disparities The fixated object has its projections in the fovea of each eye. When another object has projections that fall on corresponding points in the retinæ, the angular distance from the projections to the foveæ is the (angular) *disparity*. It is easily shown that the disparity increases with the distance to the fixation point in the physical world. Also, the sign of the disparity determines if the object is in front or behind the fixation point.

The binocular cues will be further discussed in Section 4.2.2.

2.3 Overview of auditory-visual perception

We have seen so far that both vision and audition can potentially give us information on the spatial location of an object. This information therefore has to be integrated in the brain to form the percept of a single audiovisual object. The material in this section is largely inspired from [Kohlrausch and van de Par, 2005]. To investigate how normal human subjects integrate the information coming from different sensory sources, traditional experiments present the subjects with a situation in which there is a discrepancy between two or more sensory modalities. This discrepancy result in an intersensory bias. This bias can sometimes rule out one modality in favor of the other, a phenomenon called *capture*. More often, however, the overall percept is an intermediary percept, resulting from multisensory integration. The amount of bias, whether it be 100% as in capture, or less, depends on the stimuli and the experimental design.

Amongst the stimulus variables, Radeau and Bertelson [1977] distinguish between structural factors and cognitive factors. The structural factors relate to the way the stimuli are delivered. This includes the magnitude of the intersensory discrepancy and the relative timing between the stimuli. The cognitive factors include the subject's awareness of the discrepancy, the “unity” assumption, which may convince the subject that there is a single physical event, and the compellingness of the situation, when more cues are redundant. Amongst the response variables, the type of response and the timing given to the subject to answer influence the most the intersensory bias.

Concerning the integration of the multisensory information, Welch and Warren [1980] introduced three potential integration schemes:

- the modality precision hypothesis, which states that the bias will be towards the modality which is the most precise. This hypothesis is supported by the phenomenon of visual capture of sound localization, the ventriloquism effect, discussed in Section 2.3.2.
- the directed attention hypothesis, which explains the bias by a different allocation of attention. The dominant modality is the one that receives the most attention.
- the modality appropriateness hypothesis, which states that different modalities are better suited for different tasks. For example, a spatial judgment will favor vision because vision outperforms the other modalities in this task. With this hypothesis, the precision of a modality is merely a consequence of the superiority of a certain modality at a certain task.

However, as Ernst and Bülthoff [2004] point out, this nomenclature is confusing. It is not the modality itself which is more precise or appropriate, but the estimate that is produced from the multimodal input to the brain.

In the particular case of the auditory and visual modalities, we are already familiar with the ventriloquism effect, where vision dominates audition. Vision, however, does not always dominate the auditory-visual integration. For example, the presence of either a sound or an image can improve the visual or auditory detection threshold, respectively. For instance, the ability of an irrelevant light stimulus to increase the sound detection threshold has been shown with the combination of an illuminated LED in spatial and temporal coincidence with a 200 ms broadband auditory noise at a near-threshold level [Lovelace et al., 2003]. This is also true for a sub-threshold masked visual signal, which can be more easily detected when presented in combination with a simultaneous auditory white noise burst at the same location [Frassinetti et al., 2002]. Another example is the presence of an accompanying picture in the task of identifying a sound source. When the sound and the picture are congruent (the picture represents the agent or object producing the sound) both the mean percentage of correct identification and the reaction times are better, compared to the case when the sound and the picture are incongruent [Bouchara et al., 2010].

2.3.1 Auditory-visual temporal perception

If either the modality precision hypothesis or the modality appropriateness hypothesis of multimodal integration is correct, then temporal perception is an example where the auditory modality should dominate. Recanzone [2003] has shown that the presentation of two consecutive series of four beeps and flashes, where only the second series of beeps is presented at a higher frequency, results in the perception of the second series of flashes at a higher frequency. Thus the visual rate perception is captured by an auditory distractor. This effect is not dependent on the spatial alignment of the two stimuli or the intensity or bandwidth of the auditory signal. It is, however, strongly dependent on the difference of rates between the two presentations.

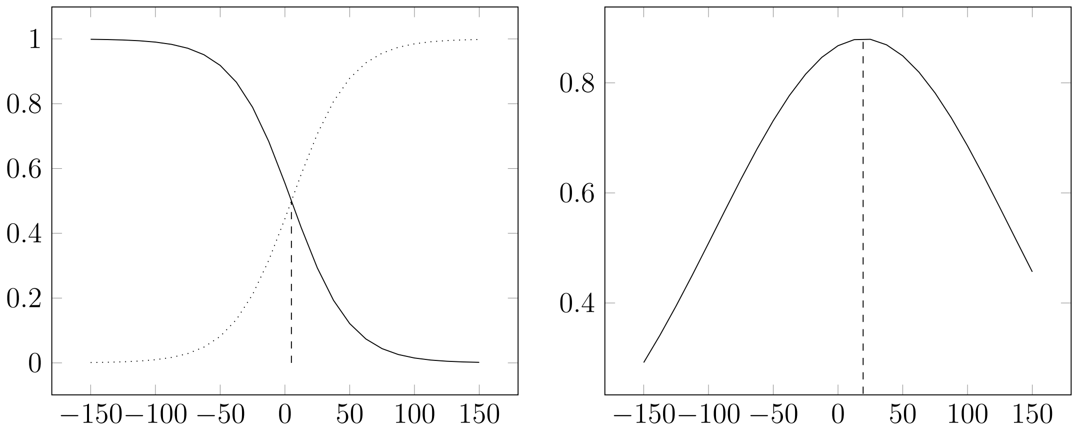
The perception of auditory-visual simultaneity is another important topic for research. Depending on the stimuli and the experimental method, the time-window of simultaneity varies. In the following, a negative time delay means that the sound precedes the image. At the 75% level, Hirsh and Sherrick Jr. [1961] measured a window at $[-25\text{ ms}, 25\text{ ms}]$ with a point of subjective equivalence (PSE) at 5 ms. This is shown in Figure 2.12(a). Lewald and Guski [2003] asked subjects to judge the synchrony of an AV stimulus on a scale from 1 to 9. The judgement was highest in the range $[-50\text{ ms}, 100\text{ ms}]$ with the maximum at 50 ms. Similarly, [Zampini et al., 2005] obtained a gaussian fit with mean $= 19.4$ ms and stdev $= 114$ ms. This fit is shown in Figure 2.12(b). Based on the quality judgment of an audiovisual program, the detectability range reported in the ITU recommendation BT1359 is $[-45\text{ ms}, 125\text{ ms}]$. These differences can be explained, partly by the difference in stimuli, but also by the different criteria used to define the thresholds.

2.3.2 Auditory-visual spatial perception

2.3.2.1 Ventriloquism in azimuth

When people are presented with a synchronous but spatially mismatched auditory-visual stimulus, they tend to perceive the sound coming from closer to the location of the visual stimulus, the so-called ventriloquism effect [Thurlow and Jack, 1973]. This effect decreases with increasing angular difference between the positions of the auditory and visual sources [Jackson, 1953].

Experiments previously conducted in laboratory conditions used an audiovisual stimulus consisting of a simultaneous pair of brief, simple, and arbitrary stimuli, such as an auditory beep, and a visual flash. In this case, Alais and Burr [2004] showed that a statistically optimal model approximates well the mechanism of bimodal integration. If each sensory estimate $\hat{S}_i$ is unbiased but corrupted by



(a) Precedence judgment.

(b) Simultaneity judgment.

Figure 2.12: Response pattern in temporal judgment tasks. (a) Judgment about whether "audio comes first" (continuous line) or "video comes first" (dotted line). Data from [Hirsh and Sherrick Jr., 1961]. (b) Judgment about the simultaneity of the stimuli. Data from [Zampini et al., 2005].

a gaussian white noise with variance σ_i^2, then the maximum-likelihood estimate (MLE) of the integrated perception is

$$\hat{S} = \sum_i w_i \hat{S}_i \text{ with } w_i = \frac{1/\sigma_i^2}{\sum_j 1/\sigma_j^2}. \tag{2.2}$$

Thus, the variance of an auditory-visual percept is given by

$$\sigma_{AV}^2 = \frac{\sigma_A^2 \sigma_V^2}{\sigma_A^2 + \sigma_V^2} \leq \min\left(\sigma_A^2, \sigma_V^2\right) \tag{2.3}$$

provided that the sensory noises are independent [Ernst and Banks, 2002]. In another experiment, Battaglia et al. [2003] had to include a Bayesian prior, thereby generalizing the MLE model, to accurately model the data from their auditory-visual localization task. In the experiment reported in Chapter 8, the additional auditory ambient noise serves as a way to decrease the reliability of the auditory spatial information through an increase of σ_A^2. The sensory integration should therefore rely more on the visual stimulus, and the stimulus integration should be more resilient to the angular error between the sound and the image.

The magnitude of the auditory-visual integration has been found to depend on both spatial relations and temporal relations of the unimodal stimuli. The auditory-visual window of integration of arbitrary stimuli extends up to about

100 ms in time and 3° in space [Lewald et al., 2001]. It is centered around 0° in space and about 50 ms in time, when the audio arrives after the image [Lewald and Guski, 2003; Slutsky and Recanzone, 2001]. Slutsky and Recanzone [2001] have shown that the effect of a temporal disparity on bimodal integration is greatest when the spatial (angular) error between the sound and the image is below the (spatial) threshold of integration. However, the effect of a temporal disparity is not significant below a 50 ms time delay.

When the stimuli are more natural, i.e. carry more information, such as for a speaking character, then the "unity assumption" must be taken into account. The unity assumption arises from properties shared by the unimodal stimuli such as space location, temporal rate, size, shape, ... [Welch, 1999]. The more numerous the shared properties, the stronger the association of the two stimuli. Conversely, if more properties give conflicting cues, the integration is lessened.

Therefore, when more natural stimuli are used, such that the unity assumption holds, the bimodal integration is maintained at much larger angles of discrepancy than those obtained with arbitrary stimuli. Simply by letting participants assume that the arbitrary stimuli had a common cause, Lewald and Guski [2003] increased the spatial window to about 12°. The temporal window, also, can be enlarged. Using, a speech stimulus, van Wassenhove et al. [2007] obtained a 200 ms time window (see also Section 2.3.3).

2.3.2.2 Ventriloquism in distance

The question of whether a similar ventriloquism effect exists in auditory distance perception is still open. Results in existing studies are contradictory and seem to heavily depend on the experimental design.

The first indication on this topic is that provided by Gardner [1968], in an anechoic environment. Participants faced a loudspeaker at eye-level. Directly behind this loudspeaker were located several other loudspeakers at increasing distances, in such a way that only the closest loudspeaker could be seen. When a speech signal was played by the loudspeaker farthest from the participant (at about 9 m) at a comfortable level (65 dB(B) measured at the participant's position), the participants chose the closest loudspeaker as the source, without exception. Gardner called this the proximity-image effect. Note that this effect only worked when the sound was played at a constant, reasonable level [Gardner, 1969]. When several levels were tested successively, the chosen apparent source moved away from the participant with a decreasing sound level. Because these experiments were conducted in an anechoic environment, the participants missed an important cue for distance perception, namely the direct-to-reverberant ratio.

Zahorik [2001] reexamined the results from [Gardner, 1968] using a similar apparatus and a similar method, but in a semi-reverberant environment. Half the subjects did not see the apparatus. The other half saw the layout before the experiment. When participants were blindfolded, the auditory distance estimation was in alignment with the literature ($k = 0.92$, and $a = 0.66$). When additional visual cues were available, the distance estimation was better ($k = 0.90$, and $a = 0.78$) and the variability of the judgments was lowered. These results therefore contradicts Gardner's proximity-image capture hypothesis. In addition, Zahorik did not find any evidence supporting the visual capture of auditory distance in his data.

Calcagno et al. [2012] again reexamined the results from [Gardner, 1968] in a darkened, semi-reverberant room. Participants therefore had no information on the layout of the room and could not see the loudspeaker during the experiment. Instead, the visual targets were LEDs aligned on a line parallel to the loudspeaker array. In addition, only one mobile loudspeaker was used to prevent shadow filtering of the signal coming from loudspeakers behind the closest one. The mobile loudspeaker was moved in distance between trials, while two other loudspeakers played a masking signal. Highly accurate distance estimation ($k = 1.14$, and $a = 0.89$) was achieved when participants could see the test room for 5 min with lights on before the actual experiment, in the dark, suggesting that visual cues affect auditory distance perception. Some authors suggest the behavior of auditory distance perception at high distances (compression and underestimation) lies in the breakdown of the direct-to-reverberant ratio cue, when the soundfield is entirely dominated by the reverberant portion. The results in [Calcagno et al., 2012] suggest that this behavior can be counteracted by visual cues (i.e. non-auditory cues), even before the experiment.

Recent experiments investigated directly the question of ventriloquism by presenting stimuli in unimodal conditions, bimodal congruent conditions, and bimodal incongruent conditions.

Bowen et al. [2011] investigated the possible ventriloquism in both azimuth and depth. They presented visual stimuli on an stereoscopic 3D (s-3D) screen with active glasses, and the auditory stimuli with binaural sound, recorded directly on the participants. They presented the visual stimuli on three lines at 0° and $\pm 8.78°$ azimuth. Three depth planes were tested, at 55, 69, and 83 cm (the screen plane being at 70 cm). Although they only included results from four participants, the bias of the auditory perception towards the visual stimulus was of the same order of magnitude, reaching 84% in azimuth and 72% in depth. In addition, the Bayesian weights corresponding to Equation (2.2) showed that σ_V was approximately equal to σ_{AV} in both cases, indicating that the Bayesian framework was equally applicable in azimuth and depth.

Turner et al. [2011] presented the image of a mobile phone on a 46-inch 3DTV. The mobile phone appeared to ring from one of two sound positions (hidden loudspeakers). One sound source was congruent to the visual object, and the other was 25 cm in front of the visual object. Fifteen participants judged in a two-alternative forced choice (2-AFC[*]) experiment in which presentation the phone was displayed closer, while in reality, only the sound position varied between presentations. The results showed that participants did perceive the phone to be closer.

Cote et al. [2012] presented congruent and incongruent stimuli on a 2.4×1.8 m^2 s-3D screen, with binaural sound. Their results suggest that, when the sound source is located behind the perceived visual object, the auditory-visual source is located at the position of the perceived visual object. The perceived auditory-visual object is slightly (but significantly) brought to the front when the sound source is in front of the perceived visual object. From the available data in the article, we computed that the bias in depth towards the visual stimulus was around 50%.

Corrigan et al. [2013] presented congruent and incongruent stimuli on a 46-inch 3DTV. The visual stimuli were an s-3D picture of a loudspeaker and an s-3D video of a talking person. The corresponding auditory stimuli were dry recordings convolved with B-format room impulse responses, presented binaurally. Participants judged whether the auditory stimulus was in front, at the same depth, or behind the visual stimulus. Three different visual distances (2, 4, and 8 m) were tested, and fourteen auditory distances. When the visual stimulus was at 2 m (subsequently 4 and 8 m) from the viewer, the auditory stimulus was perceived as congruent at distances between 1 and 2 m (between 3 and 6 m, and between 6 and 11 m, respectively, for greater distances). However, we noticed the presence of window violation in the visual stimulus. Therefore, the depth estimation might be biased. Window violation occurs when an object appears in one image in the s-3D pair, but not in the other. The fact that the edge of the picture cuts off the object is interpreted as an occlusion by the brain, which is in conflict with the

[*] In a 2-AFC experiment, participants are presented consecutively with two different stimuli. They have to decide which of the two corresponds better to a given criteria.

presence of the object in the other retinal image.

2.3.3 Auditory-visual perception of speech

The ability to see the lips of a speaker while hearing the speaker's voice has been shown to improve the sensitivity to acoustic information in the presence of noise, *i.e.* the ability to detect speech in noise [Grant, 2001]. It has also been shown to improve intelligibility, *i.e.* the ability to understand what is said, by comparison to auditory-only situations [Schwartz et al., 2004; Sumby and Pollack, 1954].

In order to obtain this increase in intelligibility, the delay between the auditory stimulus and the visual stimulus has to remain close to zero. McGrath and Summerfield [1985] measured the just-noticeable audiovisual delay with a speech-like stimulus (a Lissajou curve resembling lips and a 120 Hz triangular wave attenuated at onset and offset). The just-noticeable delay was 79 ms when the audio came first, and 138 ms when the video came first. By comparison, the ITU-T recommendation BT1359 concerning television programs suggests that the delay should be kept below 90 ms (audio first) or 185 ms (video first).

The discrepant presentation of speech and lip movement can result in the McGurk effect, a famous auditory-visual speech illusion [McGurk and MacDonald, 1976]: when presented with a visual stimulus consisting of a talking face repeating the syllable /ga/ while listening to an audio track consisting in a voice repeating the syllable /ba/, subjects reported hearing the syllable /da/. Of course, this effect disappeared in unimodal trials.

2.3.4 Auditory-visual interaction in presence of ambiguity

Sekuler et al. [1997] presented to participants a visually ambiguous sequence. Two visually identical particles in motion move steadily towards each other, coincide on the screen for a variable period of time, and then regain their initial position (Figure 2.13). This sequence can be interpreted in two ways: (1) each particle returns to its initial position, which is called the *bouncing* scenario, and (2) each particle arrives at the other side of the picture, which is called the *streaming* scenario. The addition of a click noise when the particles coincide increases significantly the perception of the sequence as a bouncing scenario.

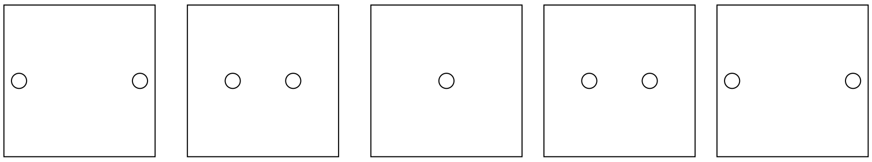


Figure 2.13: Illustration of an ambiguous visual sequence (from left to right). Two particles progress towards one another. In the middle picture, they coincide. In the absence of any other cue, one cannot separate the case where each particle regains its initial position (*bouncing*) from the case where each particle arrives at the other side of the picture (*streaming*). The addition of a click noise when the particles coincide increases significantly the perception of the sequence as a bouncing scenario [Sekuler et al., 1997].

2.3.5 Auditory-visual illusion

An auditory-visual illusion that cannot be placed in any of the previous sections is the multiple flash illusion [Shams et al., 2002]. When a single flash of light is accompanied by several auditory beeps, subjects perceive the single flash as a sequence of flashes. This phenomenon is particular in that it occurs for an unambiguous visual stimulus. Researchers also tried to present multiple flashes accompanied by one auditory beep [Shimojo and Shams, 2001]. This does not result in any illusion.

2.3.6 Perception of audiovisual quality

A temporal asynchrony [ITU R. BT1359] or a spatial incongruence [Komiyama, 1989] can degrade the perceptual quality of an audiovisual television program. Beerends and De Caluwe [1999] investigated whether viewers could effectively discard the sound when making a judgment about the picture in an audiovisual content, and, conversely, whether they could discard the picture when making a judgment on sound quality. Their results show that this is not the case: the perceived quality of sound is influenced by the quality of the picture (1.2 points on a 9-point scale), and, to a lesser degree, the perceived quality of the picture is influenced by the quality of sound (0.2 points on a 9-point scale). The correlation between the perceived visual quality and the perceived audiovisual quality was 0.92. The correlation between the perceived audio quality and the perceived audiovisual quality was significantly lower, at only 0.35.

More recently, Gaston et al. [2010] showed with a same/different paradigm that the addition of a degradation of the audio track could increase or decrease the sensitivity to a degradation of the video track. For example, the presence of all three audio degradations tested (high frequency boost, high frequency cut, and lowered MP3 bitrate) lowered the sensitivity of the viewers to an increase in brightness.

Bibliography

Alais, D., Burr, D., Feb. 2004. The ventriloquist effect results from near-optimal bimodal integration. Current Biology 14 (3), 257–262.
http://dx.doi.org/10.1016/j.cub.2004.01.029 30

Asano, F., Suzuki, Y., Sone, T., Jul. 1990. Role of spectral cues in median plane localization. J. Acoust. Soc. Am. 88 (1), 159–168.
http://dx.doi.org/10.1121/1.399963 16

Atchison, D. A., Jul. 1995. Accommodation and presbyopia. Ophthalmic and Physiological Optics 15 (4), 255–272.
http://dx.doi.org/10.1016/0275-5408(95)00020-E 21

Battaglia, P. W., Jacobs, R. A., Aslin, R. N., Jul. 2003. Bayesian integration of visual and auditory signals for spatial localization. Journal of the Optical Society of America 20 (7), 1391–1397. 31

Beerends, J. G., De Caluwe, F. E., 1999. The influence of video quality on perceived audio quality and vice versa. J. Audio Eng. Soc 47 (5), 355–362.
http://www.aes.org/e-lib/browse.cfm?elib=12105 36

Begault, D. R., 1994. 3-D sound for virtual reality and multimedia. Academic Press Professional, Inc. 15

Blauert, J., 1969. Sound localization in the median plane. Acustica 22 (4), 205–213. 15

Blauert, J., 1997. Spatial hearing: the psychophysics of human sound localization. MIT Press. 14

Bouchara, T., Giordano, B. L., Frissen, I., Katz, B. F. G., Guastavino, C., May 2010. Effect of signal-to-noise ratio and visual context on environmental sound identification. In: Audio Eng. Soc. Conv. 128.
http://www.aes.org/e-lib/browse.cfm?elib=15446 29

Bowen, A. L., Ramachandran, R., Muday, J. A., Schirillo, J. A., Oct. 2011. Visual signals bias auditory targets in azimuth and depth. Experimental Brain Research 214 (3), 403–414.
http://dx.doi.org/10.1007/s00221-011-2838-1 33

Buell, T. N., Griffin, S. J., Bernstein, L. R., 2008. Listeners' sensitivity to "onset/offset" and "ongoing" interaural delays in high-frequency, sinusoidally amplitude-modulated tones. J. Acoust. Soc. Am. 123 (1), 279–294.
http://dx.doi.org/10.1121/1.2816399 15

Calcagno, E. R., Abregú, E. L., Eguía, M. C., Vergara, R., 2012. The role of vision in auditory distance perception. Perception 41 (2), 175 – 192.
http://dx.doi.org/10.1068/p7153 33

Carlile, S., Delaney, S., Corderoy, A., Feb. 1999. The localisation of spectrally restricted sounds by human listeners. Hearing Research 128 (1–2), 175–189.
http://dx.doi.org/10.1016/S0378-5955(98)00205-6 16

Chittka, L., Brockmann, A., Apr. 2005. Perception space - the final frontier. PLoS Biol 3 (4), e137.
http://dx.doi.org/10.1371/journal.pbio.0030137 12

Coleman, P. D., 1962. Failure to localize the source distance of an unfamiliar sound. J. Acoust. Soc. Am. 34 (3), 345–346.
http://dx.doi.org/10.1121/1.1928121 19

Coleman, P. D., 1963. An analysis of cues to auditory depth perception in free space. Psychological Bulletin 60 (3), 302–315.
http://dx.doi.org/10.1037/h0045716 17, 18

Coleman, P. D., 1968. Dual role of frequency spectrum in determination of auditory distance. J. Acoust. Soc. Am. 44 (2), 631–632.
http://dx.doi.org/10.1121/1.1911132 18

Cook, M., Jan. 1978. The judgment of distance on a plane surface. Perception & Psychophysics 23 (1), 85–90.
http://dx.doi.org/10.3758/BF03214300 25

Corrigan, D., Gorzel, M., Squires, J., Boland, F., Mar. 2013. Depth perception of audio sources in stereo 3D environments. In: Woods, A. J., Holliman, N. S.,

Favalora, G. E. (Eds.), Proc. SPIE 8648. Burlingame, CA, p. 864816.
http://dx.doi.org/10.1117/12.2000713 34

Cote, N., Koehl, V., Paquier, M., Apr. 2012. Ventriloquism effect on distance auditory cues. In: d'Acoustique, S. F. (Ed.), Acoustics 2012 Nantes. Nantes, France, SP-G01: Sound perception SP-G01: Sound perception.
http://hal.archives-ouvertes.fr/hal-00810668 34

Cutting, J. E., Vishton, P. M., 1995. Perceiving layout and knowing distances: the integration, relative potency and contextual use of different information about depth. In: Epstein, W., Rogers, S. (Eds.), Handbook of perception and Cognition, Academic Press Edition. Vol. 5: Perception of Space and Motion. San Diego, CA, pp. 69–117. 25, 26, 27

Da Silva, J. A., Apr. 1985. Scales for perceived egocentric distance in a large open field: Comparison of three psychophysical methods. The American Journal of Psychology 98 (1), 119–144.
http://dx.doi.org/10.2307/1422771 25

Daum, S. O., Hecht, H., Jul. 2009. Distance estimation in vista space. Attention, Perception, & Psychophysics 71 (5), 1127–1137.
http://dx.doi.org/10.3758/APP.71.5.1127 25

Ernst, M. O., Banks, M. S., Jan. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415 (6870), 429–433.
http://dx.doi.org/10.1038/415429a 31

Ernst, M. O., Bülthoff, H. H., Apr. 2004. Merging the senses into a robust percept. Trends in Cognitive Sciences 8 (4), 162–169.
http://dx.doi.org/10.1016/j.tics.2004.02.002 29

Fincham, E. F., Walton, J., Jun. 1957. The reciprocal actions of accommodation and convergence. The Journal of Physiology 137 (3), 488–508.
http://jp.physoc.org/content/137/3/488 21

Flannery, R., Butler, R. A., Sep. 1981. Spectral cues provided by the pinna for monaural localization in the horizontal plane. Perception & Psychophysics 29 (5), 438–444.
http://dx.doi.org/10.3758/BF03207357 15

Frassinetti, F., Bolognini, N., Làdavas, E., Dec. 2002. Enhancement of visual perception by crossmodal visuo-auditory interaction. Experimental Brain Research 147 (3), 332–343.
http://dx.doi.org/10.1007/s00221-002-1262-y 29

Fukusima, S. S., Loomis, J. M., Da Silva, J. A., 1997. Visual perception of egocentric distance as assessed by triangulation. Journal of Experimental Psychology: Human Perception and Performance 23 (1), 86–100.
http://dx.doi.org/10.1037//0096-1523.23.1.86 25

Gamble, E. A., 1909. Intensity as a criterion in estimating the distance of sounds. Psychological Review 16 (6), 416–426.
http://dx.doi.org/10.1037/h0073666 17

Gardner, M. B., 1968. Proximity image effect in sound localization. J. Acoust. Soc. Am. 43 (1), 163–163.
http://link.aip.org/link/?JAS/43/163/1 32, 33

Gardner, M. B., 1969. Distance estimation of 0° or apparent 0°-oriented speech signals in anechoic space. J. Acoust. Soc. Am. 45 (1), 47–53.
http://dx.doi.org/10.1121/1.1911372 19, 32

Gardner, M. B., 1973. Some monaural and binaural facets of median plane localization. J. Acoust. Soc. Am. 54 (6), 1489–1495.
http://dx.doi.org/10.1121/1.1914447 16

Gaston, L., Boley, J., Selter, S., Ratterman, J., May 2010. The influence of individual audio impairments on perceived video quality. In: Audio Engineering Society Convention 128.
http://www.aes.org/e-lib/browse.cfm?elib=15447 36

Grant, K. W., 2001. The effect of speechreading on masked detection thresholds for filtered speech. The Journal of the Acoustical Society of America 109 (5), 2272–2275.
http://dx.doi.org/10.1121/1.1362687 35

Hirsh, I. J., Sherrick Jr., C. E., 1961. Perceived order in different sense modalities. Journal of Experimental Psychology 62 (5), 423–432.
http://dx.doi.org/10.1037/h0045283 30, 31

Interrante, V., Ries, B., Anderson, L., 2006. Distance perception in immersive virtual environments, revisited. In: Proceedings of the IEEE conference on Virtual Reality. VR 06. IEEE Computer Society, Washington, DC, USA, p. 3–10.
http://dx.doi.org/10.1109/VR.2006.52 25

Interrante, V., Ries, B., Lindquist, J., Kaeding, M., Anderson, L., Apr. 2008. Elucidating factors that can facilitate veridical spatial perception in immersive virtual environments. Presence: Teleoperators & Virtual Environments 17 (2), 176–198.
http://dx.doi.org/10.1162/pres.17.2.176 25

Jackson, C. V., 1953. Visual factors in auditory localization. Quarterly Journal of Experimental Psychology 5 (2), 52–65.
http://dx.doi.org/10.1080/17470215308416626 30

Kennel, G., Jul. 2012. Color and Mastering for Digital Cinema. CRC Press. 21

King, R. B., Oldfield, S. R., 1997. The impact of signal bandwidth on auditory localization: Implications for the design of three-dimensional audio displays. Human Factors 39 (2), 287. 16

Kohlrausch, A., van de Par, S., Jan. 2005. Audio-visual interaction in the context of multi-media applications. In: Blauert, J. (Ed.), Communication Acoustics. Springer Berlin Heidelberg, pp. 109–138. 28

Komiyama, S., 1989. Subjective evaluation of angular displacement between picture and sound directions for HDTV sound systems. J. Audio Eng. Soc. 37 (4), 210–214.
http://www.aes.org/e-lib/browse.cfm?elib=6094 36

Langendijk, E. H. A., Bronkhorst, A. W., 2002. Contribution of spectral cues to human sound localization. J. Acoust. Soc. Am. 112 (4), 1583–1596.
http://dx.doi.org/10.1121/1.1501901 16

Leibowitz, H. W., Shiina, K., Hennessy, R. T., Nov. 1972. Oculomotor adjustments and size constancy. Perception & Psychophysics 12 (6), 497–500.
http://dx.doi.org/10.3758/BF03210943 28

Lewald, J., Ehrenstein, W. H., Guski, R., Jun. 2001. Spatio-temporal constraints for auditory-visual integration. Behavioural Brain Research 121 (1-2), 69–79.
http://dx.doi.org/10.1016/S0166-4328(00)00386-7 32

Lewald, J., Guski, R., May 2003. Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. Cognitive Brain Research 16 (3), 468–478.
http://dx.doi.org/10.1016/S0926-6410(03)00074-0 30, 32

Little, A. D., Mershon, D. H., Cox, P. H., 1992. Spectral content as a cue to perceived auditory distance. Perception 21 (3), 405 – 416.
http://dx.doi.org/10.1068/p210405 18

Loomis, J. M., Da Silva, J. A., Fujita, N., Fukusima, S. S., Nov. 1992. Visual space perception and visually directed action. Journal of Experimental Psychology. Human Perception and Performance 18 (4), 906–921. 25

Lovelace, C. T., Stein, B. E., Wallace, M. T., Jul. 2003. An irrelevant light enhances auditory detection in humans: a psychophysical analysis of multisensory integration in stimulus detection. Cognitive Brain Research 17 (2), 447–453.
http://dx.doi.org/10.1016/S0926-6410(03)00160-5 29

McFadden, D., Pasanen, E. G., 1976. Lateralization at high frequencies based on interaural time differences. J. Acoust. Soc. Am. 59 (3), 634–639.
http://dx.doi.org/10.1121/1.380913 14

McGrath, M., Summerfield, Q., 1985. Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. The Journal of the Acoustical Society of America 77 (2), 678–685.
http://dx.doi.org/10.1121/1.392336 35

McGurk, H., MacDonald, J., Dec. 1976. Hearing lips and seeing voices. Nature 264 (5588), 746–748.
http://dx.doi.org/10.1038/264746a0 35

Mershon, D. H., King, L. E., Nov. 1975. Intensity and reverberation as factors in the auditory perception of egocentric distance. Perception & Psychophysics 18 (6), 409–415.
http://dx.doi.org/10.3758/BF03204113 18, 19

Middlebrooks, J. C., Makous, J. C., Green, D. M., 1989. Directional sensitivity of sound-pressure levels in the human ear canal. J. Acoust. Soc. Am. 86 (1), 89–108.
http://dx.doi.org/10.1121/1.398224 12

Musicant, A. D., Butler, R. A., 1984. The influence of pinnae-based spectral cues on sound localization. J. Acoust. Soc. Am. 75 (4), 1195–1200.
http://dx.doi.org/10.1121/1.390770 15

Møller, H., Sørensen, M. F., Hammershøi, D., Jensen, C. B., May 1995. Head-related transfer functions of human subjects. J. Audio Eng. Soc. 43 (5), 300–321.
http://www.aes.org/e-lib/browse.cfm?elib=7949 17

Ono, H., Nov. 1979. Axiomatic summary and deductions from Hering's principles of visual direction. Perception & Psychophysics 25 (6), 473–477.
http://dx.doi.org/10.3758/BF03213825 23

Osterberg, G., 1935. Topography of the layer of rods and cones in the human retina. Acta Ophthalmologica 13 (Supplement 6), 1–97.
http://dx.doi.org/10.1001/jama.1937.02780030070033 22

Perrett, S., Noble, W., Jan. 1997a. The contribution of head motion cues to localization of low-pass noise. Perception & Psychophysics 59 (7), 1018–1026.
http://dx.doi.org/10.3758/BF03205517 15

Perrett, S., Noble, W., 1997b. The effect of head rotations on vertical plane sound localization. J. Acoust. Soc. Am. 102 (4), 2325–2332.
http://dx.doi.org/10.1121/1.419642 17

Radeau, M., Bertelson, P., 1977. Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. Attention, Perception, & Psychophysics 22 (2), 137–146.
http://dx.doi.org/10.3758/BF03198746 28

Rayleigh, L., 1907. XII. On our perception of sound direction. Philosophical Magazine Series 6 13 (74), 214–232.
http://dx.doi.org/10.1080/14786440709463595 15

Recanzone, G. H., Jan. 2003. Auditory influences on visual temporal rate perception. Journal of Neurophysiology 89 (2), 1078–1093.
http://dx.doi.org/10.1152/jn.00706.2002 30

Roffler, S. K., Butler, R. A., 1968. Factors that influence the localization of sound in the vertical plane. J. Acoust. Soc. Am. 43 (6), 1255–1259.
http://dx.doi.org/10.1121/1.1910976 16

Roorda, A., 2002. Human Visual System – Image Formation. In: Hornak, J. P. (Ed.), Encyclopedia of Imaging Science and Technology, John Wiley & Sons Edition. Vol. 1. New York, pp. 539–557. 21

Schwartz, J.-L., Berthommier, F., Savariaux, C., Sep. 2004. Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition 93 (2), B69–B78.
`http://dx.doi.org/10.1016/j.cognition.2004.01.006` 35

Sekuler, R., Sekuler, A. B., Lau, R., Jan. 1997. Sound alters visual motion perception. Nature 385 (6614), 308–308.
`http://dx.doi.org/10.1038/385308a0` 35, 36

Shams, L., Kamitani, Y., Shimojo, S., Jun. 2002. Visual illusion induced by sound. Cognitive Brain Research 14 (1), 147–152.
`http://dx.doi.org/10.1016/S0926-6410(02)00069-1` 36

Shaw, B. K., McGowan, R. S., Turvey, M., 1991. An acoustic variable specifying time-to-contact. Ecological Psychology 3 (3), 253–261.
`http://dx.doi.org/10.1207/s15326969eco0303_4` 18

Shimojo, S., Shams, L., Aug. 2001. Sensory modalities are not separate modalities: plasticity and interactions. Current Opinion in Neurobiology 11 (4), 505–509.
`http://dx.doi.org/10.1016/S0959-4388(00)00241-5` 36

Slutsky, D. A., Recanzone, G. H., Jan. 2001. Temporal and spatial dependency of the ventriloquism effect. Neuroreport 12 (1), 7–10. 32

Spector, R. H., 1990. Visual fields. In: Walker, H. K., Hall, W. D., Hurst, J. W. (Eds.), Clinical Methods: The History, Physical, and Laboratory Examinations, 3rd Edition. Butterworths, Boston.
`http://www.ncbi.nlm.nih.gov/books/NBK220/` 20

Speigle, J. M., Loomis, J. M., 1993. Auditory distance perception by translating observers. In: Virtual Reality, 1993. Proceedings., IEEE 1993 Symposium on Research Frontiers in. pp. 92–99.
`http://dx.doi.org/10.1109/VRAIS.1993.378257` 18

Stern, R. M., Brown, G. J., Wang, D., Oct. 2006. Binaural sound localization. In: Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley-IEEE Press, New York, p. 395. 14

Stewart, G. W., May 1920a. The function of intensity and phase in the binaural location of pure tones. I. Physical Review 15 (5), 425–431.
http://dx.doi.org/10.1103/PhysRev.15.425 14

Stewart, G. W., May 1920b. The function of intensity and phase in the binaural location of pure tones. II. Physical Review 15 (5), 432–445.
http://dx.doi.org/10.1103/PhysRev.15.432 15

Stockman, A., Sharpe, L. T., 2000. The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. Vision research 40 (13), 1711–1737. 22

Sumby, W. H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. The Journal of the Acoustical Society of America 26 (2), 212–215.
http://dx.doi.org/10.1121/1.1907309 35

Thurlow, W. R., Jack, C. E., Jun. 1973. Certain determinants of the "ventriloquism effect". Percept. Motor Skill 36, 1171–1184.
http://dx.doi.org/10.2466/pms.1973.36.3c.1171 19, 30

Thurlow, W. R., Runge, P. S., 1967. Effect of induced head movements on localization of direction of sounds. J. Acoust. Soc. Am. 42 (2), 480–488.
http://dx.doi.org/10.1121/1.1910604 15, 17

Tobias, J. V., Schubert, E. D., 1959. Effective onset duration of auditory stimuli. J. Acoust. Soc. Am. 31 (12), 1595–1605.
http://dx.doi.org/10.1121/1.1907665 15

Turner, A., Berry, J., Holliman, N., Feb. 2011. Can the perception of depth in stereoscopic images be influenced by 3D sound? Proceedings of SPIE 7863, 786307.
http://dx.doi.org/10.1117/12.871960 34

van Wassenhove, V., Grant, K. W., Poeppel, D., Jan. 2007. Temporal window of integration in auditory-visual speech perception. Neuropsychologia 45 (3), 598–607.
http://dx.doi.org/10.1016/j.neuropsychologia.2006.01.001 32

Vishwanath, D., Blaser, E., Aug. 2010. Retinal blur and the perception of egocentric distance. Journal of Vision 10 (10).
http://dx.doi.org/10.1167/10.10.26 24

von Noorden, G. K., Campos, E. C., 2002. Binocular vision and space perception. In: Binocular Vision and Ocular Motility, 6th Edition. Mosby, pp. 7–37. 23

Wandell, B. A., 1995. Foundations of vision. Sinauer Associates. 22

Welch, R. B., 1999. Meaning, attention, and the "unity assumption" in the intersensory bias of spatial and temporal perceptions. In: Aschersleben, G., Bachmann, T., Müsseler, J. (Eds.), Cognitive Contributions to the Perception of Spatial and Temporal Events. Vol. 129 of Advances in Psychology. North-Holland, pp. 371–387. 32

Welch, R. B., Warren, D. H., 1980. Immediate perceptual response to intersensory discrepancy. Psychological Bulletin 88 (3), 638–667.
http://dx.doi.org/10.1037/0033-2909.88.3.638 28

Wiener, F. M., Ross, D. A., 1946. The pressure distribution in the auditory canal in a progressive sound field. J. Acoust. Soc. Am. 18 (2), 401–408.
http://dx.doi.org/10.1121/1.1916378 12

Wightman, F. L., Kistler, D. J., 1999. Resolution of front–back ambiguity in spatial hearing by listener and source movement. J. Acoust. Soc. Am. 105 (5), 2841–2853.
http://dx.doi.org/10.1121/1.426899 15

Wikipedia, 2013a. Auditory system — Wikipedia, The Free Encyclopedia. [Online; accessed 3-April-2013].
http://en.wikipedia.org/w/index.php?title=Auditory_system&oldid=547093343 12

Wikipedia, 2013b. Eye — Wikipedia, The Free Encyclopedia. [Online; accessed 3-April-2013].
http://en.wikipedia.org/w/index.php?title=Eye&oldid=548312637 20

Willemsen, P., Gooch, A. A., Thompson, W. B., Creem-Regehr, S. H., 2008. Effects of stereo viewing conditions on distance perception in virtual environments. Presence: Teleoperators and Virtual Environments 17 (1), 91–101.
http://dx.doi.org/10.1162/pres.17.1.91 25

Zahorik, P., 2001. Estimating sound source distance with and without vision. Optometry & Vision Science 78 (5). 19, 33

Zahorik, P., Brungart, D. S., Bronkhorst, A. W., May 2005. Auditory distance perception in humans: A summary of past and present research. Acta Acust. united with Acust. 91, 409–420(12). 17

Zampini, M., Guest, S., Shore, D. I., Spence, C., Apr. 2005. Audio-visual simultaneity judgments. Perception & Psychophysics 67 (3), 531–544. http://dx.doi.org/10.3758/BF03193329 30, 31

This page intentionally left blank.

Chapter
3

Spatial sound reproduction

Highlights

✓ The reproduction of spatial sound to the moviegoer is considered.

✓ Stereo sound is discussed.

✓ A nomenclature of one and two-channel audio systems is given.

✓ A short historical review describes the technological evolution of sound systems in movie theaters.

✓ The different 3D audio technologies are discussed.

Contents

This chapter focuses on the reproduction of spatial sound to the spectator in a cinema. First, stereo sound is discussed and a nomenclature of one and two-amplifying channel audio systems is given. Then, a short historical review describes the technological evolution of sound systems in movie theaters. Finally, different 3D audio technologies are discussed.

3.1 Two-channel sound reproduction

One must understand a few concepts about current spatial sound reproduction before discussing more involved audio techniques. Most sound reproduction nowadays, at least on television and the radio, is stereophonic. This is popularly abbreviated as stereo sound. Stereo sound refers to a specific way of capturing, transmitting, and reproducing a sound to the listener, which uses two channels to convey the information. A channel is understood as "... an electric circuit carrying a current having a definite form depending upon the original sounds in the studio." (Blumlein [1936])

Section 3.1.1 discusses stereo sound. Section 3.1.2 describes other two-channels reproductions systems and gives a global nomenclature.

3.1.1 Stereo sound

When two loudspeakers are fed the same signal, possibly with a small time delay and/or level difference between the two channels, one perceives a *phantom source* through stereo sound. The time delay and/or the level difference determine the placement of the phantom source between the two loudspeakers (Figure 3.1). More precisely, in a standard setup ($\theta_0 = 30°$), for shifts inferior to about 23°, the relationship between the shift and both time delay and level difference is linear and the shift S, in degree, can be expressed as

$$S(\Delta t, \Delta L) = 44 \times \Delta t + 2.2 \times \Delta L, \tag{3.1}$$

where Δt is expressed in ms and ΔL is expressed in dB [Theile, 2000].

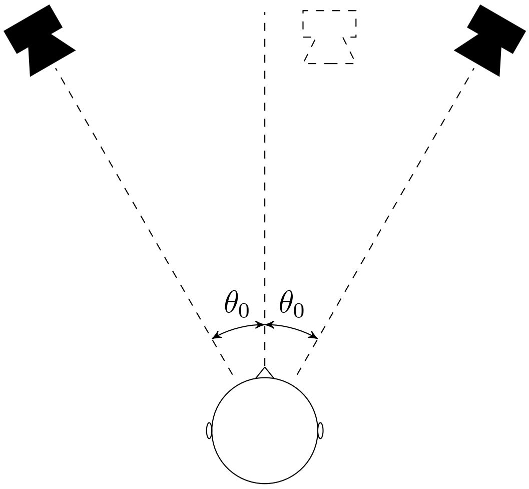


Figure 3.1: The usual stereo setup. When $\theta_0 = 30°$, the setup is compatible with the ITU standard for 5.1 sound reproduction [ITU, 2012]. A phantom source (dashed shape) is reproduced at an angle of 10° from the forward direction. After Streicher and Everest [2006].

The development of stereo sound reproduction started in the 1930's. It resulted from two apparently different approaches.

The first approach is that described in the patent of Blumlein [1936], which claims a system for capturing, transmitting, and reproducing sound which relies on Rayleigh's duplex theory [1896]. The recording system consists in two identical microphones, possibly separated by a baffle, spaced apart by a typical distance of 20 cm, both facing the same direction. At low frequencies, Blumlein found that, using a post-processing step known as Blumlein shuffling, a difference in phase at the input, i.e. in the microphone signals, turns into an amplitude difference at the output. By subsequently playing the signals on a stereo loudspeaker pair, the necessary interaural time differences (ITD, see Section 2.1.1) are reproduced at the listener's ears. Blumlein shuffling filters the sum-and-difference form of two stereo signals [Gerzon, 1994]. The introduction of a phase shift in the path of the difference signal converts a difference in phase between the two input stereo signals into a difference in amplitude. However, Blumlein shuffling has practically no effect at high frequencies. At high frequencies, the baffle between the microphones acts as an acoustic obstacle and the difference in sound pressure level between the microphone signals goes through electrical amplification before being reproduced by the loudspeakers in order to produce the necessary interaural level difference (ILD) at the listener's ears.

The second approach to stereophonic sound is the engineering compromise of

Steinberg and Snow [1934], based on the idea of an ideal acoustic curtain [Fletcher, 1934]. This curtain is made of a myriad of tiny microphones, and is placed in front of the scene to be recorded. In the reproduction space, each microphone signal, by essence different from the others, is fed to a corresponding loudspeaker and the combination of all the loudspeakers recreates a true audio perspective to the ears of the listener, who uses natural hearing to perceive the recreated soundscape. The compromise is then to reduce the number of microphones and loudspeakers to only two or three to capture and reproduce the sound scene. As Snow [1953] acknowledged, the idea of an ideal acoustic curtain is fundamentally different from the engineering compromise of two or three-channel stereo. The perception of the latter is based on a different phenomenon in the brain, called the *precedence effect* [Litovsky et al., 1999; Wallach et al., 1949]. The precedence effect is the perceptual phenomenon responsible for the grouping of similar sounds separated by short time delays (below around 1 ms). In a reverberant environment, for example, the sounds coming from a source and reflected from the walls, floor, and ceiling are not heard as separate auditory events. Instead, the direct sound and its reflections are fused, and the localization of the source is dominated by the direction of the first incident wave. The acoustic curtain approach, however, is similar in concept to the spatial audio technique known as Wave Field Synthesis, which is discussed in Section 3.3.4.

The perceived direction of origin resulting from a pair of amplitude panned stereo signals (left, L, and right, R) can be computed, based on an energetic sound field analysis, and is dependent on the cross-correlation ϕ_{LR} of the signals [Merimaa, 2007]. The perceived direction of the phantom source θ is given by

$$\frac{\tan\theta}{\tan\theta_0} = \frac{m_{LR} - m_{RL}}{m_{LR} + m_{RL} + 2\,\mathfrak{Re}\{\phi_{LR}\}} \tag{3.2}$$

where θ_0 is 30°, assuming the standard stereo layout, and $m_{LR} = 1/m_{RL}$ is the magnitude ratio between the two loudspeaker gains. The directions θ obtained with this equation are plotted in Figure 3.2 for several values of the cross-correlation ϕ_{LR}. In the case of the amplitude panning of one input signal ($\phi_{LR} = 1$), we can see that the relationship is almost linear, as stated in Equation (3.1). In addition, Equation (3.2) reduces to the tangent panning law [Pulkki and Karjalainen, 2001]

$$\frac{\tan\theta}{\tan\theta_0} = \frac{g_L - g_R}{g_L + g_R} \tag{3.3}$$

where g_L and g_R are the gains applied to the unique input signal.

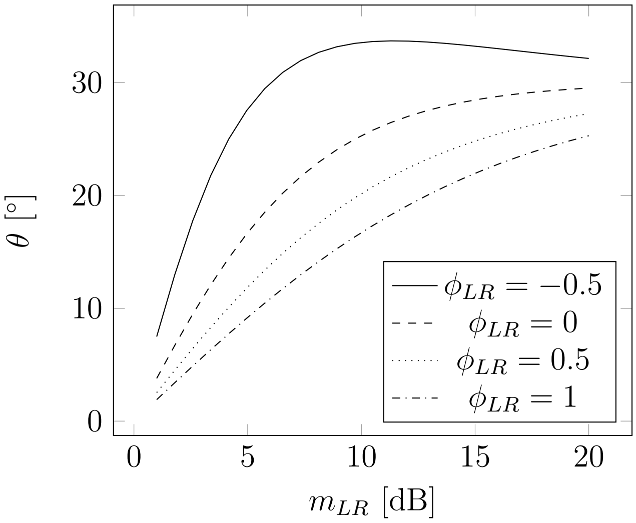


Figure 3.2: Direction of arrival θ of the net sound energy as a function of the magnitude ratio m_{LR} for stereo loudspeakers at ±30°. The different curves correspond to different values of the cross-correlation coefficient ϕ_{LR}. After [Merimaa, 2007].

We now discuss the capabilities of stereo sound in terms of cues present in the reproduced sound field. This discussion is summarized in Table 3.1. The focus is on intensity panning only, because it produces narrower phantom sources [Lee and Rumsey, 2004].

Table 3.1: Spatial cues and spatial attributes of a sound field reproduced with stereo sound.

(a) Spatial cues provided in the sound field reproduced with stereo sound.

	Stereo (amplitude panning)
ITD	correct below 1.1 kHz
ILD	correct above 2.6 kHz
Spectral cues	no
Distance cues	simulated

(b) Spatial attributes of sound reproduced with stereo sound.

	Stereo (amplitude panning)
Azimuth	$\left[-30°, 30°\right]$
Elevation	no
Near field	no
Distance, depth	simulated
Spatial impression	simulated
Envelopment	no

Azimuth reproduction In anechoic conditions, Pulkki and Karjalainen [2001] asked participants to pan phantom sources to try and reproduce the sound direction of two real sources at ±15°. The results were also compared to a computational model of auditory localization [Pulkki et al., 1999]. Results show that, for broadband stimuli, the ITD is correctly reproduced by the stereophonic setup up to about 1.1 kHz. The ILD deviates more from the intended panning angle, and oscillates, but suggests roughly the correct direction at frequencies below 0.5 kHz and above 2.6 kHz. The discrepant cues produced in the range 1.1 kHz to 2.6 kHz suggest different directions, and are therefore thought to increase the perceived source width.

Elevation reproduction Obviously, sources reproduced with stereo sound are confined to the horizontal plane. However, the phantom source may be perceived with a slight elevation (up to a few degrees), depending on the frequency content of the signal and the characteristics of both the loudspeakers and the reproduction room.

Distance reproduction Much like a picture, in which depth is perceived at the physical distance of the printed object, auditory distance reproduction in stereo systems allows for the perception of auditory depth at the physical distance of the loudspeakers. All distance cues related to the sound spectrum are reproduced (intensity, air absorption, and direct-to-reverberant ratio).

Because accurate stereo reproduction is limited to a sweet spot in space, erroneous distance cues are produced when the listener moves.

3.1.2 Nomenclature of two-channel reproduction systems

Stereo is not the only way to convey spatial sound over two channels. Therefore, it is important to differentiate between the possible chains involving one or two amplifying channels. These systems are also depicted in Figure 3.3. In the following nomenclature, adapted from [Snow, 1953; Streicher and Everest, 2006], the suffix *-aural* indicates reproduction via headphones, while the suffix *-phonic* indicates that the intended reproduction is via loudspeakers.

Monaural The sound is recorded with one microphone, and one channel is used to convey the sound to one ear via an earphone. Before electrical amplification was available, this was the standard. Nowadays, this term is used to describe the reproduction of a unique signal at both ears via headphones.

Binaural The sound is recorded by placing two microphones in the ears either of a dummy head or of a real person. The real person can be the intended listener or someone else. Two amplifying channels convey the sound to each corresponding ear at the listener's end via headphones.

Monophonic The term monophonic sound, or "mono", is used when considering sound recorded with one or several microphones, transmitted on one channel, and reproduced on one or several loudspeakers. The emphasis is on the single processing and transmission channel. When several microphones have captured the sound, a mixing stage is necessary.

Stereophonic Two channel stereophonic sound, or "stereo", refers to the capture of sound by two carefully placed microphones, the transmission over two channels, and the reproduction over two carefully placed loudspeakers to a listener at a fixed location. Several microphone arrangements exist [Streicher and Everest, 2006], and the associated processing step of the two signals may vary.

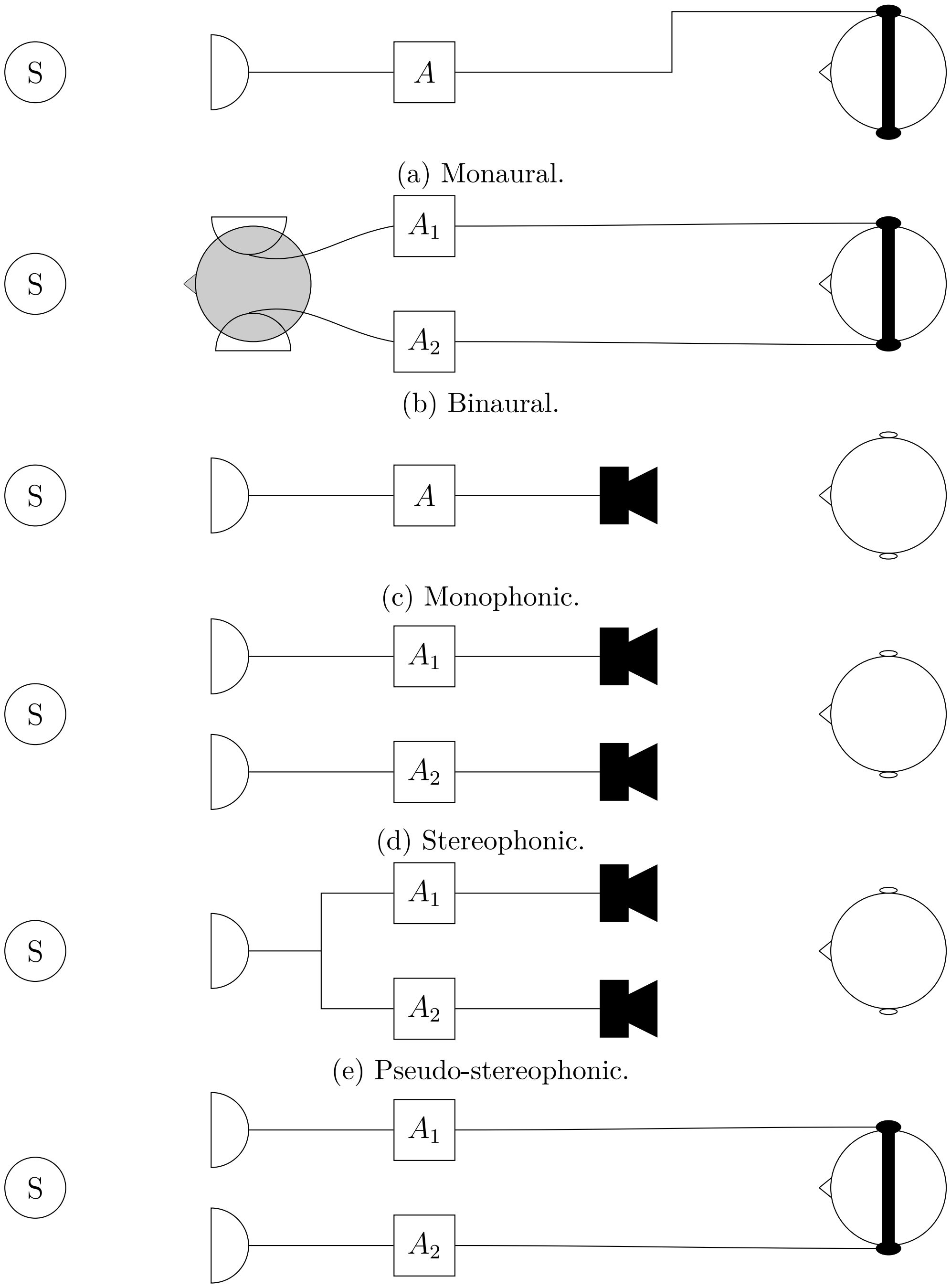
Pseudo-stereophonic In this case, the source is recorded with a single microphone. Different gains and delays in the two amplifying channels will result in a phantom source that moves between the two loudspeakers used for reproduction.

Biphonic This is nowadays a very frequent situation when sound intended for stereophonic reproduction is listened to via headphones, for example with a mobile phone or an MP3 player.

With this nomenclature in mind, we now turn to the historical review of the different technologies that appeared in movie theaters over time.

3.2 A brief history of sound reproduction in movie theaters

In this brief historical record, we focus on the technological aspect of sound reproduction in movie theaters. We do not consider creative aspects as well as the financial aspects. We acknowledge that these last two factors have had a large influence on how sound technology evolved over the course of time. Still, the goal here is not to study how (or why) sound reproduction technology became what it is today but rather to describe the different milestones in technology that



(a) Monaural.

(b) Binaural.

(c) Monophonic.

(d) Stereophonic.

(e) Pseudo-stereophonic.

(f) Biphonic.

Figure 3.3: Classification of the various forms of sound reproduction. See the text for more information. After Snow [1953]; Streicher and Everest [2006].

were reached since the first sound recording/playing device, the *phonograph*, was invented by Thomas Edison in 1877. As Theile [1990] states, the purpose of cinema sound is not about achieving the optimum naturalness of the soundscape but rather about producing spatial effects to a large audience in an economic way.

3.2.1 From silent movies to talkies

Ironically, the phonograph marks the beginning of the history of sound for moving pictures before the actual invention of the moving picture. Indeed, an article in *Scientific American* [1877] envisioned the combination of Edison's phonograph with the projection of stereoscopic images from the beginning:

> It is already possible by ingenious optical contrivances to throw stereoscopic photographs of people on screens in full view of an audience. Add the talking phonograph to counterfeit their voices, and it would be difficult to carry the illusion of real presence much further. (*Scientific American* [1877])

However, the audience would have to wait until 1891 to see moving pictures. Edison, again, designed an apparatus for one spectator (the *Kinetoscope*), and later in 1895, the Lumière brothers invented a device capable of capturing, processing, and projecting moving images for a larger audience (the *Cinématographe*).

The first movies projected in theaters were sometimes accompanied by live sound performed by an orchestra or a pianist [Altman, 1995]. Until around 1910, the movie experience became closer and closer to live theater performance as professional actors spoke lines in sync with the images, carefully hidden behind the screen. In parallel, inventors around the world experimented with the synchronization between a disc containing a soundtrack and the movie projection. Since these audio systems aimed primarily at reproducing the human voice, thereby replacing the live actor, the first loudspeakers moved from a position next to the projector (behind the audience) to a position near the screen.

Producers, however, favored music accompaniment from the early 1910s until the time in the mid-1920s, when Bell Labs developed the *Vitaphone*. This system used electrical amplification to bring the sound to the audience (Figure 3.4). The first feature film with synchronized sound was "Don Juan" in 1926. The soundtrack contained music and sound effects. One year later, "The Jazz Singer", the first feature film with synchronized dialogs, was released, again using Vitaphone sound.

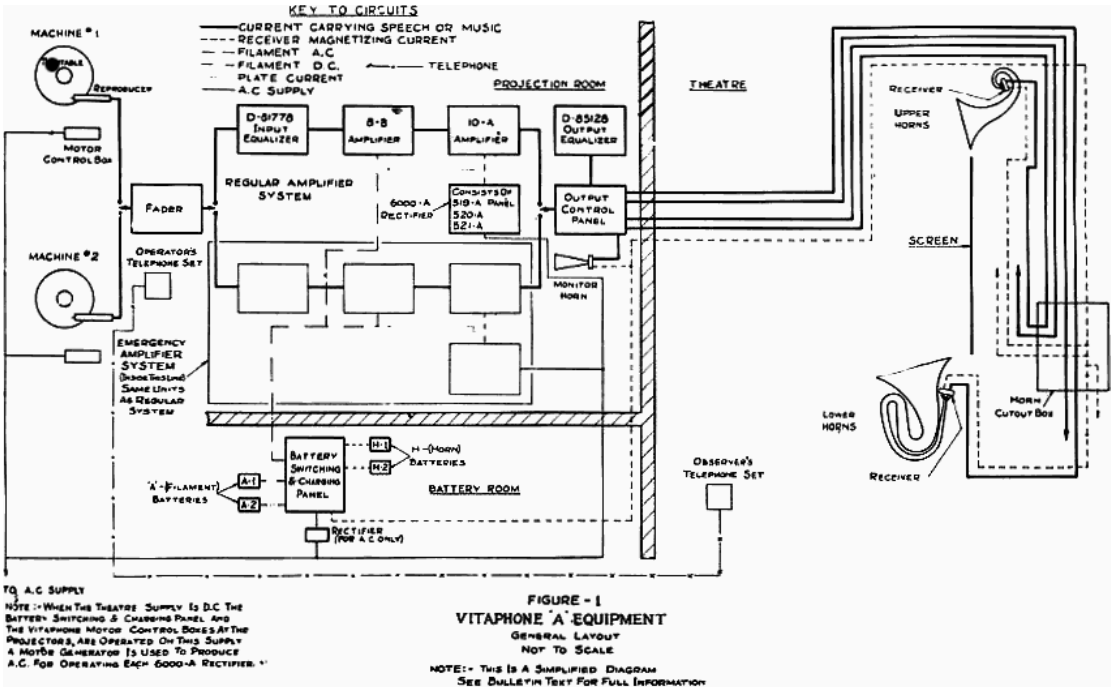


Figure 3.4: A Vitaphone system diagram, circa 1927. © 2000 The American WideScreen Museum.

3.2.2 From monophonic to stereophonic sound

The electrical amplification of sound signals pushed research forward towards stereophonic reproduction. At the same time, the speakers moved behind the screen for a better association between the sound and the image. However, stereo sound did not reach theaters directly, mainly because the 1929 crisis and World War II prevented studios and theater owners to invest at that time.

3.2.3 The development of multiphonic sound

The development of multiphony started with Disney's "Fantasia", in 1940. A particular sound system, *Fantasound*, was conceived for the reproduction of this movie. Two projectors were needed to play the movie, one for the image and a back-up mono soundtrack, and one for the three-channel stereo soundtrack, combined to a synchronization track, all optically encoded. The three channels drove a stereo system behind the screen. In addition, a combination of the left and right signals drove loudspeakers surrounding the room. The movie was presented as a roadshow in 1940 and 1941. However, due to the complexity of the hardware and the political turmoil of the time, the movie was re-released in mono, which was easier to distribute, and *Fantasound* was not used for any other movie.

Although the use of *Fantasound* was short-lived, engineers kept the idea of a three-channel stereo system behind the screen. After World War II, improvements in sound reproduction happened around the audience, where dedicated channels were used to reproduce ambience and sound effects. In 1952, the new *Cinerama* system combined three projectors illuminating a huge curved screen and multi-channel sound. A five-channel stereo system was used behind the screen and the sixth channel drove loudspeakers in the room. Shortly after, Fox's *CinemaScope* used four tracks (on 35 mm film) or six tracks (on 70 mm film), again with one channel controlling the loudspeakers in the room, and the others controlling a front stereo system.

Magnetic recording of the soundtrack was very expensive, and prevented the generalization of surround sound in movie theaters. This was the case until the first episode of the "Star Wars" saga, which used *Dolby Stereo*, an optical encoding of four tracks on 35 mm film. This encoding combined a noise reduction technique (called *Dolby A*) to a 4:2:4 matrixing. This means that four tracks were recorded, encoded into two channels, and decoded back to four channels in the theater. Hence the name "stereo", which indicates that the four tracks only occupy the space of two tracks on the film. These four tracks were divided into a three channel stereo system at the front (left, L, center, C, and right, R) of the theater, and a surround channel at the back. A conventional loudspeaker arrangement is illustrated in Figure 3.5.

3.2.4 The digital sound era

The sound of movies went through digitalization long before the image. Dolby launched *Dolby Digital* with the release of "Batman Returns", in 1992. But it was rather "Jurassic Park", one year later, that encouraged theater owners to upgrade to digital sound. Digital sound at that time was still printed on film, just as in *Dolby Stereo*. Both systems could actually live next to each other, as well as next to their competitors *SDDS* (Sony Dynamic Digital Sound), and a time code for *DTS* (Digital Theater System). The latter stored the sound information on an additional CD-ROM. A picture showing these four systems on a film track is shown in Figure 3.6.

Dolby Digital and DTS surround allowed up to six channels. These are commonly referred to as 5.1. Five channels drive full range loudspeakers, from about 20 Hz to about 20 kHz. The five tracks divide in a three channel stereo system at the front (L, C, and R) of the theater, and two surround channels at the back, at

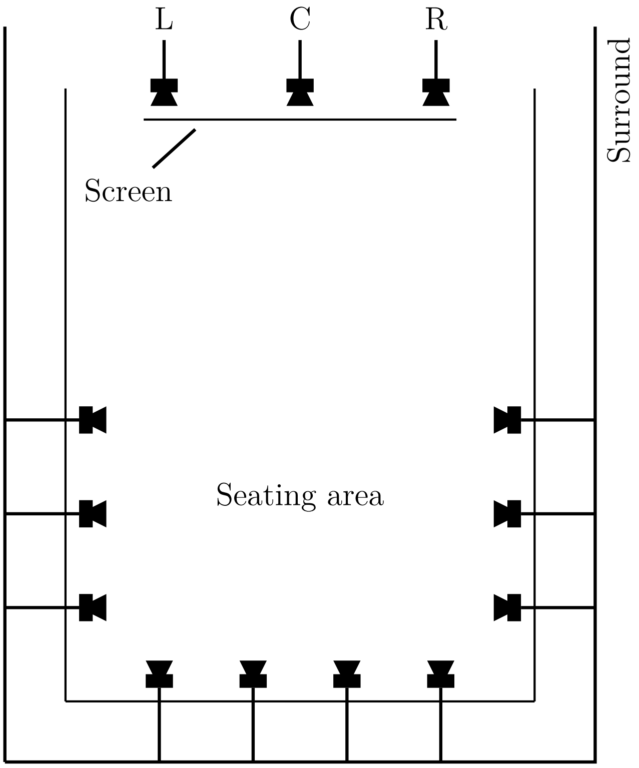


Figure 3.5: Arrangement of loudspeakers for conventional motion picture sound. The surround channel may be divided into several channels, depending on the encoding. After Streicher and Everest [2006].

the left (LS) and the right (RS) of the room. The sixth channel, 0.1, is a special channel that drives a subwoofer, with a range of about 20 Hz to 120 Hz, or about one tenth of the normal range in logarithmic units. This channel has a particular use in cinema sound as it conveys the low-frequency effects (LFE). SDDS has the particularity to add two channels at the front, between the three stereo channels already defined.

In 1999, Dolby, working with THX, released its new format called *Dolby Digital Surround EX* for the movie “Star Wars Episode I: The Phantom Menace.” The new format added a center surround channel, matrixed in the left and right surround channels for backward compatibility with 5.1.

Recently, the channel count has stabilized to 7.1, although, contrary to 5.1, this layout has not been standardized. The two channels in addition to 5.1 can be placed on the sides (Dolby, DTS) or at the front (SDDS). Dolby and DTS have also released lossless audio codecs, respectively called *Dolby TrueHD* and *DTS HD Master Audio*.

Along with the new available channels came the challenge of recording for such

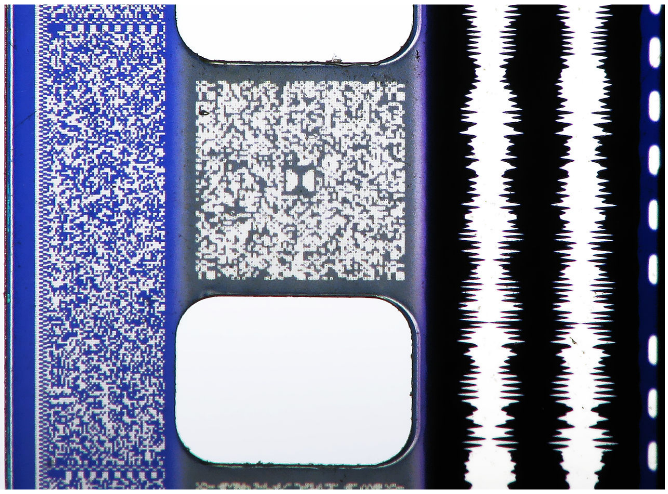


Figure 3.6: A photo of a 35 mm film print featuring four audio formats – from left to right: SDDS (blue area to the left of the sprocket holes), Dolby Digital (grey area between the sprocket holes labeled with the Dolby Double-D logo in the middle), analog optical sound (the two white lines to the right of the sprocket holes), and the DTS time code (the dashed line to the far right.) License: Creative Commons Attribution-Share Alike 3.0 Unported license.

formats. Different microphone arrangements exist to capture a natural sound field for subsequent 5.1 reproduction. Some arrangements focus on the quality of the spatial image, mainly sought in the frontal direction (INA 3, near-coincident line, Decca-Tree, and OCT, see [Theile, 2000] for a discussion), and some combine a frontal spatial image with an enveloping ambience at the back into one arrangement (INA 5, Fukada Tree, and OCT-Surround, see [Theile, 2000] for a discussion). Most of the time, such a microphone rig is used in combination with spot microphones placed close to particular sources of interest, resulting in dry signals, and the various signals are mixed to produce the final soundtrack, spread on the whole loudspeaker layout.

3.2.5 A glimpse into the near future

Several formats, available for production today, go beyond 7.1. This increase in channel count is mainly an attempt to increase the area of correct reproduction, or *sweet-spot*, of the reproduction system [Hamasaki et al., 2004]. More recently, the focus has also been on adding height information [Hamasaki et al., 2006; Van Daele and Van Baelen, 2012].

Advertising it as “twice as good as 5.1”, Tomlinson Holman introduced the

10.2 format [DellaSala, 2006]. It includes the ITU 5.1 layout and adds to it a back surround, the two additional stereo channels of SDDS, and two channels above the L and R channels of ITU 5.1, as well as a second LFE channel to form a stereo pair.

For movie theaters, Auro Technologies proposes two formats: 11.1 and 13.1. 11.1 consists in the addition of height channels above each full range channel in the 5.1 layout, as well as a channel above the audience. The addition of a back surround channel and its height counterpart results in a 13.1 system.

NHK, the japanese national broadcast company, introduced along with their ultra-high definition TV a 22.2-channel sound system [Hamasaki et al., 2004]. This system consists of three layers of loudspeakers: the upper layer with nine channels, the middle layer (at the listener's ears level) with ten channels, and the lower layer with three channels at the front and 2 channels for low frequency effects.

3.3 3D audio technologies

The meaning of the term "3D audio" is not consistently defined in the literature. In this thesis, we use the following definition: 3D audio techniques aim at achieving the best spatialization possible, i.e. the reproduction at the listener's ears of the correct sensation of the following three factors: (1) the direction of the source, (2) the distance to the source, and (3) the associated room effect.

In this section, we review the following five 3D audio reproduction techniques: (1) Vector Base Amplitude Panning (VBAP), (2) binaural techniques, (3) transaural techniques, (4) Wave Field Synthesis (WFS), and (5) Ambisonics.

3.3.1 Vector Base Amplitude Panning (VBAP)

Principles Vector Base Amplitude Panning (VBAP) [Pulkki, 1997] extends the idea behind stereophony by adding a third loudspeaker above the stereo pair such that it is equidistant from the listener and from each of the two stereo loudspeakers. A complete 3D soundscape is achieved by placing several such triplets of loudspeakers on a sphere surrounding the listener. The listening room is assumed to be not too reverberant.

The source is restricted to lie outside of the sphere but can be created along any direction by the three closest loudspeakers. Each loudspeaker gain is then simply obtained by linear algebra involving only its direction with respect to the listener and the direction of the phantom source.

We consider the situation depicted in Figure 3.7. If the unit-length vector

$$\mathbf{l}_i = \begin{bmatrix} l_{i1} \\ l_{i2} \\ l_{i3} \end{bmatrix} \tag{3.4}$$

points to the i^{th} loudspeaker $\mathbf{l}_i$, and the unit-length vector

$$\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} \tag{3.5}$$

is the desired phantom source direction, then we can write

$$\mathbf{p} = g_1\mathbf{l}_1 + g_2\mathbf{l}_2 + g_3\mathbf{l}_3, \tag{3.6}$$

or, equivalently, $\mathbf{p} = \begin{bmatrix} \mathbf{l}_1 & \mathbf{l}_2 & \mathbf{l}_3 \end{bmatrix} \mathbf{g}$ and we can solve for the gain factors $\mathbf{g}^T = \begin{bmatrix} g_1 & g_2 & g_3 \end{bmatrix}$ of the three loudspeakers,

$$\mathbf{g} = L_{123}^{-1}\mathbf{p} = \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ l_{12} & l_{22} & l_{32} \\ l_{13} & l_{23} & l_{33} \end{bmatrix}^{-1} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}, \tag{3.7}$$

where $L_{123} = \begin{bmatrix} \mathbf{l}_1 & \mathbf{l}_2 & \mathbf{l}_3 \end{bmatrix}$, and where its inverse L_{123}^{-1} exists if the three loudspeakers define a basis that spans the 3D space.

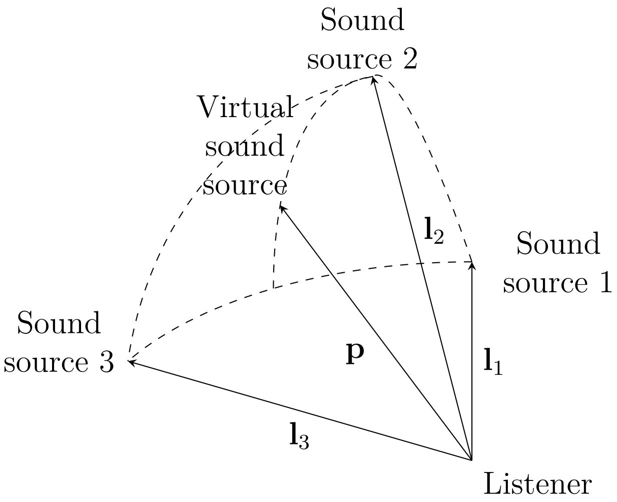


Figure 3.7: Geometry for VBAP with three loudspeakers. After [Pulkki, 2001].

We now discuss the capacities of VBAP in terms of cues present in the reproduced sound field. This discussion is summarized in Table 3.2.

Table 3.2: Spatial cues and spatial attributes of a sound field reproduced with VBAP.

(a) Spatial cues provided in the sound field reproduced with VBAP, when the centroid of the triplet is near the median plane.

	VBAP
ITD	yes (low freq.)
ILD	yes (high freq.)
Spectral cues	no
Distance cues	simulated

(b) Spatial cues provided in the sound field reproduced with VBAP, when the centroid of the triplet is not near the median plane.

	VBAP
ITD	biased (low. freq)
ILD	no
Spectral cues	no
Distance cues	simulated

(c) Spatial attributes of sound reproduced with VBAP.

	VBAP
Azimuth	yes (may be biased)
Elevation	yes (individual)
Near field	no
Distance, depth	simulated
Spatial impression	simulated
Envelopment	simulated

Azimuth reproduction A distinction must be made between triplets with their centroid in the median plane and the others [Pulkki, 2001]. When the centroid is in the median plane, the horizontal localization is accurate thanks to correct low-frequency ITD and roughly correct high-frequency ILD. When the triplet is moved away from the median plane, the low-frequency ITD biases the perceived azimuth towards the median plane. In this case, the ILD is heavily distorted.

Elevation reproduction The elevation is consistently perceived, although the results are individual and vary from subject to subject [Pulkki, 2001].

Distance reproduction Since it is based on a generalization of stereo amplitude panning (see Section 3.1.1), VBAP can only simulate auditory depth. However, due the potential presence of loudspeakers around the listener, the simulation of early reflections is possible.

3.3.2 Binaural techniques

Principles Binaural techniques aim at producing a correct sound field directly at the listener's ears. Headphones are therefore the natural reproduction system for binaural signals as they offer complete channel separation and a relative acoustic protection from the environment (although this may not be entirely desirable as will be seen later).

Sound spatialization, however, is achieved based on cues that are mostly derived from the interaction of sound with the listener's head and torso. Since using headphones bypasses this interaction, it is necessary to make sure that spatialization is included in the signal fed to each ear.

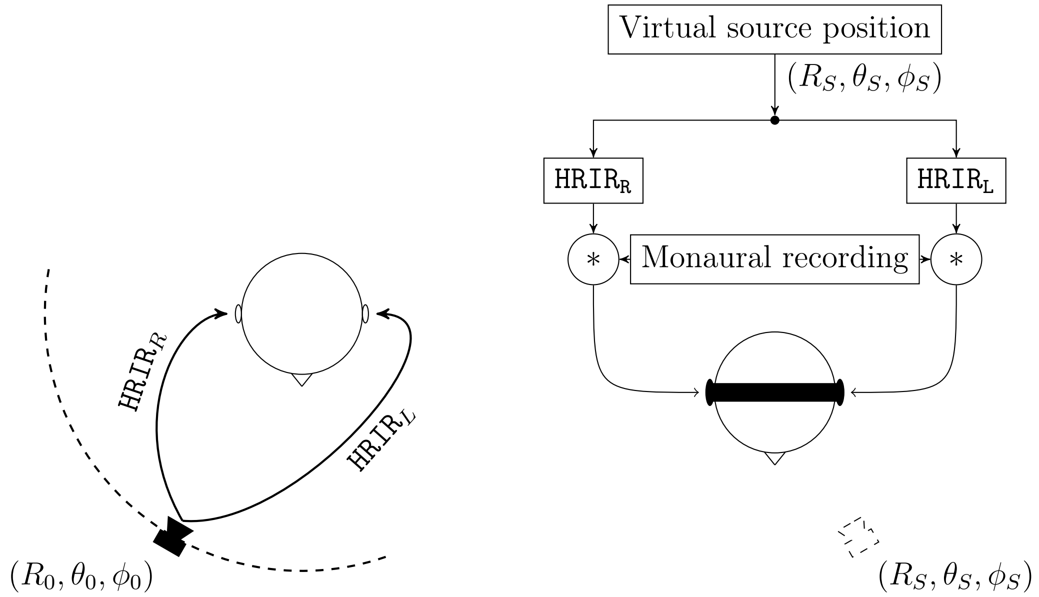
Two methods come to mind for recording and reproducing such signals. On the one hand, one can record two signals with two microphones placed in a subject's ears or with two microphones that simulate ears on a dummy head. In this case, the recorded signal contains the spatial information. On the other hand, it is possible to process a monophonic signal that does not contain any spatial information, to accurately simulate the effect of propagation of the signal from the source to the listener (Figure 3.8). This processing consists in a convolution with the Head Related Impulse Responses (HRIRs), or, in the frequency domain, by a multiplication with the Head Related Transfer Functions (HRTFs) which are the Fourier transforms of the HRIRs. In the first case, the listener hears through the ears used at the recording (either his/her own ears or someone else's), whereas in the second case it is possible to hear through one's own ears, provided that personalized HRTFs are available. Using someone else's HRTFs when listening to a binaural recording can lead to errors in spatial localization judgment [Møller et al., 1996]. However, training with feedback can lead to learning and improved localization performances [Mendonça et al., 2012; Parseihian and Katz, 2012].

The complete recording and playback chain has to be calibrated in order to reproduce the correct pressure at the eardrum. When the recording is made outside of blocked ear canals, the equalization function G can be expressed as [Blauert, 2005]

$$G = \frac{1}{M \cdot \mathrm{PTF}} \cdot \mathrm{PDR}, \tag{3.8}$$

where M is the recording microphone's transfer function, PTF is the headPhone Transfer Function, measured at the position in the ear canal where the recording is made, and PDR is the Pressure Division Ratio.

Headphones rarely present a uniform frequency response. Their PTF has


HRIR$_R$
HRIR$_L$
(R_0, θ_0, ϕ_0)
Virtual source position
(R_S, θ_S, ϕ_S)
HRIR$_R$
HRIR$_L$
*
Monaural recording
*
(R_S, θ_S, ϕ_S)

(a) HRIR measurement.

(b) Binaural synthesis.

Figure 3.8: A method for binaural reproduction.

smooth variations at low frequency and rather marked peaks at high frequencies [Møller et al., 1995]. To ensure that the headphones reproduce the correct pressure at the eardrum, it is necessary to compensate for their behavior.

PDR is best understood using an electrical equivalent for the free-field sound transmission to the external ear (Figure 3.9). The Thévenin equivalent pressure source P_{Th} and its impedance Z_{Th} model the complete sound field outside the ear canal. When the ear canal is blocked, P_{Th} and P_{entrance} are respectively measured outside and inside the blockage. The ear canal is modeled by a two-port element loaded by the eardrum impedance. The impedance seen from the entrance of the ear canal is called $Z_{\text{ear canal}}$.

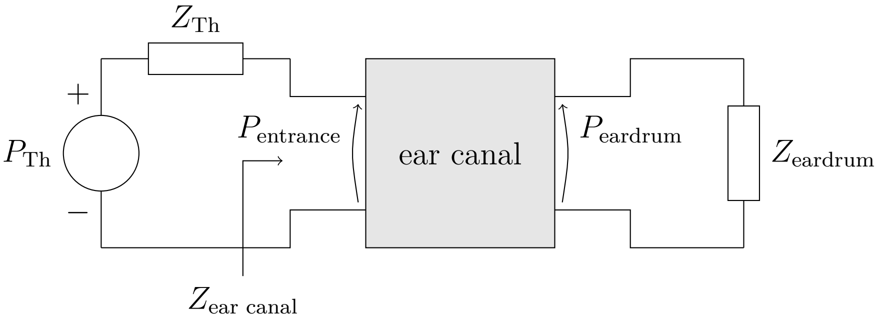


Figure 3.9: Electrical equivalent for the model of free-field sound transmission to the human external ear. After Blauert [2005].

During recording, the source impedance Z_{Th} is the radiation impedance $Z_{\text{radiation}}$

as seen from the entrance of the ear canal looking out into the free field. We have

$$\frac{P^{\text{rec}}_{\text{entrance}}}{P^{\text{rec}}_{\text{Th}}} = \frac{Z_{\text{ear canal}}}{Z_{\text{ear canal}} + Z_{\text{radiation}}}. \tag{3.9}$$

During reproduction, Z_{Th} is the headphone impedance $Z_{\text{headphone}}$ and

$$\frac{P^{\text{rep}}_{\text{entrance}}}{P^{\text{rep}}_{\text{Th}}} = \frac{Z_{\text{ear canal}}}{Z_{\text{ear canal}} + Z_{\text{headphone}}}. \tag{3.10}$$

PDR is the ratio of Equations (3.9) and (3.10),

$$\text{PDR} = \frac{Z_{\text{ear canal}} + Z_{\text{headphone}}}{Z_{\text{ear canal}} + Z_{\text{radiation}}}. \tag{3.11}$$

Determining PDR would thus theoretically require three acoustic impedance measurements. Headphones that have a PDR equal to unity are called Free-air Equivalent Coupling (FEC) headphones. Although not every headphones exhibit this property, Møller et al. [1995] have shown that, in practice, the deviation induced by the PDR is small compared to the PTF. Therefore, the PDR is rarely taken into account in practical applications.

We now discuss the capacities of binaural sound in terms of cues present in the reproduced sound field. This discussion is summarized in Table 3.3.

Table 3.3: Spatial cues and spatial attributes of a sound field reproduced binaurally.

(a) Spatial cues provided in the sound field reproduced binaurally.

	Binaural sound
ITD	yes
ILD	yes
Spectral cues	yes
Distance cues	yes

(b) Spatial attributes of sound reproduced binaurally.

	Binaural sound
Azimuth	yes
Elevation	yes (improved with head-tracking)
Near field	yes
Distance, depth	yes
Spatial impression	yes
Envelopment	yes

Azimuth and elevation reproduction Several studies have proven that, provided the HRTFs are individualized, the localization performance is almost identical for real and binaural sources in the free-field [Bronkhorst, 1995; Langendijk and Bronkhorst, 2000; Wightman and Kistler, 1989; Zahorik et al., 1995]. Still, three differences are observed: (1) front/back confusions happen around twice as often with virtual sources than with real sources, even when the HRTFs are personalized, (2) reverberation needs to be simulated, as HRTFs are most often recorded in an anechoic environment, and (3) while the azimuth localization is almost perfect, elevation yields a higher localization error. This third point seems to indicate that high-frequency localization cues, useful for elevation judgments, are biased in the HRTFs, due to the HRTF measurement method [Bronkhorst, 1995].

Head tracking should be provided to ensure sound is perceived as coming from the right quadrant [Begault et al., 2001]. Indeed, head movements have been shown to reduce reversal rates when localizing real sources, mainly by resolving the ambiguity caused by the cone of confusion [Wightman and Kistler, 1999].

Distance reproduction Most HRTFs are collected at a fixed distance, most often at around 2 m, because the functions do not vary with distance beyond approximately 1 m [Brungart and Rabinowitz, 1999]. As a result, a free-field virtual sound source is localized at an approximately constant distance when presented through headphones; see for example the reference case in [Parseihian et al., 2012]. Even when HRTFs are used, sound can appear to be located inside the head [Hartmann and Wittenberg, 1996]. Kim and Wang [2003] have shown that an appropriate reproduction equalization can lead to the externalization of sound. Early reflections up to 80 ms also contribute positively to the externalization of sound sources [Begault et al., 2001].

3.3.3 Transaural techniques

Principles We use the term transaural when the binaural recording is reproduced through loudspeakers that act as "virtual headphones." In this case, channel separation is lost and interchannel crosstalk cancels the localization cues present in the binaural signal. An adequate processing is therefore required.

According to Figure 3.10, the signals Y_L and Y_R produced by the loudspeakers are different from the ear signals Z_L and Z_R because of crosstalk [Lentz, 2006].

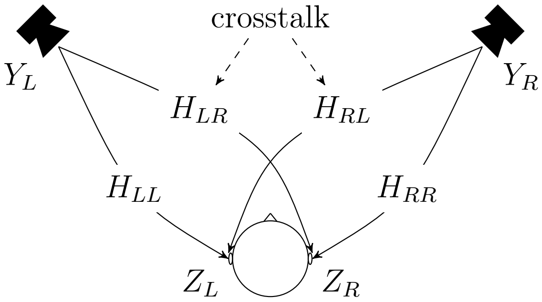


Figure 3.10: Crosstalk in transaural reproduction. After Lentz [2006].

Mathematically, we have

$$Z_L = H_{LL} \cdot Y_L + H_{RL} \cdot Y_R \tag{3.12}$$

$$Z_R = H_{RR} \cdot Y_R + H_{LR} \cdot Y_L, \tag{3.13}$$

where the functions H_{XY} are HRTFs normalized with respect to the free-field response at the center of the head, with no head present [Gardner, 1997]. The effects of loudspeaker responses and propagation to the ears are not taken into account in these relations.

Solving Equations (3.12) and (3.13) for Y_L and Y_R yields the expression

$$\begin{aligned} Y_L = & \frac{H_{RR}}{H_{LL}H_{RR} - H_{LR}H_{RL}} Z_L \\ & - \frac{H_{RL}}{H_{LL}H_{RR} - H_{LR}H_{RL}} Z_R \end{aligned} \tag{3.14}$$

and a similar one for Y_R.

This solution requires that the listener be located at a particular position, since the sweet-spot is only a few centimeters wide in the left – right direction [Vorländer, 2008]. To allow for, say, head rotation, head tracking is necessary, as well as the use of adaptive filtering to dynamically change the HRTFs [Gardner, 1997]. In that case, the solution is stable only in the angle spanned by the loudspeakers. Instability arises from the inherently iterative structure of the crosstalk cancellator. In a real-world situation, the presence of noise can lead to sound coloration, ringing, or even range overflows (the amplitude of sound exceeds the maximum value admissible by the sound device) [Lentz, 2006].

Azimuth, elevation, and distance reproduction Since transaural sound aims at reproducing binaural sound with loudspeakers, the performance in terms of reproduction can theoretically reach that of binaural reproduction, and we therefore refer the reader to Table 3.3.

3.3.4 Wave Field Synthesis

Principles According to Huygens' principle, each point of a wavefront can be considered to be a secondary source. Therefore, by producing the correct pressure at each of these points, it is possible to reconstruct the whole wavefront with a loudspeaker array. This is the idea behind Wave Field Synthesis (WFS), originally suggested by Berkhout [1988]. The comparison between a real source and a WFS virtual source is shown in Figure 3.11. Here, we are interested in solutions of the wave equation of the form

$$p(\vec{r}, t) = P(\vec{r})e^{j\omega t}, \tag{3.15}$$

where ω is the angular frequency ($\omega = 2\pi f$). When no sound source is present at $\vec{r}$, P satisfies the homogeneous Helmholtz wave equation

$$\nabla^2 P(\vec{r}) + \left(\frac{\omega}{c}\right)^2 P(\vec{r}) = 0, \tag{3.16}$$

where c is the speed of sound, in a closed volume Ω surrounded by a surface Λ.

If the sound source is located at r_0 outside of Ω, then the pressure at a point $\vec{r}$ inside the volume is given by [Cox and D'Antonio, 2004]

$$P(\vec{r}, \vec{r}_0) = \int_\Lambda \left[G(\vec{r}|\vec{r}_q)\frac{\partial P(\vec{r}_q)}{\partial \vec{n}_q} - P(\vec{r}_q)\frac{\partial G(\vec{r}|\vec{r}_q)}{\partial \vec{n}_q} \right] \mathrm{d}\,\Lambda, \tag{3.17}$$

where $\vec{r}_q$ denotes a generic position on Λ and $\vec{n}_q$ the outward pointing normal at $\vec{r}_q$, and $G(\vec{r}|\vec{r}_q)$ is the Green function, which is the solution of the inhomogeneous Helmholtz equation

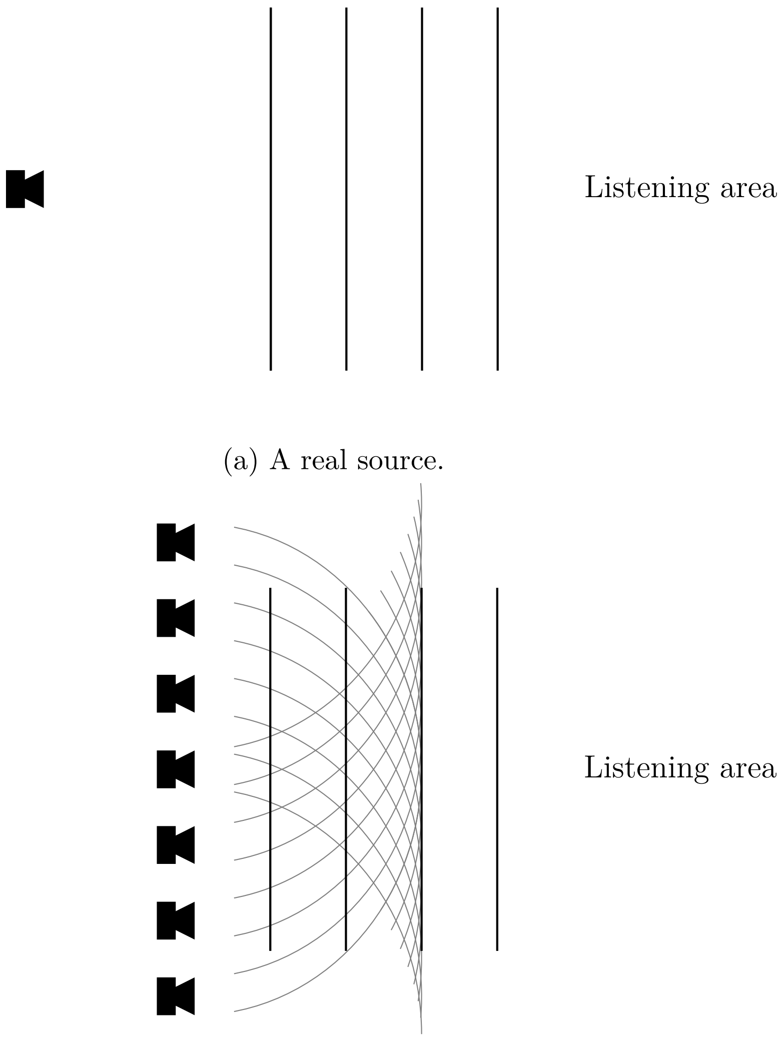
$$\nabla^2 G(\vec{r}|\vec{r}_q) + \left(\frac{\omega}{c}\right)^2 G(\vec{r}|\vec{r}_q) = -\delta(\vec{r} - \vec{r}_q) \tag{3.18}$$

where δ is the Dirac "delta function" of appropriate dimensionality. In three dimensions, we have

$$G(\vec{r}|\vec{r}_q) = \frac{e^{-jk|\vec{r}-\vec{r}_q|}}{4\pi|\vec{r} - \vec{r}_q|} + F, \tag{3.19}$$

where F is any function satisfying the homogeneous Helmholtz wave equation given in (3.16).

The first term in the integral in Equation (3.17) corresponds to a dipole source



(a) A real source.

(b) A WFS loudspeaker array.

Figure 3.11: Comparison between the production of a plane wave by a real source and a WFS loudspeaker array. With a WFS loudspeaker array, the contributions from all loudspeakers add to produce the target wavefront in the listening area.

distribution, and the second term corresponds to a monopole source distribution [Spors and Ahrens, 2008]. These sources are referred to as *secondary sources*, while the original sound source producing the target sound field is referred to as the *primary source*. The equation means that appropriate monopole and dipole source distributions placed on the boundary Λ determine the sound field in the volume Ω. For practical implementations, only monopole sources are considered because they can be implemented by traditional (boxed) loudspeakers. Thus, the Green function is chosen to cancel the second term in Equation (3.17).

For practical implementation, the reproduction is often limited to a linear array of loudspeakers, because most information is considered to be included in the horizontal plane at the height of the listener's ears [Berkhout et al., 1993]. Therefore, three additional operations are needed: (1) the geometry of Λ is particularized to an infinite planar distribution, (2) the infinite planar distribution is reduced to a continuous line source, and (3) the continuous line is discretized.

In practice, therefore, the sound field produced by a WFS array differs from the sound field of a single punctual source because of the finiteness of the array (which causes diffraction), the discreteness of the array (which causes spatial aliasing), and the linearity of the array (which causes cylindrical wavefronts instead of spherical wavefronts). In addition, the room in which the array is placed causes unwanted reflections. We now discuss each of these drawbacks.

Diffraction effects Limiting the array in length causes diffraction: the wavefront is correctly synthesized, but delayed secondary wavefronts are created that can be perceived either as coloration of the first wavefront, or as echoes. To limit the impact of diffraction, one applies weights on the loudspeaker gains which lower the contribution of loudspeakers near the edges of the array. This method is referred to as *tapering*. This is a special case of spatial windowing [Vogel, 1993, Section 3.4].

Unfortunately, because the number of effectively active loudspeakers to reproduce a sound source is limited by tapering, the method has the undesirable effect to reduce the area of correct reproduction [Boone et al., 1995].

Spatial aliasing As long as the frequency is below a certain aliasing frequency f_{Nyq}, the reproduction is stable in the whole playback area. However, because the discretization leads to spatial sampling of the sound field, the distance Δx separating the successive loudspeakers must be kept small. For a given Δx, the reproduction of the sound field is incorrect at frequencies above

$$f_{\mathrm{Nyq}} = \frac{c}{2\Delta x}. \tag{3.20}$$

This implies a spacing Δx of at most 8.5 mm to achieve a correct reproduction up to 20 kHz.

In order to evaluate the impact of the spatial aliasing, Start [1997, Section 6.5] measured with a KEMAR dummy head [Gardner and Martin, 1995] the sound field produced by a single loudspeaker and a virtual source on a WFS array with $\Delta x = 11$ cm. The stimuli were broadband and band-limited white noises from 100 Hz to 8000 Hz, and from 100 Hz to 1500 Hz, respectively. Then, the recorded (binaural) signals were played to participants through headphones. A 2-AFC paradigm (see for example [Wichmann and Hill, 2001]) was used to evaluate the minimum audible angle (MAA) for each stimulus. No difference was found between real and virtual sources, both for the broadband stimulus (MAA = 0.8°) and the band-limited stimulus (MAA = 1.1°). Furthermore, the doubling of Δx led to an increase of the MAA of around 0.5° for both type of stimulus. These findings first supported the compromise to limit the aliasing frequency to about 1500 Hz in WFS arrays.

Verheijen [1998] performed an experiment where participants directly faced the WFS loudspeaker array, comparing the localization accuracy of virtual WFS sources and real sources. With a loudspeaker spacing of 22 cm, the mean RMS error was 3.2°, with a standard deviation of the error of 1.4°. This was only slightly higher than the results for real sources, which were 2.6° and 1°, respectively.

Above the sampling frequency, the reproduced sound field consists of the correct synthesized direction superposed to aliased contributions. Because the effect of aliasing on the sound field depends on the frequency of the signal, the erroneous contributions do not bias the perception of direction towards any particular direction. Instead, they blur the correct direction and decrease the localization accuracy.

To prove this last point, Wittek [2007, Chapter 7] compared the sound field produced by a single loudspeaker (equivalent to a real source), a WFS array with loudspeakers separated by $\Delta x = 4.2$ cm, a second WFS array with $\Delta x = 12.7$ cm, and phantom sources produced by a stereophonic pair. The stimulus was a series of pink noise bursts. The participants were asked to point with a laser to the direction of the sound source (directional accuracy). They also graded on a 5-point scale how well they could pinpoint the location of the real and virtual sound sources (apparent source width). The scale ranged from very bad to very good. In terms of directional accuracy, all systems under test could accurately reproduce sound direction. In terms of apparent source width, the WFS array with the smallest Δx performed statistically significantly better than the other, indicating that spatial aliasing does create localization blur. Compared to a single loudspeaker, however, both WFS arrays performed slightly, but statistically significantly, worse.

Two reasons may explain why directional accuracy is maintained in WFS systems with an aliasing frequency at around 1.5 kHz. First, Wightman and Kistler [1992] have shown that the ITD is the dominant cue in auditory localization. They presented synthesized binaural signals to subjects and asked them to judge the source direction in spherical coordinates. The stimuli were high-passed trains of Gaussian noise. Results showed that ITD dominated localization when frequencies below 5 kHz were available. Other available cues included ILD and spectral cues. For several subjects, perception was already accurate when the signal was high-passed filtered at 2.5 kHz. Therefore, it seems that, as long as the ITD is correctly reproduced at the listener's ears, the azimuthal localization of the sound source is accurate. A second contributing factor to the accurate estimation of source direction in WFS systems is the precedence effect (see Section 3.1.1). Indeed, at any frequency, the first signal arriving from a non-focused source (a source located outside of the array) to the listener's ears always comes from the loudspeaker in the direction of the source. This is because it is the loudspeaker with the smallest distance, and thus the smallest time delay, on the path between the virtual source and the listener. As a result, in WFS systems, the precedence effect gives a cue to the correct direction of a non-focused source at all frequencies, including those above the aliasing frequency [Wittek, 2007, Section 4.3].

Amplitude error As an example of this error, we consider the reproduction of a plane wave. Because sound reproduction on the WFS array is limited to the horizontal plane, the wavefronts are cylindrical. This results in a 3 dB amplitude decay per doubling distance, instead of a constant amplitude for the ideal plane wave. We refer the reader to [Sonke, 2000, Section 4.3], which provided a thorough analysis of this discrepancy.

Room reflections The reflections from the room which encloses the WFS array can lead to erroneous distance and direction perception. In the field of auralization, where the goal is to simulate acoustically a virtual room, both the real and the simulated rooms contribute to the perception, which is undesirable. Adequate equalization can decrease the influence of the reproduction room, using the WFS array itself [Corteel and Nicol, 2003].

Azimuth, elevation, and distance reproduction The spatial cues and spatial attributes reproduced by WFS are summarized in Table 3.4. The content from this table generally follows from the previous discussion.

Table 3.4: Spatial cues and spatial attributes of a sound field reproduced with WFS.

(a) Spatial cues provided in the sound field reproduced with WFS.

	WFS
ITD	correct below f_{Nyq}
ILD	correct below f_{Nyq}
Spectral cues	no
Distance cues	correct below f_{Nyq}

(b) Spatial attributes of sound reproduced with WFS.

	WFS
Azimuth	correct in the acoustic window
Elevation	no
Near field	yes (focused source)
Distance, depth	yes
Spatial impression	yes (assuming a 360° window)
Envelopment	yes (assuming a 360° window)

3.3.5 Ambisonics

Principles Ambisonics is a technique based on Gerzon's work [1985]. It is designed to reproduce plane waves coming from any direction. Recent mathematical developments generalize the concept to point sources [Zotter et al., 2009]. Instead of using integral theorems as in WFS, the solution to Equation (3.17) is given in polar coordinates using a Fourier-Bessel series expansion of the pressure. Accordingly, the array geometries are usually cylindrical or spherical. Hence, a planar ambisonic array is usually circular, and a fully 3D array is usually spherical.

Here, we limit our analysis to a planar array [Bamford and Vanderkooy, 1995]. The reader is invited to consult [Daniel, 2001] for a development of the full 3D case.

The pressure at (r, θ) caused by a plane wave coming from the direction ψ (Figure 3.12) is described by the following phasor:

$$P_{\psi}(r, \theta) = P_{\psi,\max} e^{jkr\cos(\theta-\psi)}. \tag{3.21}$$

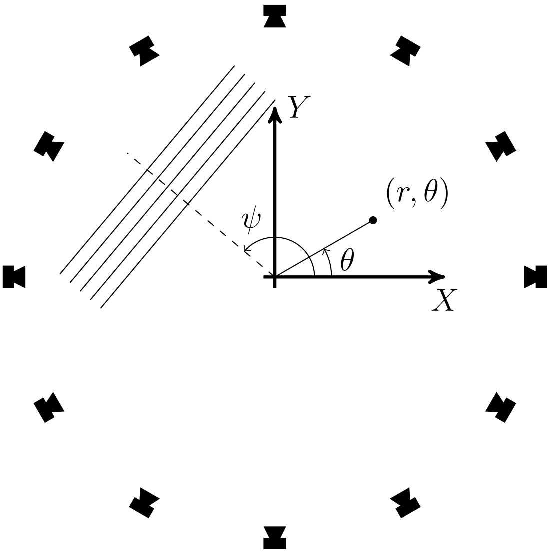


Figure 3.12: A plane wave coming from the direction ψ is produced at (r, θ) by a circular loudspeaker array. After Bamford and Vanderkooy [1995].

The Fourier-Bessel series expansion of this wave is given by

$$P_\psi = P_{\psi,\max} J_0(kr) + P_{\psi,\max} \left[\sum_{m=1}^{\infty} 2j^m J_m(kr) \cos\Big(m(\psi - \theta)\Big) \right], \tag{3.22}$$

where J_0 and J_m are cylindrical Bessel functions of the first kind.

Assuming that the loudspeakers produce plane waves at the listener's location, then the n^{th} loudspeaker produces a signal P_n from its position ϕ_n that can be written

$$P_n = P_{n,\max} J_0(kr) + P_{n,\max} \left[\sum_{m=1}^{\infty} 2j^m J_m(kr) \cos\Big(m(\phi_n - \theta)\Big) \right], \tag{3.23}$$

and the total pressure is obtained by taking the sum of the P_n's over all loudspeakers,

$$P = \sum_n P_{n,\max} J_0(kr) + \sum_{m=1}^{\infty} 2j^m J_m(kr) \left[\sum_n P_{n,\max} \cos\Big(m(\phi_n - \theta)\Big) \right]. \tag{3.24}$$

The infinite series in Equation (3.24) is limited in practice to a certain maximum value of m, noted M, called the order of the expansion. As we should have $P_\psi = \sum_n P_n$ we can easily match up the corresponding terms in Equations (3.22) and (3.24). Using the trigonometric identity for $\cos(a+b)$, we have

$$P_{\psi,\max} = \sum_n P_{n,\max} \tag{3.25}$$

$$P_{\psi,\max}\cos(m\psi) = \sum_n P_{n,\max}\cos(m\phi_n) \tag{3.26}$$

$$P_{\psi,\max}\sin(m\psi) = \sum_n P_{n,\max}\sin(m\phi_n), \tag{3.27}$$

and these equations must be satisfied for $m = 1, \ldots, M$. Each loudspeaker signal $P_{n,\max}$ can be determined by solving the whole system of equations consisting of Equations (3.25) – (3.27) for $m = 1, \ldots, M$.

The order M determines the theoretical number of channels L required for playback by

$$L = 2M + 1 \tag{3.28}$$

in this case, and by

$$L = (M+1)^2 \tag{3.29}$$

in the 3D case. When more loudspeakers are available, a least-squares solution can be computed.

Consequently, the first order expansion ($M = 1$) in 3D has four components: the absolute pressure (W) and the three pressure gradients in three orthogonal directions (X, Y, and Z). This expansion is known as B-Format. The SoundField Microphone natively captures a sound field in B-Format [Farrar, 1979]. Extensions of the reproduction to $M > 1$ are referred to as Higher-Order Ambisonics, or HOA.

In practice, the sound field produced by an ambisonic array differs from the sound field of a single punctual source because of the limited angular resolution, resulting from the truncation of the wavefield expansion, and the discreteness of the array, which results in spatial aliasing [Spors and Ahrens, 2008; Zotter et al., 2009]. However, the contribution of the discretization of the array contributes only slightly to the shrinking of the sweet-spot, in comparison to the truncation error [Zotter et al., 2009].

Truncating the expansion reduces the sweet-spot when the frequency increases. However, the representation remains valid whatever the frequency, provided that the listener is at the sweet-spot. Also, at a given frequency, the size of the sweet-spot increases with the order of the expansion. Taking into account only the truncation error, Ward and Abhayapala [2001] have shown that taking

$$M = \lceil kr \rceil \tag{3.30}$$

as a rule of thumb limits the reproduction error to about 4% in a sphere of radius r. For example, 25 loudspeakers are sufficient to allow accurate reproduction for one listener ($r = 0.1$ m) up to 2 kHz.

In addition, the room in which the array is placed causes unwanted reflections, but these can be taken into account [Poletti, 2005].

The spatial cues and spatial attributes reproduced by Ambisonics are summarized in Table 3.5. Compared to WFS, there is little published work available that evaluates the localization performance of HOA systems. This is a current ongoing research effort.

Table 3.5: Spatial cues and spatial attributes of a sound field reproduced with Ambisonics.

(a) Spatial cues provided in the sound field reproduced with Ambisonics.

	Ambisonics
ITD	correct in the sweet-spot
ILD	correct in the sweet-spot
Spectral cues	correct in the sweet-spot
Distance cues	correct in the sweet-spot

(b) Spatial attributes of sound reproduced with Ambisonics (in the sweet-spot).

	Ambisonics
Azimuth	yes
Elevation	yes
Near field	yes (NFC-HOA)
Distance, depth	yes
Spatial impression	yes
Envelopment	yes

Azimuth reproduction Because of the form of the equations, Ambisonics is better suited for circular (2D) or spherical (3D) array geometries. In all cases, the listener is surrounded by loudspeakers and the system can reproduce a source at any azimuth.

In the horizontal plane, Benjamin et al. [2010] have shown, using a virtual experiment, that the ITDs are correctly reproduced at $M = 1$ by an octagonal array of loudspeakers, at least in the tested range 100 Hz to 800 Hz. This result does not depend on the shape of the array. The ILDs, averaged from 1 kHz to 3.2 kHz, are generally larger than that produced by natural hearing, and they change more rapidly near the front and near the back.

Bertet et al. [2007] have compared subjectively the horizontal localization accuracy for 3D ambisonic microphones with $M = 1, \ldots, 4$. Results show that the accuracy increase with the order M. At $M = 4$, the median angle error stays below 5° at any angle. Localization blur, measured by the interquartile range, decreases with M, but remains important at lateral angles. An objective analysis of the ITD corroborates these results: as the order increases, the ITD matches more and more that of a real source. At lateral positions, the difference with the real source is more important than at the front or the back.

Satongar et al. [2012] have shown that the mean ITD across all frequency bands (up to 800 Hz) and all azimuths for an hectagon of loudspeakers decreases with increasing order. This is also the case for the mean ILDs (up to 5 kHz). The corresponding numerical data is presented in Table 3.6(a). In addition, the same authors computed additional data for several off-centered positions. These are presented in Table 3.6(b). While the error is small at the sweet-spot, both ITDs and ILDs are biased at positions outside the sweet-spot.

Table 3.6: ITDs and ILDs averaged across frequency bands and azimuths for ambisonic reproduction systems with varying order [Satongar et al., 2012].

(a) At the sweet-spot.

Order M	Mean ITD error (µs)	Mean ILD error (dB)
1	32	3.2
2	6.6	2.2
3	1.4	1.4

(b) Average values of several off-center positions.

Order M	Mean ITD error (µs)	Mean ILD error (dB)
1	537	5.6
2	460	5.0
3	392	4.2

Wierstorf et al. [2013] compared the horizontal localization performance of WFS and HOA using dynamic binaural synthesis [Völk et al., 2008]. The simulated array of loudspeakers was circular, with a fixed radius of 3 m, and the number of loudspeakers in the array was varied with values of 53, 28, and 14. Participants pointed towards either a point source or the direction of a planar wave (white noise pulse signal). Sixteen positions of the participant were simulated, in the left half of the (circular) listening area. Compared to WFS, the localization performance of HOA was slightly worse, especially outside the central region of the listening area. The overall localization performance for the 56 loudspeaker array was 2.8° and 4.4° for the point source and the plane wave direction, respectively. The corresponding values for the WFS reproduction were 0.8° and 1.2°, respectively. Lowering the number of loudspeakers in the array led to the perception of more than one auditory event for some participants. Note that the results concerning HOA are deemed preliminary by the authors of the article, because only 3 participants took part in the experiment, whereas data from twelve participants was collected in WFS.

Elevation reproduction Power et al. [2012] compared the performance in vertical localization of 1st, 2nd, and 3rd order Ambisonics. For real sources, the participants in their panel, who were experienced listeners, reached a localization error in the median plane of about 10°. In average, the localization error of virtual sound sources was about 25° at 1st and 2nd order, and 15° at 3rd order Ambisonics.

Distance reproduction The paradigm shift allowing distance reproduction with Ambisonics started by considering point sources instead of plane waves [Zotter et al., 2009]. This resulted in the formulation of near-field compensated HOA (NFC-HOA) [Daniel, 2003].

Using the direct blind walking method, Kearney et al. [2012] compared the distance perception of real sources with the distance perception of sources rendered with a virtual 1st, 2nd, and 3rd order Ambisonics array. To simulate HOA with headphones, the perceptual approach described in [Merimaa and Pulkki, 2004] was used. The array consisted in an octagon at ear level and a cube (4 loudspeakers at the bottom and 4 at the top). Four distances were tested: 2, 4, 6, and 8 m. Results show that the distance to the source in 1st order Ambisonics reproduction matches the perception of the distance to the real source. No statistically significant improvement is observed with increasing order.

Bibliography

Altman, R., 1995. The sound of sound: A brief history of the reproduction of sound in movie theaters. Cineaste 21 (1-2), 68–71. 57

Bamford, J. S., Vanderkooy, J., Oct. 1995. Ambisonic sound for us. In: AES 99th Convention Preprint. New York.
http://www.aes.org/e-lib/browse.cfm?elib=7628 75, 76

Begault, D. R., Wenzel, E. M., Anderson, M. R., Oct. 2001. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. J. Audio Eng. Soc. 49 (10), 904–916.
http://www.aes.org/e-lib/browse.cfm?elib=10175 68

Benjamin, E., Heller, A., Lee, R., Nov. 2010. Why ambisonics does work. Audio Engineering Society.
http://www.aes.org/e-lib/browse.cfm?elib=15664 78

Berkhout, A. J., Dec. 1988. A holographic approach to acoustic control. J. Audio Eng. Soc. 36 (12), 977–995.
http://www.aes.org/e-lib/browse.cfm?elib=5117 70

Berkhout, A. J., de Vries, D., Vogel, P., May 1993. Acoustic control by wave field synthesis. J. Acoust. Soc. Am. 93 (5), 2764–2778.
http://dx.doi.org/10.1121/1.405852 72

Bertet, S., Daniel, J., Gros, L., Parizet, E., Warusfel, O., 2007. Investigation of the perceived spatial resolution of higher order ambisonics sound fields: A subjective evaluation involving virtual and real 3D microphones. In: Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments.
http://www.aes.org/e-lib/browse.cfm?elib=13925 79

Blauert, J., Jul. 2005. Communication acoustics. Springer.
http://dx.doi.org/10.1007/b139075 65, 66

Blumlein, A. D., Nov. 1936. Sound-transmission, sound-recording, and sound-reproducing system. 50, 51

Boone, M. M., Verheijen, E. N. G., Van Tol, P. F., 1995. Spatial sound-field reproduction by wave-field synthesis. J. Audio Eng. Soc 43 (12), 1003–1012.
http://www.aes.org/e-lib/browse.cfm?elib=7920 72

Bronkhorst, A. W., 1995. Localization of real and virtual sound sources. J. Acoust. Soc. Am. 98 (5), 2542–2553.
http://dx.doi.org/10.1121/1.413219 68

Brungart, D. S., Rabinowitz, W. M., 1999. Auditory localization of nearby sources. head-related transfer functions. J. Acoust. Soc. Am. 106 (3), 1465–1479.
http://dx.doi.org/10.1121/1.427180 68

Corteel, E., Nicol, R., May 2003. Listening room compensation for wave field synthesis. what can be done? Audio Engineering Society.
http://www.aes.org/e-lib/browse.cfm?elib=12330 74

Cox, T. J., D'Antonio, P., 2004. Acoustic Absorbers and Diffusers: Theory, Design and Application. Taylor & Francis. 70

Daniel, J., Jul. 2001. Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia. Ph.D. thesis, Université Paris 6, Paris. 75

Daniel, J., 2003. Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format. In: Audio Engineering Society Conference: 23rd International Conference: Signal Processing in Audio Recording and Reproduction.
http://www.aes.org/e-lib/browse.cfm?elib=12321 80

DellaSala, G., Sep. 2006. Introducing the 10.2 surround format. Audioholics. 62

Farrar, K., Oct. 1979. Soundfield microphone. Wireless World, 48–50. 77

Fletcher, H., Jan. 1934. Auditory perspective - basic requirements. Electrical Engineering, 9–11. 52

Gardner, W. G., Sep. 1997. 3-D audio using loudspeakers. Ph.D. thesis, Massachusetts Institute of Technology. 69

Gardner, W. G., Martin, K. D., 1995. HRTF measurements of a KEMAR. J. Acoust. Soc. Am. 97 (6), 3907–3908.
http://dx.doi.org/10.1121/1.412407 72

Gerzon, M. A., Nov. 1985. Ambisonics in multichannel broadcasting and video. J. Audio Eng. Soc. 33 (11), 859–871.
http://www.aes.org/e-lib/browse.cfm?elib=4419 75

Gerzon, M. A., Jun. 1994. Applications of Blumlein shuffling to stereo microphone techniques. J. Audio Eng. Soc. 42 (6), 435–453.
http://www.aes.org/e-lib/browse.cfm?elib=6939 51

Hamasaki, K., Hiyama, K., Nishiguchi, T., Okumura, R., 2006. Effectiveness of height information for reproducing the presence and reality in multichannel audio system. In: Audio Eng. Soc. Conv. 120.
http://www.aes.org/e-lib/browse.cfm?elib=13483 61

Hamasaki, K., Nishiguchi, T., Hiyama, K., Ono, K., 2004. Advanced multichannel audio systems with superior impression of presence and reality. In: Audio Eng. Soc. Conv. 116.
http://www.aes.org/e-lib/browse.cfm?elib=12756 61, 62

Hartmann, W. M., Wittenberg, A., Jun. 1996. On the externalization of sound images. J. Acoust. Soc. Am. 99 (6), 3678–3688.
http://dx.doi.org/10.1121/1.414965 68

ITU, Aug. 2012. Recommendation BS.775. Multichannel stereophonic sound system with and without accompanying picture. ITU-R. 51

Kearney, G., Gorzel, M., Rice, H., Boland, F., Jan. 2012. Distance perception in interactive virtual acoustic environments using first and Higher Order Ambisonic sound fields. Acta Acust. united with Acust. 98 (1), 61–71.
http://dx.doi.org/10.3813/AAA.918492 80

Kim, S., Wang, S., Dec. 2003. A Wiener filter approach to the binaural reproduction of stereo sound. J. Acoust. Soc. Am. 114 (6), 3179–3188.
http://dx.doi.org/10.1121/1.1624070 68

Langendijk, E. H. A., Bronkhorst, A. W., Jan. 2000. Fidelity of three-dimensional-sound reproduction using a virtual auditory display. J. Acoust. Soc. Am. 107 (1), 528–537.
http://dx.doi.org/10.1121/1.428321 68

Lee, H.-K., Rumsey, F., May 2004. Elicitation and grading of subjective attributes of 2-channel phantom images. In: Audio Eng. Soc. Conv. 116.
http://www.aes.org/e-lib/browse.cfm?elib=12718 53

Lentz, T., Apr. 2006. Dynamic crosstalk cancellation for binaural synthesis in virtual reality environments. J. Audio Eng. Soc. 54 (4), 283–294.
http://www.aes.org/e-lib/browse.cfm?elib=13677 68, 69

Litovsky, R. Y., Colburn, H. S., Yost, W. A., Guzman, S. J., 1999. The precedence effect. J. Acoust. Soc. Am. 106 (4), 1633–1654.
http://dx.doi.org/10.1121/1.427914 52

Mendonça, C., Campos, G., Dias, P., Vieira, J., Ferreira, J. P., Santos, J. A., Oct. 2012. On the improvement of localization accuracy with non-individualized HRTF-Based sounds. J. Audio Eng. Soc. 60 (10), 821–830.
http://www.aes.org/e-lib/browse.cfm?elib=16555 65

Merimaa, J., Oct 2007. Energetic sound field analysis of stereo and multichannel loudspeaker reproduction. In: Audio Eng. Soc. Conv. 123.
http://www.aes.org/e-lib/browse.cfm?elib=14315 52, 53

Merimaa, J., Pulkki, V., Oct. 2004. Spatial impulse response rendering. In: Proceedings of the 7th International Conference on Digital Audio Effects (DAFx '04). Naples, Italy. 80

Møller, H., Hammershøi, D., Jensen, C. B., Sørensen, M. F., Apr. 1995. Transfer characteristics of headphones measured on human ears. J. Audio Eng. Soc. 43 (4), 203–217.
http://www.aes.org/e-lib/browse.cfm?elib=7954 66, 67

Møller, H., Sørensen, M. F., Jensen, C. B., Hammershøi, D., Jun. 1996. Binaural technique: Do we need individual recordings? J. Audio Eng. Soc. 44 (6), 451–469.
http://www.aes.org/e-lib/browse.cfm?elib=7897 65

Parseihian, G., Katz, B. F. G., 2012. Rapid head-related transfer function adaptation using a virtual auditory environment. J. Acoust. Soc. Am. 131 (4), 2948–2957.
http://dx.doi.org/10.1121/1.3687448 65

Parseihian, G., Katz, B. F. G., Conan, S., Jun. 2012. Sound effect metaphors for near field distance sonification. In: Nees, M. A., Walker, B. N., Freeman, J. (Eds.), Proc. of the 18th Int. Conf. on Auditory Display. The International Community for Auditory Display, Atlanta, GA, USA, pp. 6–13.
https://smartech.gatech.edu/handle/1853/44435 68

Poletti, M. A., 2005. Three-dimensional surround sound systems based on spherical harmonics. J. Audio Eng. Soc 53 (11), 1004–1025.
http://www.aes.org/e-lib/browse.cfm?elib=13396 78

Power, P., Davies, W., Hirst, J., Dunn, C., 2012. Localisation of elevated virtual sources in higher order ambisonic soundfields. In: Proceedings of the Institute of Acoustics. Vol. 34. pp. 149–161. 80

Pulkki, V., Jun. 1997. Virtual sound source positioning using Vector Base Amplitude Panning. J. Audio Eng. Soc. 45 (6), 456–466. 62

Pulkki, V., 2001. Localization of amplitude-panned virtual sources II: two- and three-dimensional panning. J. Audio Eng. Soc 49 (9), 753–767.
http://www.aes.org/e-lib/browse.cfm?elib=10179 63, 64

Pulkki, V., Karjalainen, M., 2001. Localization of amplitude-panned virtual sources i: Stereophonic panning. J. Audio Eng. Soc 49 (9), 739–752.
http://www.aes.org/e-lib/browse.cfm?elib=10180 52, 54

Pulkki, V., Karjalainen, M., Huopaniemi, J., Apr. 1999. Analyzing virtual sound source attributes using a binaural auditory model. J. Audio Eng. Soc. 47 (4), 203–217.
http://www.aes.org/e-lib/browse.cfm?elib=12110 54

Rayleigh, L., 1896. The theory of sound. Vol. 2. Courier Dover Publications. 51

Satongar, D., Dunn, C., Lam, Y., Li, F., 2012. Psychoacoustic evaluation of spatial audio reproduction systems. In: Proceedings of the Institute of Acoustics. Vol. 34. pp. 77–86. 79, 261

Scientific American, Dec. 1877. The Talking Phonograph. Scientific American. 57

Snow, W. B., 1953. Basic Principles of Stereophonic Sound. Journal of the SMPTE 61, 567–587. 52, 54, 56

Sonke, J.-J., Oct. 2000. Variable acoustics by wave field synthesis. Ph.D. thesis, Delft University of Technology, Delft, The Netherlands. 74

Spors, S., Ahrens, J., Oct. 2008. A comparison of Wave Field Synthesis and Higher-Order Ambisonics with respect to physical properties and spatial sampling. In: Audio Eng. Soc. Conv. 125.
http://www.aes.org/e-lib/browse.cfm?elib=14708 71, 77

Start, E. W., Jun. 1997. Direct sound enhancement by Wave Field Synthesis. Ph.D. thesis, TU Delft, The Nederlands. 72

Steinberg, J. C., Snow, W. B., Jan. 1934. Auditory perspective - physical factors. Electrical Engineering, 12–17. 52

Streicher, R., Everest, F. A., 2006. The new stereo soundbook, 3rd Edition. Audio Engineering Associates, Pasadena. 51, 54, 55, 56, 60

Theile, G., 1990. On the performance of two-channel and multi-channel stereophony. In: Audio Eng. Soc. Conv. 88.
http://www.aes.org/e-lib/browse.cfm?elib=5807 57

Theile, G., Feb 2000. Multichannel natural recording based on psychoacoustic principles. In: Audio Eng. Soc. Conv. 108.
http://www.aes.org/e-lib/browse.cfm?elib=9182 50, 61

Van Daele, B., Van Baelen, W., Feb. 2012. Productions in AURO-3D. Tech. Rep. Rev 0.6, AURO Technologies NV. 61

Verheijen, E. N. G., 1998. Sound reproduction by Wave Field Synthesis. Ph.D. thesis, Delft University of Technology. 73

Vogel, P., Dec. 1993. Application of Wave Field Synthesis in room acoustics. Ph.D. thesis, Delft University of Technology, Delft, The Netherlands. 72

Vorländer, M., 2008. Auralization. Springer.
http://dx.doi.org/10.1007/978-3-540-48830-9 69

Völk, F., Konradl, J., Fastl, H., Jul. 2008. Simulation of wave field synthesis. In: Acoustics 2008 Paris. Paris, France, pp. 1165–1170.
http://scitation.aip.org/content/asa/journal/jasa/123/5/10.1121/1.2933196 79

Wallach, H., Newman, E. B., Rosenzweig, M. R., Jul. 1949. The precedence effect in sound localization. The American Journal of Psychology 62 (3), 315–336.
http://dx.doi.org/10.2307/1418275 52

Ward, D. B., Abhayapala, T., 2001. Reproduction of a plane-wave sound field using an array of loudspeakers. IEEE Transactions on Speech and Audio Processing 9 (6), 697–707. 78

Wichmann, F. A., Hill, N. J., Nov. 2001. The psychometric function: I. Fitting, sampling, and goodness of fit. Perception & Psychophysics 63 (8), 1293–1313. 73

Wierstorf, H., Raake, A., Spors, S., 2013. Localization in Wave Field Synthesis and Higher Order Ambisonics at different positions within the listening area. In: Proceedings of the AIA-DAGA 2013 Conference on Acoustics. Meran, Italy. 79

Wightman, F. L., Kistler, D. J., Feb. 1989. Headphone simulation of free-field listening. II: psychophysical validation. J. Acoust. Soc. Am. 85 (2), 868–878. http://dx.doi.org/10.1121/1.397558 68

Wightman, F. L., Kistler, D. J., 1992. The dominant role of low-frequency interaural time differences in sound localization. J. Acoust. Soc. Am. 91 (3), 1648–1661. http://dx.doi.org/10.1121/1.402445 74

Wightman, F. L., Kistler, D. J., 1999. Resolution of front–back ambiguity in spatial hearing by listener and source movement. J. Acoust. Soc. Am. 105 (5), 2841–2853. http://dx.doi.org/10.1121/1.426899 68

Wittek, H., Oct. 2007. Perceptual differences between wavefield synthesis and stereophony. Ph.D. thesis, University of Surrey, UK. 73, 74

Zahorik, P., Wightman, F., Kistler, D., 1995. On the discriminability of virtual and real sound sources. In: Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics. New Paltz, NY, USA, pp. 76–79. http://dx.doi.org/10.1109/ASPAA.1995.482951 68

Zotter, F., Pomberger, H., Frank, M., 2009. An alternative ambisonics formulation: Modal source strength matching and the effect of spatial aliasing. In: Audio Eng. Soc. Conv. 126. http://www.aes.org/e-lib/browse.cfm?elib=14936 75, 77, 80

This page intentionally left blank.

Chapter

4

Principles and limitations of stereoscopic-3D visualization

Highlights

✓ The principles of s-3D visualization are described.

✓ The differences between natural viewing and s-3D viewing are discussed, as well as the potential consequences of these differences.

Contents

This chapter describes in more detail the principles of stereoscopic 3D (s-3D) visualization (Section 4.1). The whole processing chain is considered, from capture to transmission, and reproduction. Then, in Section 4.2, the differences between natural viewing and s-3D viewing are discussed, as well as the potential consequences of these differences.

4.1 Principles of stereoscopic-3D imaging

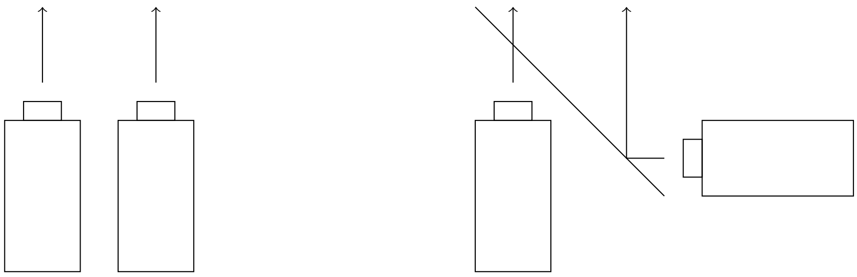
By essence, the illusion of depth perception in s-3D cinema is created by presenting a different image to each eye. Each image is physically located on the screen and all the viewers in the room look at the same image pair. In this section, we discuss how these two images can be captured, stored, transmitted digitally, and reproduced to the spectators. The reader uninterested with technical aspects can safely skip directly to the section discussing s-3D perception (Section 4.1.4).

4.1.1 Stereoscopic-3D capture

The digital capture of an s-3D image requires the use of two synchronized 2D image sensors. Ideally, the optical systems will be placed side by side, with the lenses separated by a distance similar to that between the eyes of an average person (Figure 4.1(a)). In a review by Dodgson [2004], this *interocular distance* has a mean value of 63 mm. The distance between the two camera lenses is called the *interaxial distance*. For example, in the Panasonic AG-3DA1 integrated s-3D camera, the two lenses are separated by approximately 60 mm.

However, it is not always possible to integrate the equivalent of two cameras into one. This might be because the optical systems are too large to approach the mean human interocular distance, or because one wants to control the interaxial distance for artistic reasons. The interaxial distance controls the amount of disparity of each depth plane (see Section 6.3 for the mathematical development). The greater the interaxial distance, the greater the disparity at a given depth. The solution to fine-tune the interaxial distance is to mount two regular 2D cameras on a rig. Usually, one camera in the pair is oriented directly towards the point of interest in the scene, and the other is oriented perpendicular to that direction. A beam-splitter then transmits half of the incident light to one image sensor, and reflects the other half to the second sensor (Figure 4.1(b)).

It is also possible to generate an s-3D image pair from one view only, through a process called 2D-to-3D conversion. However, the discussion of this process is outside the scope of this thesis.



(a) Side-by-side camera. (b) Beam-splitter rig camera.

Figure 4.1: Dual camera stereoscopy: two different s-3D cameras obtained from two regular 2D cameras. After [Lipton, 1982].

4.1.2 Stereoscopic-3D transmission

The movie file nowadays is transmitted digitally. In cinemas, the digital equivalent of the optical film is the Digital Cinema Package (DCP*).

For use with a PC, a stereoscopic video file can be created in a large number of ways. The simplest way, which can be achieved using regular 2D video encoding, is to encode each eye into a separate video file, and possibly combine them in a container like matroska† or MPEG's MP4 (MPEG-4 part 14). However, this method is highly inefficient in terms of storage, because a lot of information in one eye can be found in the other. Therefore, an encoding that specifically supports s-3D (or more generally multiview coding) reduces the file size.

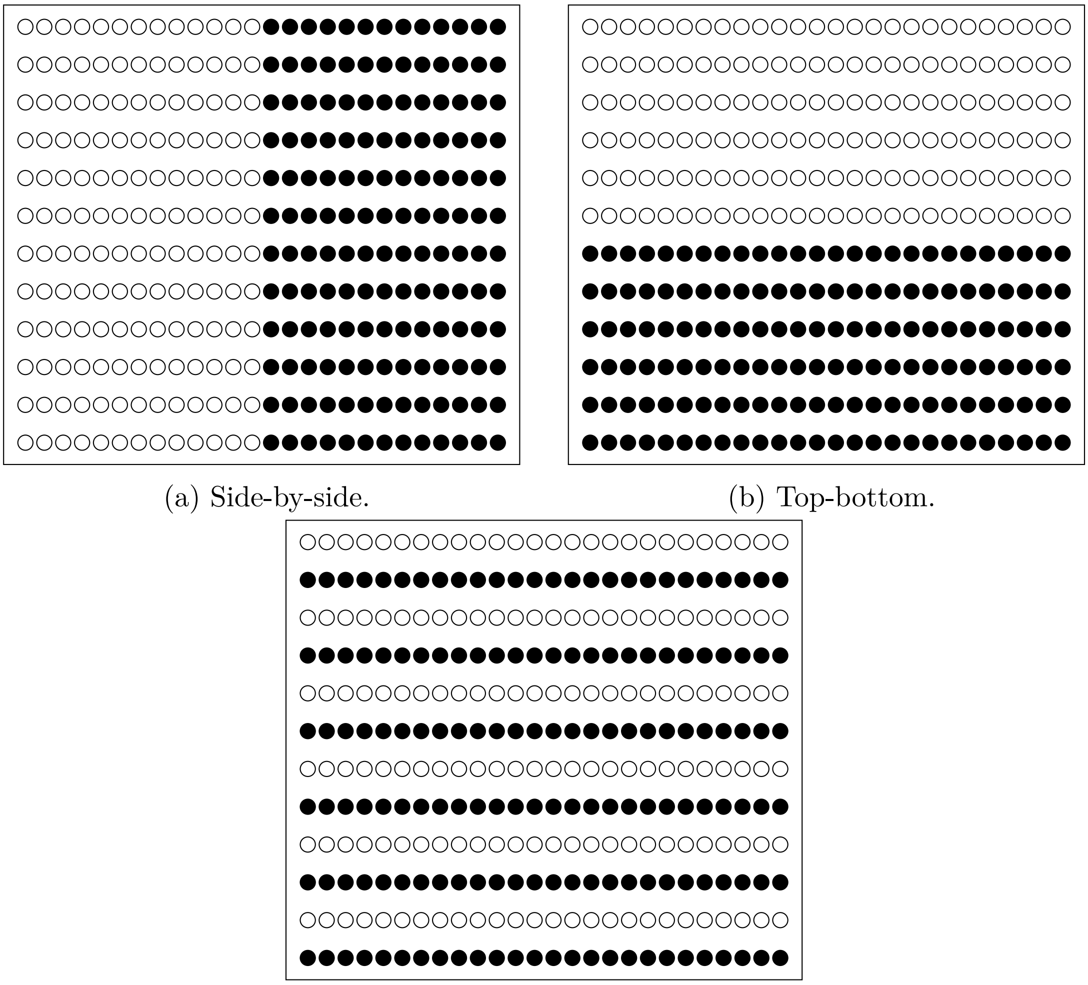
The codec H.264 (MPEG-4 part 10) supports multiview video coding, with its Annex H. A pair of corresponding 2D videos can be encoded in a single s-3D stream with double resolution. Half of the pixels in one frame of this resulting stream represents the left view and the other half represents the right view. One can visualize this by playing the s-3D stream with a regular 2D video player. The pixels from the left and right views are ordered according to a given rule. Common rules include side-by-side, top-bottom, and row-interleaved. These rules are illustrated in Figure 4.2. The s-3D stream might be reduced to the same resolution as the original 2D videos, such as 1280 × 720‡ or 1920 × 1080§. This results in a loss of half the resolution in each eye. At the same time, the s-3D video can be processed by certain hardware as if it were a regular 2D video.

*http://dcimovies.com/specification/index.html, last accessed 30/09/2013

†matroska.org, last accessed 30/09/2013

‡Also known as 720p.

§Also known as 1080p.



(a) Side-by-side.

(b) Top-bottom.

(c) Row-interleaved.

Figure 4.2: Common s-3D frame-compatible formats where a white circle represents the sample from one view and a black circle represents the sample from the other view. After [Vetro et al., 2011].

4.1.3 Stereoscopic-3D reproduction to the spectator

In the movie theater, the projection system must "bring" the two different images to the correct eye of the viewer. Three different multiplexing techniques dominate the cinema market for s-3D reproduction, namely polarization multiplexing, time multiplexing, and wavelength multiplexing.

Polarization multiplexing Inexpensive polarization filters can effectively separate two light beams with different polarizations. However, polarization technology requires a special type of screen (silver screen or aluminized screen) which preserve the polarization of the incident light. Linear polarization is used in IMAX 3D theaters and circular polarization is used in theaters equipped with RealD 3D or MasterImage systems. Circular polarization has the advantage over linear polarization that the user can rotate his head around its front/back axis without loosing any light intensity. It is interesting to note that s-3D in movie theaters can be obtained with a single projector, and the polarization of the images is time-sequential, thanks to a special apparatus (a modulator called the ZScreen for RealD 3D, and a polarization wheel for MasterImage).

Time multiplexing Time multiplexing is achieved through high frame rate projectors and active liquid crystal shutters in the glasses. The projector alternates between left views and right views faster than the frame rate of the movie itself, which is usually 24 frames per second or, very recently, 48. In a technique called *triple flash*, each view is projected three times, alternating the left and right view, in the lapse of one movie frame. Accordingly, the glasses obstruct the light in one eye and let it pass in the other. Synchronization with the projector is achieved using an infrared or a radio signal. This solution is commercialized for theaters by the company XpanD.

Wavelength multiplexing The idea behind wavelength multiplexing is to use different set of filters for the left image and the right image, at the projector, so that each of the red, green and blue components has a different wavelengths corresponding to each eye. The glasses are the complement of the filters at the projectors. This system is passive and has the advantage over polarized light that it does not require a polarizing screen. Dolby has commercialized this patented technology, developed by Infitec, under the brand Dolby 3D.

The probability to encounter one technology or another depends on the location of a given s-3D theater. According to data from Jones [2009], the northern American market is dominated by RealD polarized systems (more than 75% of the equipped rooms) while the European market is split between the different technologies: around 50% of theaters are equipped with RealD polarized systems, and the other 50% are almost equally equipped with either XpanD or Dolby 3D systems.

The reproduction equipment is responsible for two important image artifacts, namely crosstalk and flicker. These two image artifacts are discussed in Appendix B.

4.1.4 Perception of stereoscopic-3D images

The illusion of depth perception in s-3D images is created by presenting a different image to each eye. Each image is physically located on the screen. From the psycho-optical point of view, the perception of such a visual presentation can be analyzed in a geometrical fashion.

While we leave the description of a complete geometrical model to Section 6.3, we review here the effect of one important quantity: the *parallax*. The (screen) parallax is the difference in distance between points in the left and right images corresponding to the same point in the captured 3D space. The relative position of corresponding points, together with the position of the spectator's eyes, geometrically define the position of the visual percept. The geometrical approach tells us to trace a ray from each eye to its intended point on the screen. The visual percept lies at the intersection of the two rays, when there is one.

We now discuss the position of the visual percept depending on the relative position of the corresponding points. This discussion is illustrated in Figure 4.3.

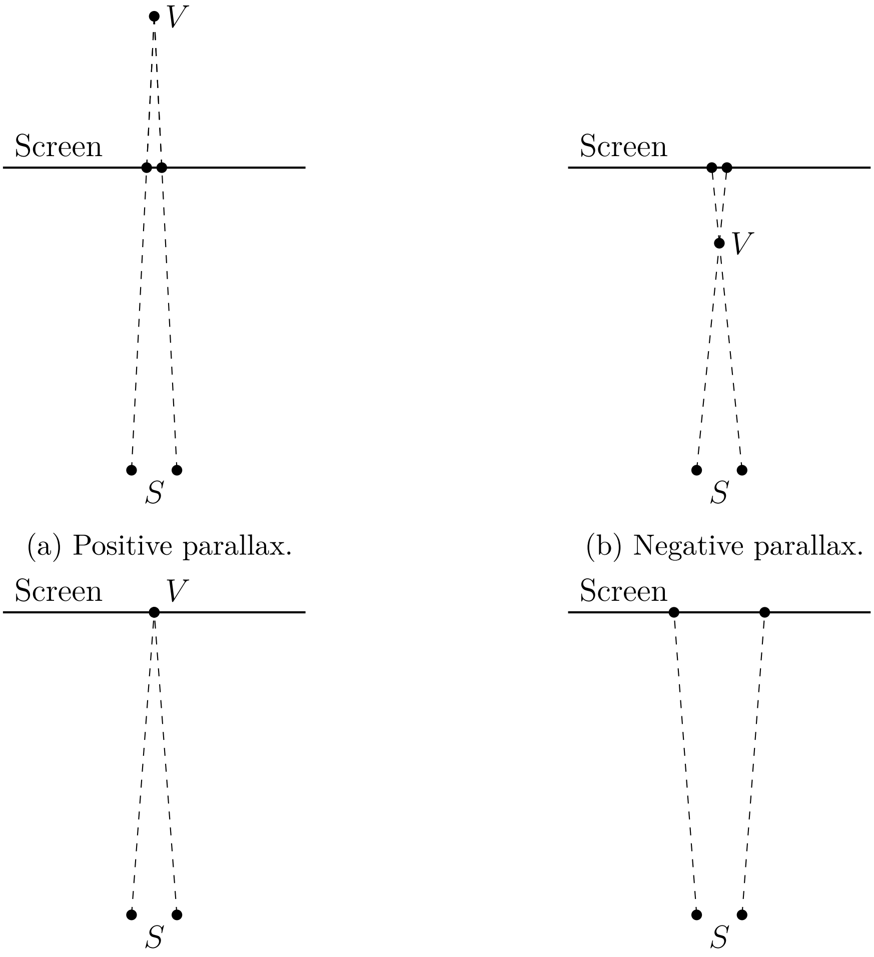
- When the point in the right image is to the right of the corresponding point in the left image (positive parallax, Figure 4.3(a)), the visual percept lies behind the screen plane. However, there is a limit to the positive parallax that can be fused under the assumption of the geometrical model. When the parallax is equal to the interocular distance, the light rays are parallel, and the percept lies at infinity. When the parallax is greater than the interocular distance, the rays do not intersect, and the stereoscopic stimulus is not fused (Figure 4.3(d)). Divergence and, consequently, double vision occur.

- When the point in the right image is to the left of the corresponding point in the left image (negative parallax, Figure 4.3(b)), the visual percept lies in front of the screen plane.

- When the point in the right image coincides with the corresponding point in the left image (zero parallax, Figure 4.3(c)), the visual percept lies in the screen plane.

4.2 Limitations of stereoscopic-3D imaging

With the recent increase of interest in s-3D technologies, concerns were raised about potential adverse effects associated with the prolonged viewing of s-3D



(a) Positive parallax.

(b) Negative parallax.

(c) Zero parallax.

(d) Divergence.

Figure 4.3: A spectator at S (with his/her two eyes represented by the two dots) perceives one s-3D object at the location V, resulting from four different parallax values: (a) positive and less than the interocular distance, (b) negative and less than the interocular distance in absolute value, (c) zero, and (d) greater than the interocular distance. After [Lipton, 1982].

images, like in s-3D movies. S-3D viewing is believed to cause *visual fatigue*, a decrease in performance of the human visual system, objectively measurable. Its perceived counterpart is called *visual discomfort*, and *asthenopia* is the all-encompassing medical term [Lambooij et al., 2009].

Several features of an s-3D stimulus differ from the equivalent real world stimulus. One can logically hypothesize that the cause(s) of s-3D related symptoms is (are) to be found among them. Yano et al. [2004] and Lambooij et al. [2009] list these potential causes of visual fatigue:

- viewer's anomalies of binocular vision (stereoblindness, interocular distance, age.)

- geometrical errors. This includes keystone distortion and wrong choices in camera parameters;
- a conflict between accommodation and vergence eye movement;
- excessive screen parallax may lead to divergence, depending on the viewer's interocular distance.

Some of these potential sources have received considerable attention in the literature and therefore will be reviewed in separate sections. First, however, we review the large scale studies on symptoms reported after s-3D prolonged viewing.

4.2.1 Side effects of stereoscopic-3D movies

In a self-administered survey, participants were asked to recall the last s-3D movie they had seen [Solimini et al., 2012]. 66.4% of the 907 participants in this survey reported the experience of at least one symptom. Of these, 60.4% reported the symptom during the movie, 43.2% right after, and 15.3% two hours after the movie ended. These symptoms were mild, with tired eyes and headaches being most often reported, by 34.8% and 13.7% of the participants, respectively.

Solimini [2013] compared the intensity and frequency of visually induced motion sickness (VIMS) resulting from viewing either a 2D or an s-3D content. VIMS is the feeling of self-motion induced by the movement in an image. The symptoms are the consequence of a mismatch between the visual stimulus, the proprioceptive stimulus, and the vestibular stimulus. VIMS can produce symptoms similar to those of motion sickness, in absence of true movement. VIMS is believed to share some biological foundations with visual fatigue [Wilkins and Evans, 2010]. The 497 participants in this study were asked to go see two movies (one 2D and the other s-3D) of their choice, at the cinema of their choice, at two different days in a 3-week period. They answered a paper questionnaire both before and after the viewing. The questionnaire consisted in (1) socio-demographic questions, questions relating to possible predictors for VIMS, such as headache history, car sickness history, . . . , and (2) the simulation sickness questionnaire (SSQ) [Kennedy et al., 1993]. The SSQ is a 16-item symptom checklist which measures the VIMS. The SSQ score can be further divided into three subscales: `nausea` (measuring gastrointestinal stress), `oculomotor` (measuring disturbances of the visual system), and `disorientation` (measuring disturbances of the vestibular system). 38.1% of the participants obtained an SSQ score superior to 20, which indicates discomfort. Viewing an s-3D movie also increased the SSQ in a much larger proportion than viewing a 2D movie. The SSQ score increased from the baseline by a factor of 8.8 and 2, respectively. The symptoms, however, were self reported and the discomfort quickly wore off, once the s-3D glasses were removed. A history of headaches and car sickness also seem to contribute to the increase in SSQ score after seeing a movie, 2D or s-3D. Women, who are more susceptible to these conditions, are therefore more prone to VIMS. As Howarth [2011] notes about VIMS, however, what is attributed here to s-3D viewing may well be attributed to simple motion in the images. Indeed, we verified this simply by consulting the movie type classification of the website IMDb. 79.9% of the people in the study saw an s-3D adventure (and comedy) animation movie ("Puss in boots"), while only 19.9% of the participants saw a 2D adventure (and action) movie ("Sherlock Holmes: a games of shadow".) All other 2D movies seen were comedies. Therefore, the difference in results might be a consequence of the comparison between comedies and action movies, rather than the difference between 2D and s-3D viewing. Further research is therefore needed to evaluate which proportion of the population suffers from s-3D induced symptoms. The study should be based on the comparison of the same movie in 2D and s-3D.

4.2.2 The vergence-accommodation conflict

We first describe the processes of accommodation and vergence in more detail. As pointed out by Shibata et al. [2011], a distinction must be made between motor and sensory aspects of accommodation and vergence. The situation is described in Table 4.1.

Table 4.1: The distinction between sensory and motor aspects of vergence and accommodation.

	Sensory stimulus	Motor response	Tolerance range
Accommodation	Blur	Adjustment of the power of crystalline lens	Depth of focus
Vergence	Binocular disparity	Eye rotations	Panum's fusional area

The stimulus to accommodation is blur, which is corrected by adjusting the optical power of the crystalline lens. When one focuses on an object, it is expected that an object closer or further will appear blurred. Within a certain range, however, the defocus on the retina does not induce a detectable blur. This range is called the *depth of focus* when it is measured on the retina, and *depth of field* when it is projected back in the object space. The depth of focus extends to about ±0.3 diopters[*] [Campbell, 1957]. The depth of focus is the vertical extent of the orange region in Figure 4.4, varying with the vergence (i.e. simulated) distance.

The stimulus to vergence is the binocular disparity. When the disparity is too large, the eyes rotate to keep the object fused. Still, when two points are horizontally separated by at most a quarter degree (or 15 arcmin) of visual angle, they still appear fused. This region in space that permits visual fusion is called *Panum's area*. It gives the *horopter*, the geometrical locus of points which activates the same corresponding points on the retina, a thickness in space. The width of Panum's area is around one-tenth the width of the depth of focus (in diopters) [Valois, 2000]. We are therefore more sensitive to an error in disparity than we are to an error in focus. Panum's area is the horizontal extent of the yellow region in Figure 4.4.

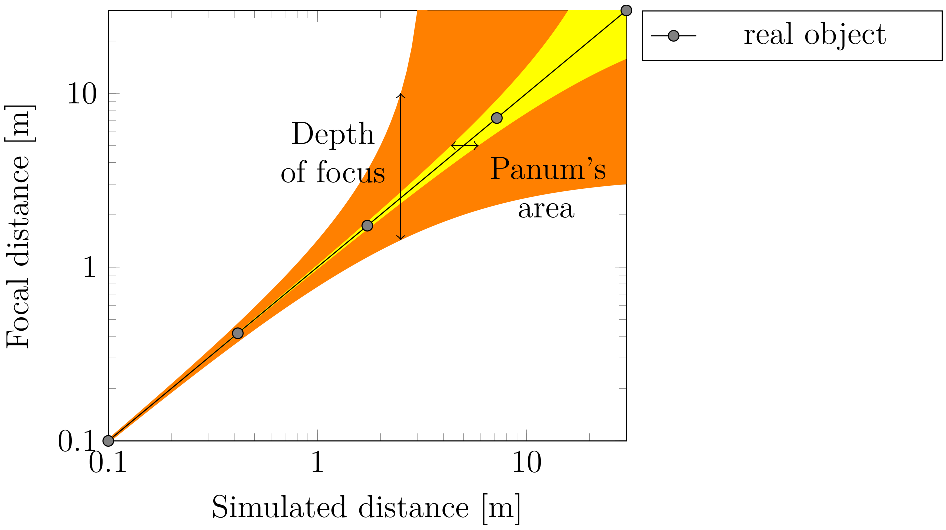


Figure 4.4: The vergence-accommodation coupling. A real object is simulated, in such a way that the vergence and the focal distance specified by the object are equal. In the orange region, an error in focus goes undetected. The vertical length of the orange region is the depth of focus (at a given simulated distance). In the yellow region, an object with non-zero disparity may still appear fused. The horizontal length of the yellow region is Panum's fusion area. After [Hoffman et al., 2008].

Accommodation, vergence, and their coupling can be modeled as two dual

[*] Diopters are inverse meters.

parallel feedback control systems that interact via cross-links (see Figure 4.5). In natural viewing conditions, the interaction of accommodation and vergence produces a clear single image. Blur does not drive the accommodation process while the defocus is within the depth of focus. Similarly, small binocular disparities, within Panum's area, do not drive the vergence system. When the object moves closer, the controllers are driven by the excess in either defocus or disparity. In the model of Figure 4.5, the motoric outputs are the sum of the controller output, the target distance, the tonic component, and the cross-link between the two systems. A negative feedback loop allows the system to find a stable state. The tonic component is the part of the system that adapts slowly to the stimulus. It can potentially activate a response during the entire stimulus duration.

An artificial stereoscopic stimulus can require a decoupling of the accommodation and the vergence. However, to maintain the fusion of the stimulus, an effort is required, which is not always possible. The *zone of clear single binocular vision* is the range of accommodation and vergence that can be achieved without excessive error in either. It is represented as the orange region in Figure 4.6. The range of accommodation and vergence that can be achieved without discomfort is called *Percival's zone of comfort*. It is represented as the yellow region in Figure 4.6. As a rule of thumb, its width is about one-third the width of the zone of clear single binocular vision (in diopters) [Hoffman et al., 2008].

The fusion effort associated to an excessive binocular parallax causes oscillations in the outputs of the two parallel systems. Excessive binocular parallax also increases the reported visual fatigue [Emoto et al., 2004]. This is why Lambooij et al. [2009] argues for a 1° limit for the disparity on the screen. Yano et al. [2004] also indicated that, in addition to stimuli that drive the visual system outside of its comfort zone, stimuli including a movement in the depth direction can also increase visual fatigue, even when the stimuli stays inside the comfort zone.

The vergence-accommodation conflict alters the perceived depth and causes fatigue [Hoffman et al., 2008]. However, as can be seen in Figure 4.6, any s-3D image further than 3 m from the viewer in a cinema theater produces accommodation and vergence stimuli that are comfortable. Therefore, this limitation of s-3D images might not be the cause of the visual discomfort reported by spectators in movie theaters [Howarth, 2011].

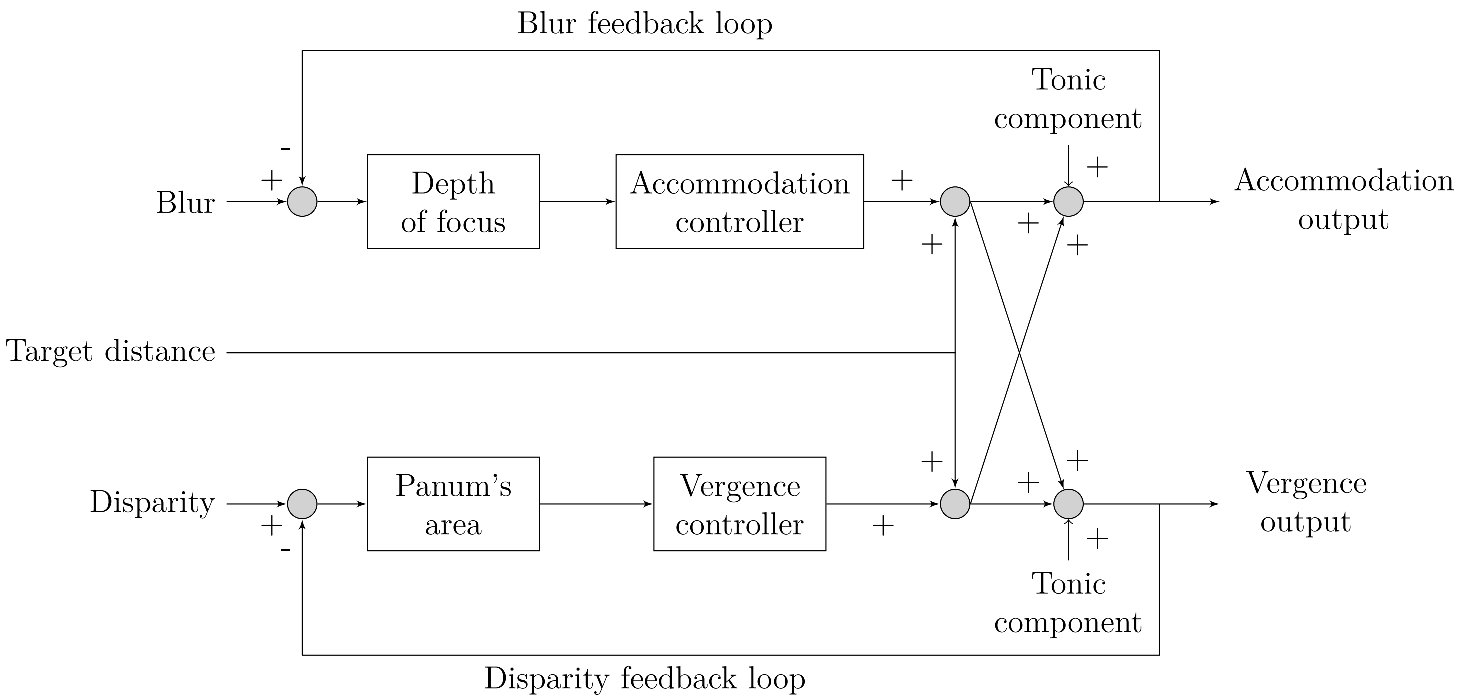


Figure 4.5: Accommodation and vergence modeled as two dual parallel feedback control systems that interact via cross-links. After [Lambooij et al., 2009; Schor and Kotulak, 1986].

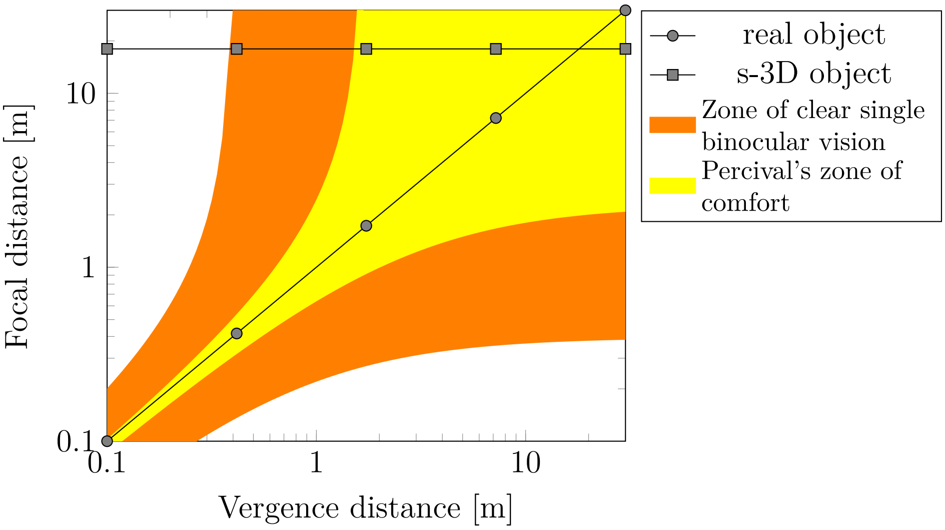


Figure 4.6: The decoupling of accommodation and vergence. An artificial stereoscopic stimulus can require a decoupling of the accommodation and the vergence. Therefore, the fusion of the stimulus requires an effort and is not always possible. The orange region represents the zone of clear single vision: the range of vergence and accommodative responses that subjects can achieve without experiencing blur and/or double vision. The yellow area represents Percival's zone of comfort: the range of responses viewers can achieve without discomfort. The circles represent a real world stimuli at five different distances, while the squares represent its s-3D image seen at 18 m from the screen. The focal distance, in the case of s-3D viewing, is the distance to the screen and the vergence distance is the simulated distance to the object of attention. After [Shibata et al., 2011].

4.2.3 The focus of the image

When the eyes examine different parts of the s-3D display, the objects at different depths do not produce varying blur because optically, the whole image is either in or out of focus. Therefore, the accommodation system receives an information indicating flatness of the scene.

In addition, in natural viewing conditions, objects located both in front and behind the object of attention are out of focus on the retina. The defocus of objects surrounding the attended object induces the perception of blur which increases with the distance to the attended object. However, on an image captured by a camera, the whole scene is usually captured in focus. Therefore the whole scene appears sharp on the retina and also indicates flatness.

Combined together, these two cues clearly affect the perception of depth [Watt et al., 2005]. However, it remains uncertain whether these can cause discomfort [Howarth, 2011]. With a display producing accurate focus cues, Akeley et al. [2004] informally observed that the visual comfort was better, compared to a traditional s-3D display.

4.2.4 Geometrical aspects

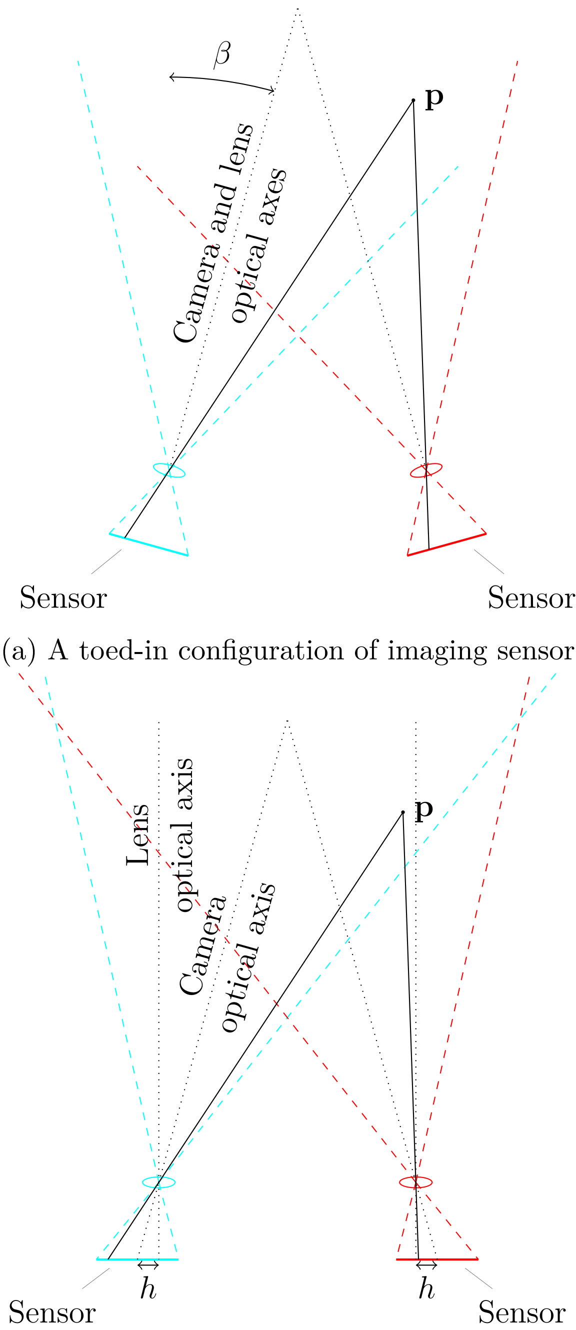
Vertical misalignment of the two images Amongst the possible geometrical errors, the vertical misalignment between the left and right image is an important factor contributing to the comfort of s-3D images. Ideally, the alignment should be perfect and the vertical shift should never be greater than 1.5% of the height of the image [Kooi and Toet, 2004; Yamanoue et al., 1998].

Toed-in and parallel camera configuration There are two principle means for creating an s-3D camera from two regular 2D imaging sensors: the toed-in camera configuration, and the parallel camera configuration [Woods, 1993]. In the toed-in camera configuration, the cameras are rotated from parallel by an angle β to achieve convergence (Figure 4.7(a)). In the parallel camera configuration, the center of each imaging sensor is shifted by a distance h away from the optical axis of the lens to achieve convergence (Figure 4.7(b)).

However, several geometrical distortions are associated to the toed-in configuration. Planes of constant depth captured by a toed-in camera are reproduced as curved. The objects close to the sides appear further away than they should be. This can also lead to unnatural depth variations of the scene when the camera pans parallel to the display. The cause of the curvature of planes of constant depth is keystone: the superposition of the two 2D images of a grid is not a grid. This is illustrated in Figure 4.8. Keystone causes unwanted horizontal as well as vertical disparities. Therefore, s-3D images generated with a toed-in camera can be uncomfortable to watch. On the contrary, a parallel camera does not produce any keystone, and is therefore favored when shooting in s-3D.

The influence of the viewpoint In a movie theater, not all spectators sit exactly in front of the middle of the screen: some sit off-axis, and the distance from the seat to the screen varies also. We review here the potential consequences of viewing an image from an unintended point of view.

The 2D camera allows to capture on a plane, its imaging sensor (CCD, CMOS, ...), a projection of a scene. Geometrically, one can trace a light ray from each



(a) A toed-in configuration of imaging sensors.

(b) A parallel configuration of imaging sensors.

Figure 4.7: Layout of imaging sensors for (a) a toed-in configuration, and (b) a parallel configuration. After Woods [1993].

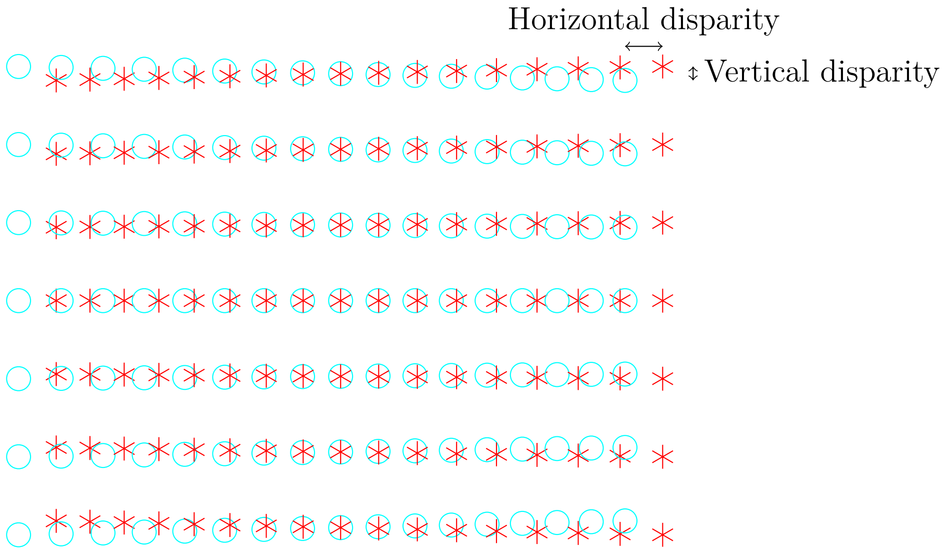


Figure 4.8: The left (circles) and right (asterisks) images in the s-3D pair resulting from a toed-in camera filming a regular grid pattern. This figure illustrates the keystone inherent to the toed-in configuration. Ideally, each asterisk in the right image should be colocated with the circle at the corresponding point in the left image. Because of keystone, however, unwanted horizontal and vertical disparities appear at the borders of the image.

pixel of the sensor, make it pass through a particular point, called the *camera center*, and record what the ray “hits” in the scene. This reasoning is accurate when the image plane is located behind the camera center, and the image of the scene is therefore upside-down. In Figure 4.9, the image plane has been placed between the camera center and the scene, which allows to recover a scaled, but not inverted, version of the scene.

Mathematically, this is called a linear perspective of the scene. As Goldstein [2005] notes, this method was rediscovered in the Renaissance as a way to duplicate reality:

> This isomorphism between a depiction in linear perspective and the retinal image means that when a perspective picture is viewed from the correct station point (with the viewer’s line of sight perpendicular to the picture and the viewer positioned at a distance so the visual angle of the picture matches the visual angle of the original scene as originally viewed by the artist), the picture will duplicate the image that the original scene cast on the artist’s retina. (Goldstein [2005])

When the viewer’s eyes are not at the correct station point, however, the reti-

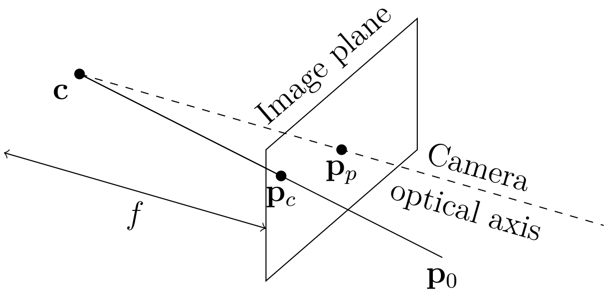


Figure 4.9: The (perspective) projection of a scene by a 2D camera. After Hartley and Zisserman [2004].

nal image suggests a scene with a different layout, but the viewer still experiences the scene in the same way. To study this effect, Vishwanath et al. [2005] presented an ovoid shape to subjects on a display which could be rotated, so that the subject was not always at the correct viewpoint. The subjects were asked to report whether the object was too wide or too narrow relative to its height to be a sphere. The researchers also conducted an experiment with a slanted plane, asking the subjects to report if the plane was too wide or too narrow to be a square. When the subjects could see the display with its frame binocularly, the aspect ratio settings were invariant over the viewing angle, even at positions |45°| away from the correct viewpoint. The results were consistent with the local-slant hypothesis. The perceptual invariance in this hypothesis is not achieved by recovering the actual center of projection, but rather by adjusting the retinal-image shape based on an evaluation of the local slant of the picture plane at the point of interest.

Following up on this experiment, Banks et al. [2009] considered the same problem, but now when viewing an s-3D image. They argued that the approach given by Woods [1993] makes a strong assumption by considering that the visual percept is not corrected for the viewpoint. Instead, it heavily relies on geometry, as will be seen in Section 6.3. This assumption had not yet been evaluated. They reproduced the first experiment using an s-3D hinge stimulus with a 90° angle. The geometrical approach indicates that the value of the hinge angle can be perceived as lower or higher, depending on the position of the viewer with respect to the correct center of projection. Following a “psychophysical procedure”, the details of which are not given in the article, the results show that viewers do not compensate for their incorrect viewpoint. Instead, the geometrical approach accounts fairly well for the results.

The two previously described experiments considered a movement of rotation of the display with respect to the viewer, so that the viewing distance was kept constant. Pollock et al. [2012] investigated the visual distortions perceived by subjects located in front or behind the center of projection. They also investigated the compromise that pairs of subjects made with regards to their respective perception when they were located at different places. Subjects had to verbally ask the experimenter to adjust the depth of a rectangle at their feet until it matched the perceived depth of a rectangle 4.9 m in front of them. The results were similar when the subjects were alone or working in pairs. After correcting for the compression of depth, which was around 86.3%, the researchers found that the distortions in depth were smaller than predicted by the geometrical model, and asymmetrical. The results were better predicted when at least one of the subjects was close to the screen. Also, when subjects worked in pairs, the time needed to reach an agreement increased with the perceived distortions.

As far as visual fatigue is concerned, Howarth [2011] notes that, even if an image may appear incorrect, there is little chance that an incorrect viewpoint will cause any physiological effect.

Bibliography

Akeley, K., Watt, S. J., Girshick, A. R., Banks, M. S., 2004. A stereo display prototype with multiple focal distances. In: ACM SIGGRAPH 2004 Papers. SIGGRAPH '04. ACM, New York, NY, USA, p. 804–813.
http://dx.doi.org/10.1145/1186562.1015804 102

Banks, M. S., Held, R. T., Girshick, A. R., Jan. 2009. Perception of 3-D layout in stereo displays. Information display 25 (1), 12–16.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3115721/ 105

Campbell, F., 1957. The depth of field of the human eye. Optica Acta: International Journal of Optics 4 (4), 157–164.
http://dx.doi.org/10.1080/713826091 98

Dodgson, N. A., May 2004. Variation and extrema of human interpupillary distance. In: Woods, A. J., Merritt, J. O., Benton, S. A., Bolas, M. T. (Eds.), Proc. SPIE 5291. San Jose, CA, pp. 36–46.
http://dx.doi.org/10.1117/12.529999 90

Emoto, M., Nojiri, Y., Okano, F., Aug. 2004. Changes in fusional vergence limit

and its hysteresis after viewing stereoscopic TV. Displays 25 (2–3), 67–76.
http://dx.doi.org/10.1016/j.displa.2004.07.001 99

Goldstein, E. B., 2005. Pictorial perception and art. In: Goldstein, E. B. (Ed.), Blackwell Handbook of Sensation and Perception. Blackwell Handbooks of Experimental Psychology. Blackwell Publishing, Oxford, UK.
http://dx.doi.org/10.1111/b.9780631206842.2005.00011.x 104

Hartley, R., Zisserman, A., 2004. Multiple View Geometry in Computer Vision, 2nd Edition. Cambridge University Press. 105

Hoffman, D. M., Girshick, A. R., Akeley, K., Banks, M. S., Mar. 2008. Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. Journal of Vision 8 (3).
http://dx.doi.org/10.1167/8.3.33 98, 99

Howarth, P. A., Mar. 2011. Potential hazards of viewing 3-D stereoscopic television, cinema and computer games: a review. Ophthalmic & Physiological Optics 31 (2), 111–122.
http://dx.doi.org/10.1111/j.1475-1313.2011.00822.x 97, 99, 102, 106

Jones, C., Dec. 2009. Roll-out and financial aspects of 3D digital cinema. In: Proceedings of 3D Stereo MEDIA 2009. Liège, Belgium. 93

Kennedy, R. S., Lane, N. E., Berbaum, K. S., Lilienthal, M. G., Jul. 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. International Journal of Aviation Psychology 3 (3), 203. 96

Kooi, F. L., Toet, A., Aug. 2004. Visual comfort of binocular and 3D displays. Displays 25 (2–3), 99–108.
http://dx.doi.org/10.1016/j.displa.2004.07.004 102

Lambooij, M., Fortuin, M., Heynderickx, I., IJsselsteijn, W., 2009. Visual discomfort and visual fatigue of stereoscopic displays: A review. Journal of Imaging Science and Technology 53 (3), 30201–1–30201–14.
http://dx.doi.org/10.2352/J.ImagingSci.Technol.2009.53.3.030201
95, 99, 100

Lipton, L., 1982. Foundations of the stereoscopic cinema. Van Nostrand Reinhold, New York, NY, USA. 91, 95

Pollock, B., Burton, M., Kelly, J., Gilbert, S., Winer, E., 2012. The right view from the wrong location: Depth perception in stereoscopic multi-user virtual environments. IEEE Transactions on Visualization and Computer Graphics 18 (4), 581–588.
http://dx.doi.org/10.1109/TVCG.2012.58 106

Schor, C. M., Kotulak, J. C., 1986. Dynamic interactions between accommodation and convergence are velocity sensitive. Vision Research 26 (6), 927–942.
http://dx.doi.org/10.1016/0042-6989(86)90151-3 100

Shibata, T., Kim, J., Hoffman, D. M., Banks, M. S., Jul. 2011. The zone of comfort: Predicting visual discomfort with stereo displays. Journal of Vision 11 (8).
http://dx.doi.org/10.1167/11.8.11 97, 101

Solimini, A. G., Feb. 2013. Are there side effects to watching 3D movies? A prospective crossover observational study on visually induced motion sickness. PLoS ONE 8 (2), e56160.
http://dx.doi.org/10.1371/journal.pone.0056160 96

Solimini, A. G., Mannocci, A., Thiene, D. D., Torre, G. L., Sep. 2012. A survey of visually induced symptoms and associated factors in spectators of three dimensional stereoscopic movies. BMC Public Health 12 (1), 779.
http://dx.doi.org/10.1186/1471-2458-12-779 96

Valois, K. K. D., Apr. 2000. Seeing. Academic Press. 98

Vetro, A., Wiegand, T., Sullivan, G., 2011. Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard. Proceedings of the IEEE 99 (4), 626–642.
http://dx.doi.org/10.1109/JPROC.2010.2098830 92

Vishwanath, D., Girshick, A. R., Banks, M. S., Oct. 2005. Why pictures look right when viewed from the wrong place. Nature Neuroscience 8 (10), 1401–1410.
http://dx.doi.org/10.1038/nn1553 105

Watt, S. J., Akeley, K., Ernst, M. O., Banks, M. S., Dec. 2005. Focus cues affect perceived depth. Journal of Vision 5 (10).
http://dx.doi.org/10.1167/5.10.7 101

Wilkins, A. J., Evans, B. J., Jul. 2010. Visual stress, its treatment with spectral filters, and its relationship to visually induced motion sickness. Applied Ergonomics 41 (4), 509–515.
http://dx.doi.org/10.1016/j.apergo.2009.01.011 96

Woods, A. J., 1993. Image distortions in stereoscopic video systems. In: Proc. of SPIE 1915. San Jose, CA, pp. 36–48.
http://dx.doi.org/10.1117/12.157041 102, 103, 105

Yamanoue, H., Nagayama, M., Bitou, M., Tanada, J., Motoki, T., Mituhashi, T., Hatori, M., 1998. Tolerance for geometrical distortions between L/R images in 3D-HDTV. Systems and Computers in Japan 29 (5), 37–48.
http://dx.doi.org/10.1002/(SICI)1520-684X(199805)29:5<37::AID-SCJ5>3.0.CO;2-O 102

Yano, S., Emoto, M., Mitsuhashi, T., Nov. 2004. Two factors in visual fatigue caused by stereoscopic HDTV images. Displays 25 (4), 141–150.
http://dx.doi.org/10.1016/j.displa.2004.09.002 95, 99

This page intentionally left blank.

Chapter

5

Combining 3D sound with stereoscopic-3D images on a single platform

Highlights

✓ We focus on the combination of 3D sound and s-3D images.

✓ The new challenges related to the combination of 3D sound with s-3D images are listed.

✓ The SMART-I^2, the 3D audiovisual platform used in this work, is described.

Contents

This chapter focuses on the combination of 3D sound and stereoscopic 3D (s-3D) images. Section 5.1 lists the new challenges related to the combination of 3D sound with s-3D images. Section 5.2 describes the SMART-I^2, the 3D audiovisual platform that is used in this work.

5.1 New challenges of stereoscopic-3D cinema

This section addresses different questions that arise when one considers combining a spatially accurate, or “3D” soundtrack with an s-3D video. First, we discuss the potential mismatch between (1) the space where the sound engineer positions sound sources (the soundscape), and (2) the space where the stereographer positions visual objects (the landscape). Then, we explain why the combination of accurate sound positioning and s-3D images can lead to errors, or *incongruence* between the sound and the image for multiple spectators.

5.1.1 Sound volumes versus image volumes

A first observation of today’s average cinema layout is that the volumes in which visual objects and auditory objects appear do not perfectly overlap [Mendiburu, 2009]. This is illustrated in Figure 5.1.

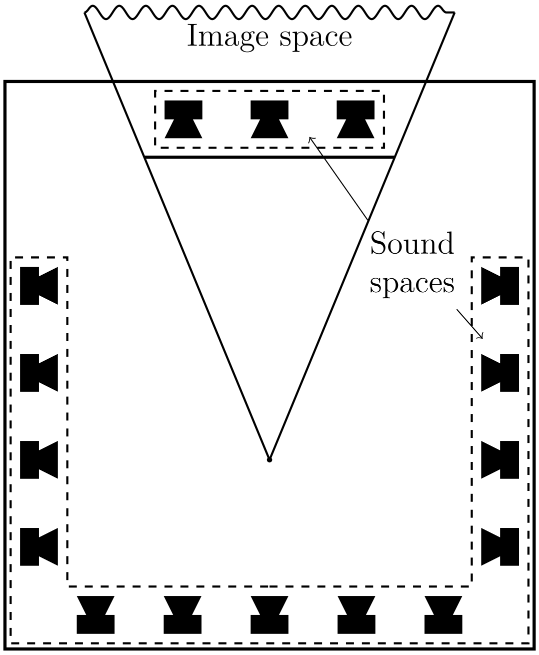


Figure 5.1: The sound and image spaces. After Mendiburu [2009].

Sound sources, on the one hand, are typically produced by three loudspeakers behind the screen, acting as a three-channel stereo source, and by a set of three linear arrays located on the sides and in the back, acting as a surround source to give the audience a sense of immersion.

Visual objects, on the other hand, can only be placed by the stereographer in our field of view. It is roughly a truncated cone with the apex at the viewer that extends towards infinity behind the screen. If needed, the volume necessary for sound reproduction can be restricted to the zone where the stereoscopy is comfortable, i.e. Percival's zone of comfort (Section 4.2.2).

Arguably, with images appearing from the front, and sound potentially appearing from all around the audience, today's layout suits the natural perception well. Regarding our goal in this thesis, however, it is currently not possible to match all visual positions with a corresponding sound source, particularly along the image depth axis. Only the sources at the screen plane can be accurately reproduced in stereo.

Ideally, one would want to render spatially precise sound sources in a volume at least corresponding to the audience's field of view. The question of whether a theater sound system should be able to render spatially precise sound at the sides and behind the audience needs to be addressed by the people involved in the process of 3D moviemaking.

5.1.2 Off-axis visual localization

Each image in an s-3D pair is physically located on the screen and all the viewers look at the same image pair. The above-mentioned truncated cone thus follows each spectator and, when the visual perception of two spectators seated at different locations in the room are compared, one concludes that the objects of the scene are not rendered at the same physical location in the room.

In Figure 5.2, two viewers A and B look at the same stereoscopic image of the sketch of a loudspeaker. Here, we use the geometrical model introduced in Section 4.1.4 (which will be detailed in Section 6.3). This visual object appears in front of the screen for both viewers, but their visual cue, i.e. the locations of their visual perception, are located at two different positions. Note that, for clarity, only one line segment of each visual cue is shown in the figure.

In combination with 3D sound, the fact that s-3D images are not consistently perceived by all spectators in a movie theater can lead to an audiovisual error for spectators seated off-axis. In Figure 5.3, two spectators at S_1 and S_2 look at the same s-3D pair of images on a screen. We assume that the images contain one object and that the images are such that the spectator at S_1, the ideal viewpoint, perceives the object as being located at V_1 (behind the screen). For the spectator seated at S_2, the visual object appears at V_2, resulting in an angular error δ between the sound and the image if the sound is positioned at $A = V_1$, the location of the object seen from the ideal viewpoint.

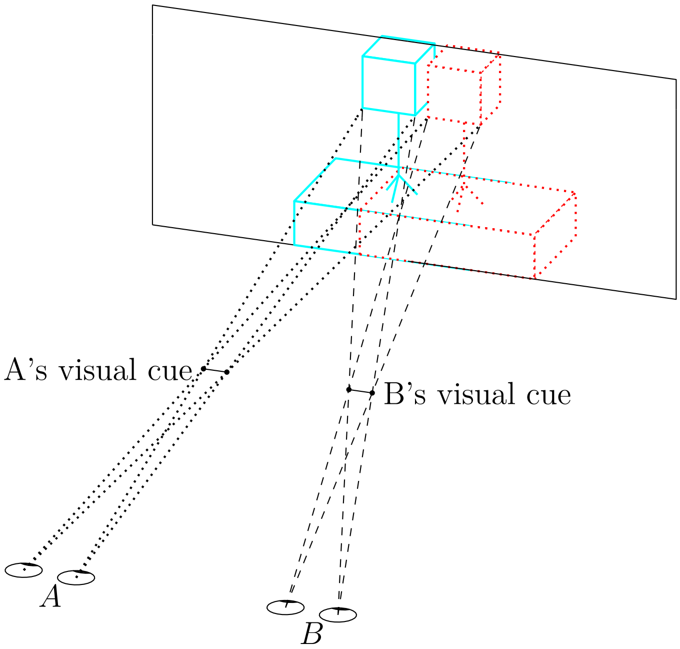


Figure 5.2: Visual localization for two spectators A and B seated at two different locations. For clarity, only one line segment of each visual cue is shown in the figure.

As will be seen in Section 6.3.4, the depth of the point-like image is proportional to the distance between the spectator and the screen. This is also shown in Figure 5.4. This means that seating closer or further from the screen respectively compresses or expands the image depth axis.

The seating location in the room is not the only factor that impacts the visual localization [Verduci, 2009]. The perceived egocentric distance from the viewer to the visual object is dependent on the viewer's *interocular distance*, i.e. the distance between his/her two eyes. The interocular distance mainly depends on ethnicity, gender, and age. Dodgson [2004] confirms this fact based the ANSUR dataset [Gordon et al., 1989]. In this particular dataset containing 3976 subjects, the rounded mean, median, and mode of the interocular distance are all 63 mm. This is slightly different from the 65 mm value which is often cited. 65 mm actually corresponds to the average interocular distance of American white males. A statistical difference (at the 99% confidence level) is found between gender and between certain racial groups. Almost all adults have an interocular distance in the range $45 - 80$ mm, and a large population is covered by the range $50 - 75$ mm. Finally, with an analysis of other datasets, Dodgson [2004] shows that the interocular distance increases from birth to late teens. A minimum value of 40 mm can be expected for children (down to five years old).

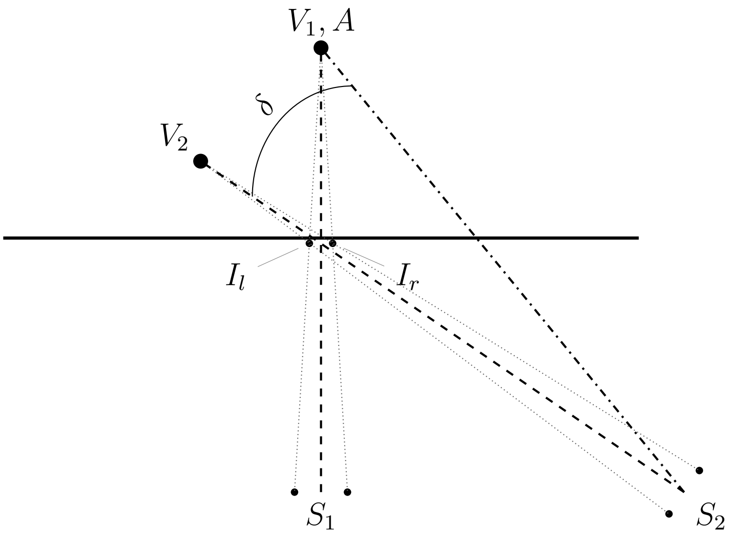


Figure 5.3: Illustration of the angular error between sound and image as a function of the seating position. The dotted lines are the light rays of the geometrical model. We assume that the images contain a single point-like object (I_l and I_r) and that the images are such that the spectator at S_1, the ideal viewpoint, perceives the object as being located at V_1 (behind the screen). As a result, the spectator at S_2 perceives the object as being located at V_2. For the spectator seated at S_2, the visual object appears at V_2, resulting in an angular error δ between the sound and the image when the sound is positioned at $A = V_1$, the location of the object seen from the ideal viewpoint.

For example, imagine a 6-year old child, a woman, and a man, each having the average interocular distance of their category (55 mm, 62 mm, and 65 mm, respectively.), sitting in the middle of the theater at 20 m from the screen, and looking at a bell that appears in front of the screen. If the man sees the bell at 6 m from him, then the woman sees it at 5.7 m (closer to her), and the child sees it at 5.1 m (closer to him).

5.2 The SMART-I^2

The SMART-I^2 system (Figure 5.5) is a high-quality 3D audiovisual interactive rendering system developed at the LIMSI-CNRS in collaboration with *sonic emotion*[*]. The 3D video and audio technologies are brought together using two Large Multi-Actuator Panels, or LaMAPs (2.6 m × 2 m), forming a "corner" that acts

[*] www.sonicemotion.com, last accessed 25/06/2013.

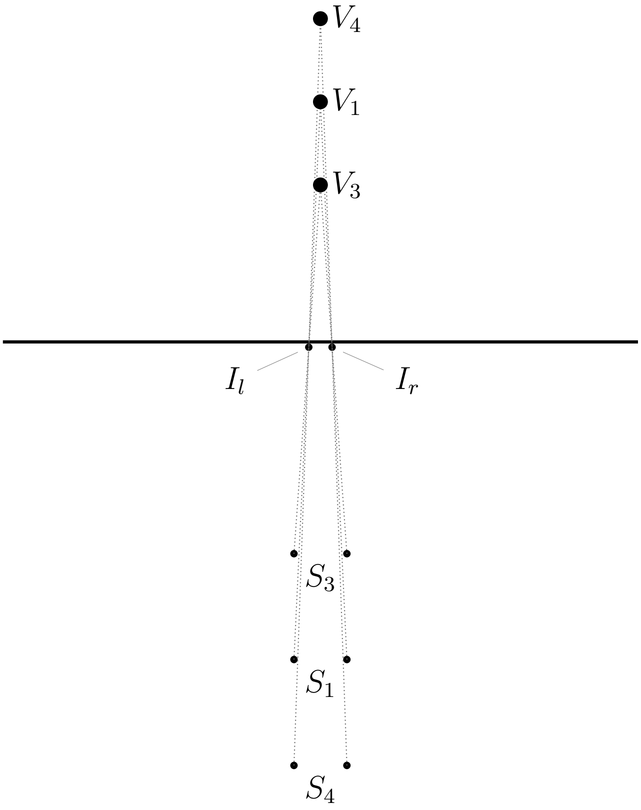


Figure 5.4: Illustration of the distance error between sound and image as a function of the seating position. The dotted lines are the light rays of the geometrical model. We assume that the images contain a single point-like object (I_l and I_r) and that the images are such that the spectator at S_1, the ideal viewpoint, perceives the object as being located at V_1 (behind the screen). As a result, the spectators at S_3 and S_4 perceive the object as being located at V_3 and V_4, respectively. Thus, seating closer or further from the screen respectively compresses or expands the image depth axis

both as projection screen, and as a 24-channel loudspeaker array. The s-3D video is presented to the user using dual-projection polarized stereoscopy. The actuators fixed at the back of each LaMAP allow for a WFS reproduction (Section 3.3.4) in a window corresponding to the s-3D video window* [Boone, 2004].

The 20 cm spacing between the actuators corresponds to an aliasing frequency of about 1.5 kHz, the upper frequency limit for a spatially correct wavefront synthesis, accounting for the size of the loudspeaker array, and the extension of the

*In the SMART-I² array, there is no loudspeaker in the traditional sense. Each actuator is fed with its own signal and excites the large panel. The whole SMART-I² configuration acts as a WFS loudspeaker array, provided there is no overlap of the regions excited by each actuator.

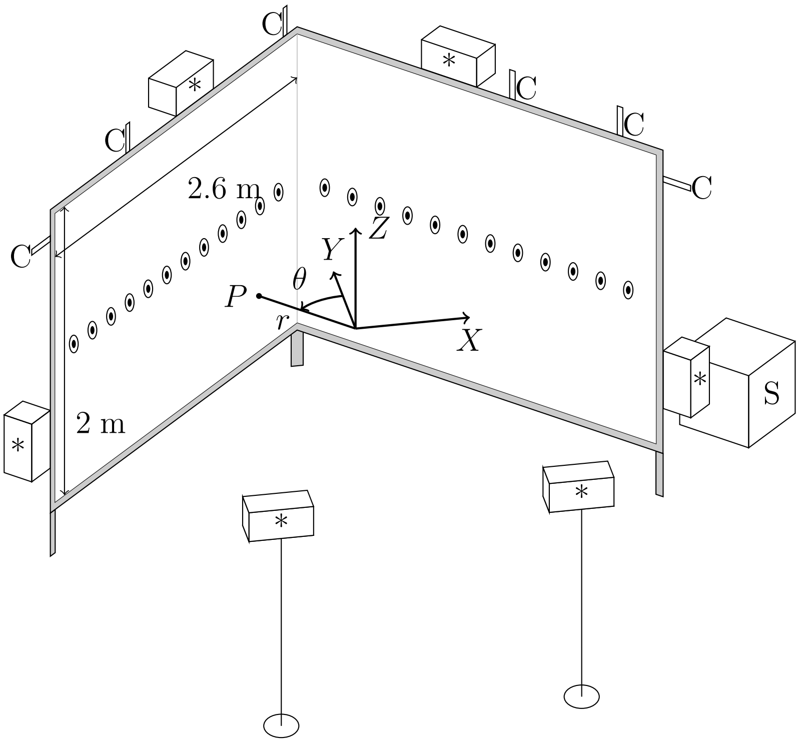


Figure 5.5: The SMART-I^2 and its coordinate axes. The origin of axes lies in the plane of the actuators. The Y axis points towards the corner, and the X axis points towards the right side of the right panel. The WFS actuators and the screens are co-located in depth. The actuators are invisible to the viewers. ⊙ – WFS actuators, **C** – IR cameras, **S** – subwoofer, $*$ – surround speakers.

listening area [Corteel, 2006]. The implementation of WFS used here is restricted to the synthesis of sound sources located in the horizontal plane [Corteel et al., 2012]. The azimuth and distance localization accuracies of sound events in the SMART-I^2 were previously verified by perceptual experiments and are globally consistent with corresponding real life localization accuracies. Rébillat et al. [2008] evaluated the azimuth localization accuracy of the WFS system in the SMART-I^2. They presented 17 virtual loudspeakers on a horizontal arc at 4 m from the listener. The loudspeakers were separated by 3°. Participants had to determine the origin of a 150 ms white noise burst. The median angular error was always less than 3°, and the variability, measured by the half inter-quantile range, was between 3 and 4°. These results are in line with the literature. Verheijen [1998] performed a similar experiment with a WFS loudspeaker array, comparing the localization accuracy of virtual WFS sources and real sources. With a loudspeaker spacing of 22 cm, the mean RMS error was 3.2°, with a standard deviation of the error of 1.4°. This was only slightly higher than the results for real sources, which were 2.6° and 1°, respectively.

Distance perception in the SMART-I^2 was evaluated by Rébillat et al. [2012]. Participants estimated the distance to virtual sources in the auditory, visual, and auditory-visual modalities. Using two methods, visual target selection and blind-walking triangulation, results were in line with the literature on real auditory source distance perception [Zahorik et al., 2005] (see Section 2.1.3). The perceived distance d_p to the auditory targets was modeled by the curve $d_p = kd_s^a$ where d_s is the simulated distance, and k and a are parameters of the model. The median values of k and a were 1.72 ± 0.09 and 0.33 ± 0.03, respectively. This was in line with the results of a review of 84 studies on auditory distance perception, where the average values for k and a were 1.32 and 0.54, respectively [Zahorik et al., 2005].

The SMART-I^2 processing architecture [Rébillat et al., 2008] is distributed, separating video, audio processing, and final WFS rendering over different applications and machines (Figure 5.6). It is therefore necessary to use some communication protocol. The Open Sound Control (OSC) format [Wright et al., 2003] is used for the audio metadata information (see Appendix C). The machine feeding the video to the four 2D projectors is called `djobi`. On the audio machine, `djoba`, a **Max/MSP**[*] patch[†] gathers all the OSC information, and directs the audio rendering. The direct sound is sent to the WFS rendering engine, sonic emotion's `Wave1`, which controls the actuators on the LaMAPs, while a **Max/MSP** based spatial audio processor (the Spat∼) generates the reverberant portion, which is then fed to the surround loudspeakers (Figure 5.5).

The Spat∼, or *Spatialisateur*, is a real-time modular spatial sound processing software system, developed under the **Max/MSP** environment by IRCAM and *Espaces Nouveaux* since the early nineties.

> A particular aim of the Spatialisateur project is to provide direct and computationally efficient control, over perceptually relevant parameters describing the interaction of each sound source with the virtual space, irrespective of the chosen reproduction format over loudspeakers or headphones. (Jot [1999])

In the SMART-I^2 system, the Spat∼ is used to generate the reverberant field only, while the WFS renders the direct sound.

The SMART-I^2 is currently capable of rendering 16 audio streams, in addition to the Spat∼ room effect channels. The spatial position of these streams can be

[*] www.cycling74.com, last accessed 25/06/2013.

[†] A program written in **Max/MSP** is called a *patch*.

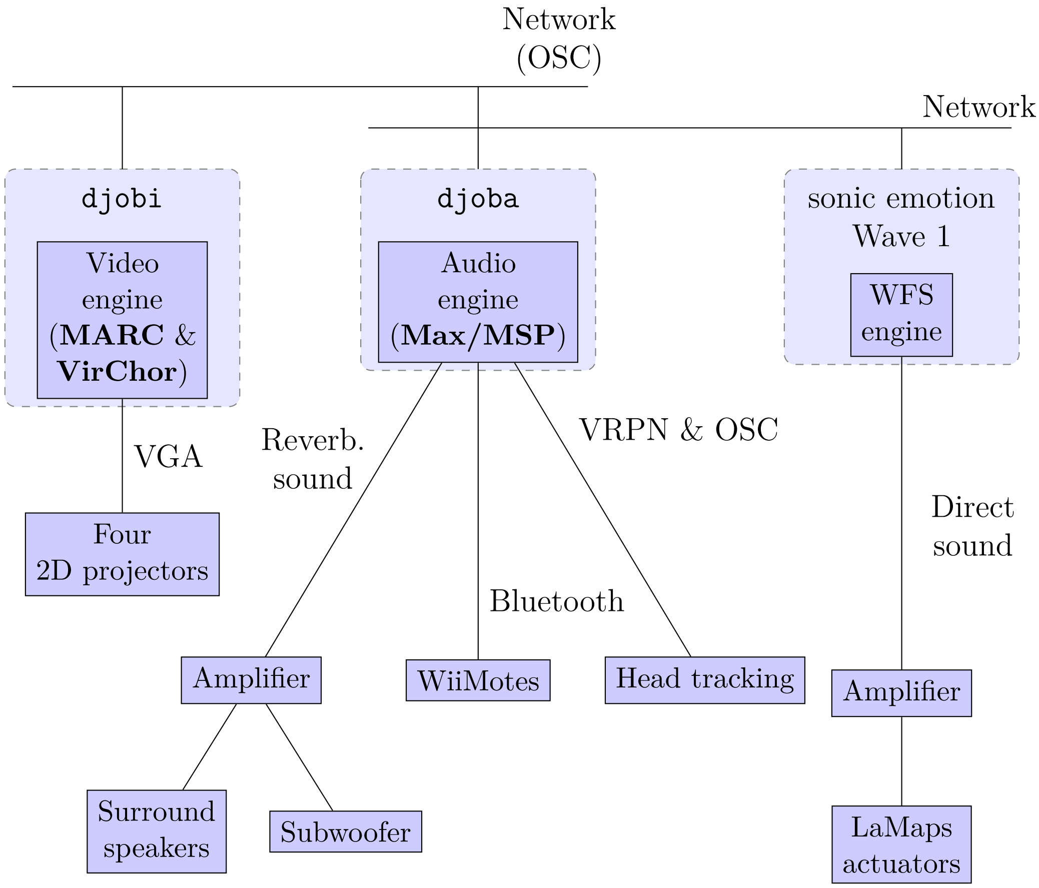


Figure 5.6: A block-diagram of the software and hardware previously available in the SMART-I^2. After Rébillat et al. [2008].

dynamically changed. In the studies of Chapters 6 and 7, the audio streams and spatial positions are controlled using a sequence table, which identifies the current audio files and their associated coordinates.

A detailed description of the SMART-I^2 and its associated modules can be found in [Rébillat et al., 2008]. The SMART-I^2 is equipped with infrared cameras to allow for the motion tracking of multiple users. This feature, however, is not used in this work.

Bibliography

Boone, M. M., 2004. Multi-Actuator Panels (MAPs) as loudspeaker arrays for Wave Field Synthesis. J. Audio Eng. Soc. 52 (7/8), 712–723.
http://www.aes.org/e-lib/browse.cfm?elib=13014 116

Corteel, É., Sep. 2006. On the use of irregularly spaced loudspeaker arrays for

Wave Field Synthesis, potential impact on spatial aliasing frequency. In: Proc. 9th Int. Conf. on Digital Audio Effects (DAFx'06). Montréal, Canada. 117

Corteel, É., Rohr, L., Falourd, X., NGuyen, K.-V., Lissek, H., Apr. 2012. Practical 3-dimensional sound reproduction using Wave Field Synthesis, theory and perceptual validation. In: Proceedings of the 11th French Congress of Acoustics and 2012 Annual IOA Meeting. Nantes, France, pp. 895–900. 117

Dodgson, N. A., May 2004. Variation and extrema of human interpupillary distance. In: Woods, A. J., Merritt, J. O., Benton, S. A., Bolas, M. T. (Eds.), Proc. SPIE 5291. San Jose, CA, pp. 36–46.
http://dx.doi.org/10.1117/12.529999 114

Gordon, C. C., Churchill, T., Clauser, C. E., Bradtmiller, B., McConville, J. T., Tebbetts, I., Walker, R. A., Sep. 1989. 1988 anthropometric survey of U.S. Army personnel: Methods and summary statistics. Final NATICK/TR-89/044, U.S. Army Natick RD&E Center, Natick, MA.
http://nsrdec.natick.army.mil/ANSURII/index.htm 114

Jot, J.-M., 1999. Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces. Multimedia Systems 7 (1), 55–69.
http://portal.acm.org/citation.cfm?id=297260 118

Mendiburu, B., Apr. 2009. 3D Movie Making: Stereoscopic Digital Cinema from Script to Screen. Focal Press. 112

Rébillat, M., Boutillon, X., Corteel, É., Katz, B. F. G., Oct. 2012. Audio, visual, and audio-visual egocentric distance perception by moving subjects in virtual environments. ACM Trans. Appl. Percept. 9 (4), 19:1–19:17.
http://dx.doi.org/10.1145/2355598.2355602 118

Rébillat, M., Corteel, É., Katz, B. F. G., Oct. 2008. SMART-I^2: Spatial Multi-User Audio-Visual Real Time Interactive Interface. In: Audio Eng. Soc. Conv. 125.
http://www.aes.org/e-lib/browse.cfm?elib=14760 117, 118, 119

Verduci, L., Dec. 2009. 3D audio and 3D images for 3D movie theater. In: 3D Stereo MEDIA conference. Liège, Belgium. 114

Verheijen, E. N. G., 1998. Sound reproduction by Wave Field Synthesis. Ph.D. thesis, Delft University of Technology. 117

Wright, M., Freed, A., Momeni, A., 2003. OpenSound Control: State of the art 2003. In: Proc. 2003 Conf. on new interfaces for musical expression (NIME-03). Montreal, Canada, pp. 153–159. 118

Zahorik, P., Brungart, D. S., Bronkhorst, A. W., May 2005. Auditory distance perception in humans: A summary of past and present research. Acta Acust. united with Acust. 91, 409–420(12). 118

This page intentionally left blank.

Part II

Contributions of the thesis

Chapter 6

The making of a 3D audiovisual content

Highlights

✓ A general mathematical model is given which transforms the 3D coordinates of a visual object captured by an s-3D camera into the 3D coordinates where the object appears when the movie is played.

✓ For an existing s-3D animation movie, an audio track is created that precisely matches the spatial positions of the visual objects.

✓ The creation of the audio track is described from the early editing stage, through the scripting process, to the sending of data to the rendering system.

Contents

6.1 Introduction

Although many movies are now produced in stereoscopic 3D (s-3D), the sound in these movies is still most often mixed in 5.1 surround. This format has rather imprecise localization capabilities. Hence, the sound mix does not provide the spectator with a 3D sound scene spatially consistent with the visual content of the movie.

We investigate in this chapter the possibility of adding an audio track that precisely matches the spatial positions of the visual objects, in both angular position and distance in space. Achieving coherence between the positions of visual and audio objects is our key goal. Note that the object need not be present in the visual field. For example, one could hear a character’s steps well after the character has left the screen. Still, the sound would originate from a likely position, say to the left or right of the screen.

The current study, published in [Évrard et al., 2011][*], is carried out on the SMART-I^2 (Section 5.2). Spatial audio here is based on Wave Field Synthesis

[*] Our contribution in this study was the selection of the movie (Section 6.2), the development of the mathematical model in Section 6.3, the software and hardware architecture (Section 6.4), and the scripting work mentioned in Section 6.5.3.

(WFS). Despite the fact that the current implementation of WFS in this setup is not a full 3D audio system, as sound sources are limited to the horizontal plane, the results are produced in a general manner that allows for expansion to a full 3D audio system, or rendering on other high-definition systems.

This study is also the first step towards an experiment using 3D audiovisual content. This experiment, which will be described in Chapter 7, will examine, in the cinema context, the perceptual differences between a traditional audio rendering, based on stereophonic principles, and a highly precise audio rendering. In addition to the description of the content creation, we give in this chapter a mathematical model which transforms the 3D coordinates of a visual object captured by an s-3D camera into the 3D coordinates where the object is perceived to appear in the physical world. This model is generally applicable to any s-3D content, and we will use it in all the subsequent chapters.

In Section 6.2, we describe the reasoning which led us to choose one particular animation movie, “Elephants Dream”. In Section 6.3, a complete geometrical model is given which allows one to express the coordinates of the objects in the movie in the SMART-I^2 coordinate axes. In Section 6.4, we describe the customization applied to the SMART-I^2 to produce the 3D audiovisual content. The creation of the new soundtrack is described in Section 6.5. Finally, we review in Section 6.6 some of the problems encountered in this work and discuss the chosen solutions.

6.2 Selecting a movie

We decided to use an animation s-3D movie to carry out this study, rather than a real-image s-3D movie. The reason was that the use of an animation movie, and, more specifically, the software that created it, allows one to automatically obtain the exact spatial information of all objects present in the scenes from the source files.

The film selected for this project is “Elephants Dream”[*], an open movie, made entirely with **Blender**, a free open source 3D content creation suite[†]. All production files necessary to render the movie video are freely available on the Internet, under a Creative Commons license. To be precise, the s-3D version has not been

[*] www.elephantsdream.org, last accessed 25/06/2013.

[†] www.blender.org, last accessed 25/06/2013. The version used was **Blender** 2.41, and the comments we make on the features of **Blender** are related to this version. **Blender** has tremendously evolved since, but the movie sources were not compatible with the newer versions.

published on the web, but Wolfgang Draxinger, the stereographer, kindly sent its sources to us.

Olaiz et al. [2009] had previously realized a similar experiment with the open source game “Yo Frankie!”, also developed with **Blender**. Spatially coherent sound was added to the game. Different rendering systems were tested, namely binaural, VBAP, and first-order Ambisonics. We were able to reuse the export scripts developed in **Python** for this previous study, but we had to make slight modifications to take into account the new variables of s-3D.

The audio track of the movie is only available in stereo or 5.1 mixes. As such, it was necessary to recreate a new audio track that was chosen to be as similar as possible to, and inspired by, the original track (Figure 6.1). For this pilot study, only the first three scenes of the movie were generated (from t = 00 min 00 s to t = 02 min 30 s).

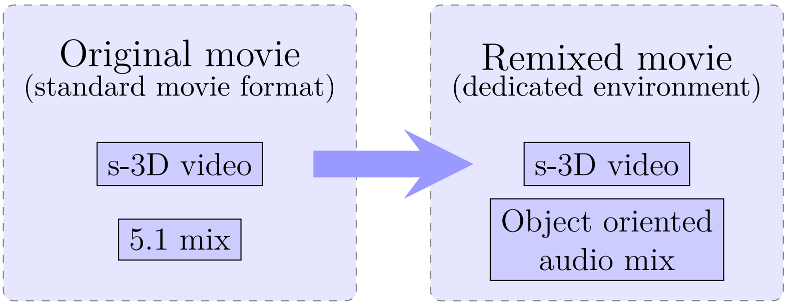


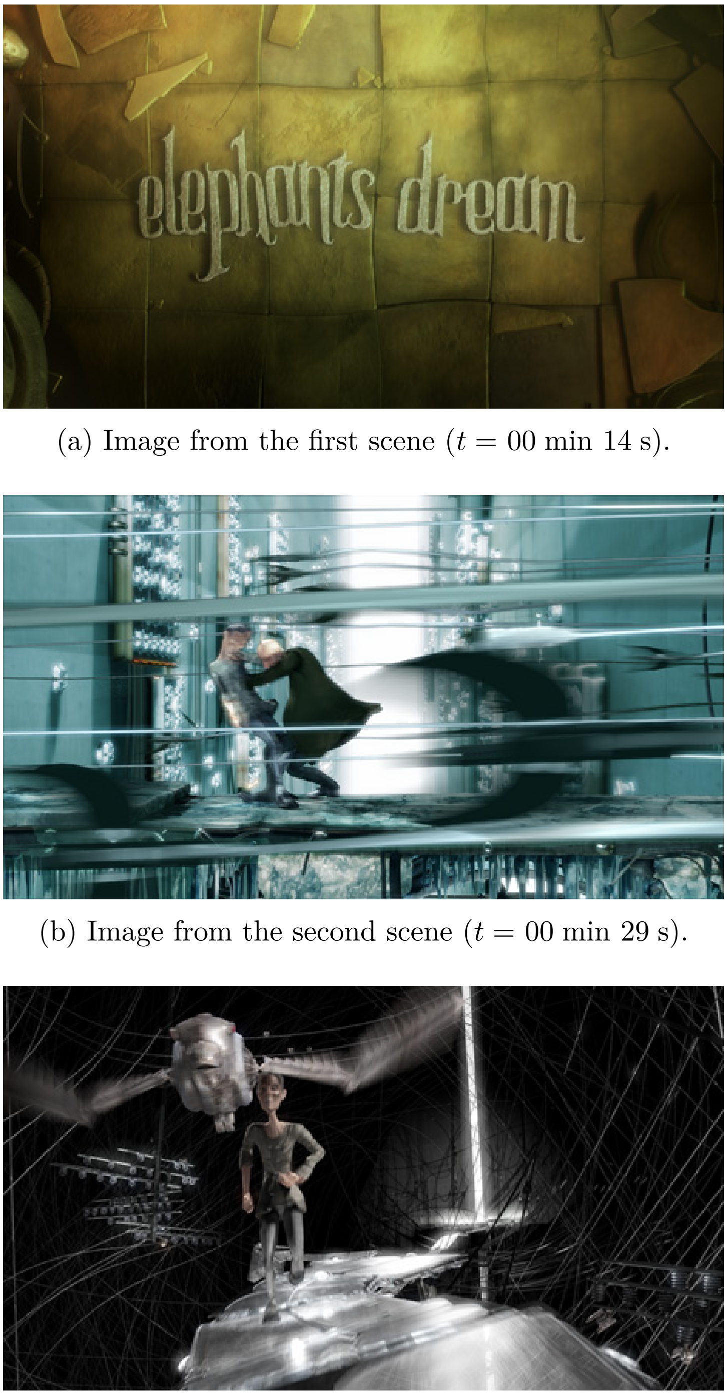
Figure 6.1: Summary of the work performed in this study.

The first scene (t = 00 min 00 s) starts with the opening credits (Figure 6.2(a)), where the camera travels upward until it reaches the first character’s reflection in water. In the second scene (t = 00 min 27 s), the two characters are attacked by flying cables, and there is a dialog (Figure 6.2(b)). The third scene (t = 01 min 10 s) consists of the two characters running through a large room, being chased by mechanical birds (Figure 6.2(c)).

In the next section, we study how to obtain the coordinates of the sound sources in the SMART-I^2 coordinate axes from the coordinates of the visual objects in the movie source files.

6.3 Precise sound positioning in the SMART-I^2

A major issue in the study was to achieve spatially coherent rendering of visual and audio objects. In common practice, the coordinate system of the graphics scene



(a) Image from the first scene (t = 00 min 14 s).

(b) Image from the second scene (t = 00 min 29 s).

(c) Image from the third scene (t = 01 min 55 s).

Figure 6.2: Three images extracted from the original “Elephants Dream” video sequence.

modeler, and that of the spectator's physical world are not the same. Changes in s-3D camera configurations produce changes in the perceived positions of visual objects relative to the spectator, which are determined relative to the plane of the projection screen (Section 4.1.4). In addition, different audio rendering methods have different coordinate system references, being either *egocentric*, for systems such as binaural, or *allocentric*, for systems such as WFS.

Sound in the SMART-I^2 is spatially positioned with respect to a coordinate basis attached to the SMART-I^2 itself (see Figure 5.5). A sound object has its coordinates expressed in meters.

To achieve spatial coherence between visual and auditory objects, it is necessary to determine where the visual objects appear in the physical space, in order to match the corresponding sound coordinates. We will see in this section that this is conveniently written as frame transformations of points expressed in homogeneous coordinates. The reader unfamiliar with these concepts will find the complete mathematical background relevant to this section in Appendix D. Woods [1993] has previously introduced this model. Here, we stress out its true geometrical origin (frame transformations) and we express it as a matrix product for efficient computation.

The coordinates of a perceived visual object are determined by reading the object's coordinates in **Blender** at each frame, and transforming them successively to the following reference frames:

Blender space: the common basis for all **Blender** objects, including the cameras.

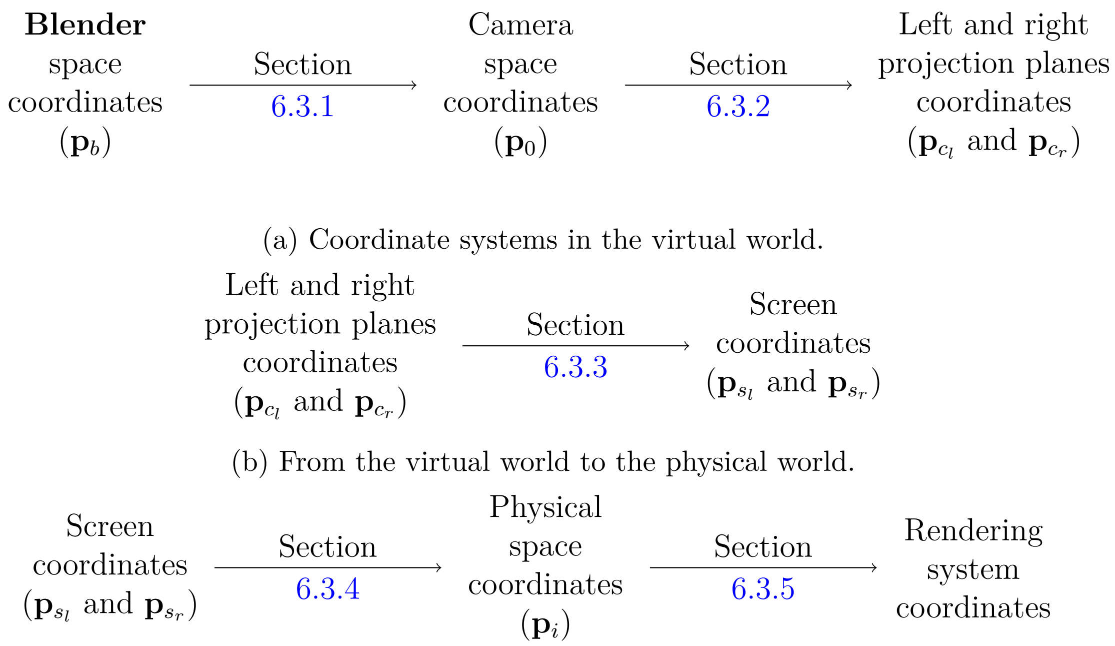
Camera space: the body-fixed camera basis, whose origin is at the center of the s-3D rig.

Projection planes: each virtual 2D camera in the s-3D rig has its own projection plane; the equivalent of the two camera sensors in a real s-3D rig.

Display system: the basis whose origin is located at the midpoint between the spectator's eyes.

In the next and final step, the display system coordinates are adapted to the audio reproduction system. The sequence of transformations is further illustrated in Figure 6.3.

In all these bases, the X-axis points towards the right of the image, and the Y-axis points towards the top of the image. The Z-axis is chosen to form a right-handed basis, and thus points from the screen towards the spectator.


Blender space coordinates ($\mathbf{p}_b$)
Section 6.3.1
Camera space coordinates ($\mathbf{p}_0$)
Section 6.3.2
Left and right projection planes coordinates ($\mathbf{p}_{c_l}$ and $\mathbf{p}_{c_r}$)
Left and right projection planes coordinates ($\mathbf{p}_{c_l}$ and $\mathbf{p}_{c_r}$)
Section 6.3.3
Screen coordinates ($\mathbf{p}_{s_l}$ and $\mathbf{p}_{s_r}$)
Screen coordinates ($\mathbf{p}_{s_l}$ and $\mathbf{p}_{s_r}$)
Section 6.3.4
Physical space coordinates ($\mathbf{p}_i$)
Section 6.3.5
Rendering system coordinates

(a) Coordinate systems in the virtual world.

(b) From the virtual world to the physical world.

(c) Coordinate systems in the physical world.

Figure 6.3: From **Blender** space to the rendering system coordinates.

The following sections describe these transformations in detail.

6.3.1 From Blender space coordinates to camera space coordinates

We express the coordinates of each object in the scene with respect to the basis attached to the active virtual camera position. This basis is the analog of the cyclopean eye (Section 2.2.1). Its origin is located at the midpoint between the centers of the two lenses.

The relation between any point $\mathbf{p}_b = \left[X_b, Y_b, Z_b, 1\right]^T$ in the **Blender** space and its equivalent $\mathbf{p}_0$ in the camera space is written as

$$\begin{bmatrix}X_0\\Y_0\\Z_0\\1\end{bmatrix} = \left[\begin{array}{c|c}\mathbf{R}_{XYZ}(\phi,\theta,\psi) & \mathbf{0}\\ \hline \mathbf{0}^T & 1\end{array}\right]\left[\begin{array}{c|c}\mathbf{I} & -\mathbf{c}_b\\ \hline \mathbf{0}^T & 1\end{array}\right]\begin{bmatrix}X_b\\Y_b\\Z_b\\1\end{bmatrix}, \tag{6.1}$$

where $\mathbf{c}_b$ is the position of the camera in the **Blender** space, and the angles $\left[\phi, \theta, \psi\right]$ define the attitude of the camera (Section D.2).

6.3.2 From camera space coordinates (3D) to left and right projection planes coordinates (2D).

The transformation from the camera space to the left and right projection planes is composed of two transformations. The first one is a change of basis from the basis attached to the active virtual camera object, to the basis linked to the center of each of the two sensors constituting the s-3D rig. The second transformation is the projection of the 3D coordinates of the scene objects onto the camera projection plane.

First, we derive a geometrical model for the s-3D rig; then, we compute the different transformation matrices.

Woods' geometrical model for the s-3D camera Woods [1993] developed a geometrical model for the s-3D camera. It is valid for both a toed-in camera configuration ($h = 0$) and a parallel camera configuration ($\beta = 0$) (Section 4.2.4).

According to this model, an s-3D camera system (Figure 6.4) is geometrically defined by

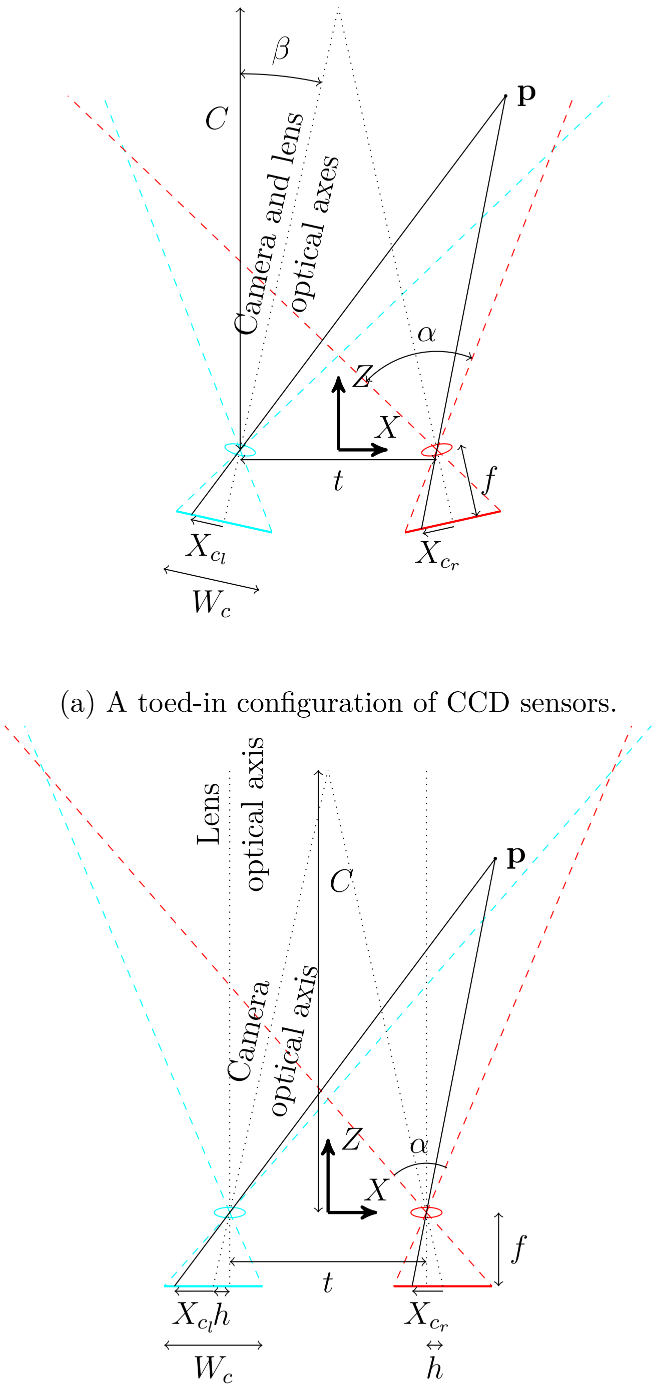
- the distance t between the cameras,
- the convergence distance C, which is the distance from the midpoint between the centers of two lenses to the point where the optical axes of the cameras intersect,
- the field of view α of the cameras, which is in turn determined by the sensor width W_c and the lens focal length f.

The convergence distance C is linked to the toed-in and parallel camera parameters β and h, respectively, through

$$C = \frac{t}{2\tan\left(\beta + \arctan\left(\frac{h}{f}\right)\right)}. \tag{6.2}$$

From camera space to each lens space These changes of basis consist in a translation from the center of the s-3D rig to each lens located at $\left[\pm\frac{t}{2}, 0, 0\right]$ and, if the camera is toed-in, a rotation by an angle $\pm\beta$ about the Y-axis. The + sign corresponds to the right lens, and the − sign to the left lens.

These changes of basis transform a 3D point $\mathbf{p}_0 = \left[X_0, Y_0, Z_0\right]^T$, expressed in the camera basis, into two 3D points $\mathbf{p}_l = \left[X_l, Y_l, Z_l\right]^T$ and $\mathbf{p}_r = \left[X_r, Y_r, Z_r\right]^T$,



(a) A toed-in configuration of CCD sensors.

(b) A parallel configuration of CCD sensors.

Figure 6.4: Camera parameters for a toed-in configuration and a parallel configuration. Note that the bases suggested in the picture do not form a right-handed coordinate system when the Y-axis points towards the reader, contrary to the bases we use in this work. After Woods [1993].

expressed in the bases of the left and right lenses respectively,

$$\begin{bmatrix} X_l \\ Y_l \\ Z_l \\ 1 \end{bmatrix} = \left[\begin{array}{c|c} \mathbf{R}_Y(-\beta) & \mathbf{0} \\ \hline \mathbf{0}^T & 1 \end{array}\right] \left[\begin{array}{c|c} & \frac{t}{2} \\ \mathbf{I} & 0 \\ & 0 \\ \hline \mathbf{0}^T & 1 \end{array}\right] \begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \\ 1 \end{bmatrix}, \tag{6.3}$$

and

$$\begin{bmatrix} X_r \\ Y_r \\ Z_r \\ 1 \end{bmatrix} = \left[\begin{array}{c|c} \mathbf{R}_Y(\beta) & \mathbf{0} \\ \hline \mathbf{0}^T & 1 \end{array}\right] \left[\begin{array}{c|c} & -\frac{t}{2} \\ \mathbf{I} & 0 \\ & 0 \\ \hline \mathbf{0}^T & 1 \end{array}\right] \begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \\ 1 \end{bmatrix}. \tag{6.4}$$

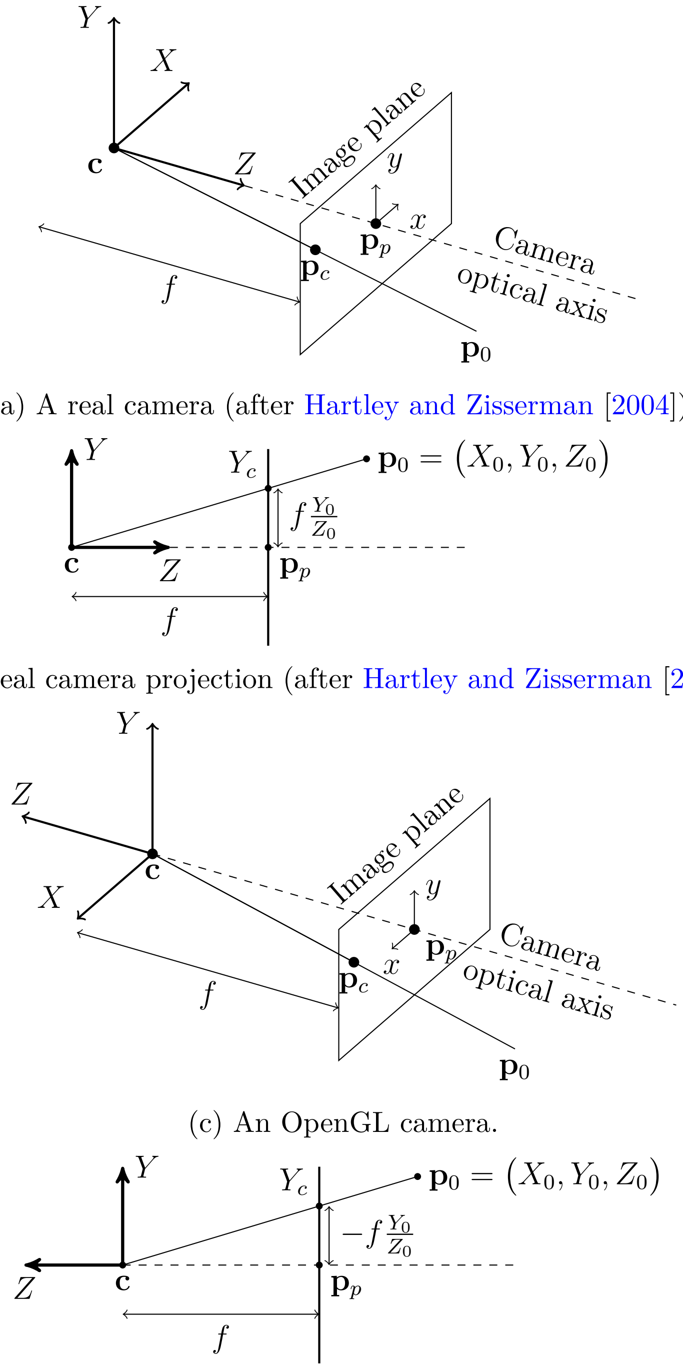
The Blender camera The virtual s-3D camera used to shoot "Elephants Dream" had a parallel configuration. The configuration is built-in to **Blender**, and does not appear before the rendering, meaning that the geometry of the virtual s-3D camera rig is not visible. Only the scene (cyclopean) camera is visible. The position of the scene camera is the center of the virtual s-3D rig. Only the convergence plane (the plane located at a distance C from the camera for the case of parallel configuration, or a curve in the case of toed-in configuration) is shown in **Blender**. The camera position and the parameters related to the s-3D configuration are varied throughout the movie. The parameters related to the s-3D configuration are also extracted as part of the coordinate transformation processing.

Woods's model relies on the traditional pinhole camera model (Figure 6.5(a)), which assumes a point lens. Mathematically, the pinhole camera model transforms a 3D point $\mathbf{p}$, expressed in the basis attached to the camera center ($\mathbf{c}$ in Figure 6.5), into the 2D point $\mathbf{p}_c = \left[X_c, Y_c\right]^T$ through

$$\begin{bmatrix} X_c \\ Y_c \end{bmatrix} \leftarrow \begin{bmatrix} X_c Z_c \\ Y_c Z_c \\ Z_c \end{bmatrix} = \begin{bmatrix} f & 0 & p_x & 0 \\ 0 & f & p_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{6.5}$$

where $\mathbf{p}_p = \left[p_x, p_y\right]^T$ is the principal point, which is the projection of the camera center on the image plane. This relation is valid when the Z-axis faces towards the point of interest $\mathbf{p}$ (Figure 6.5(b)).

However, the **Blender** camera is an OpenGL camera (Figure 6.5(c)), which means that the points of interest are in the $-Z$ direction. Therefore, the projection



(a) A real camera (after Hartley and Zisserman [2004]).

(b) Real camera projection (after Hartley and Zisserman [2004]).

(c) An OpenGL camera.

(d) Virtual camera projection.

Figure 6.5: The pinhole camera geometry.

matrix of the **Blender** camera is different from that in Eq. (6.5); it is given by

$$\begin{bmatrix} X_c Z_c \\ Y_c Z_c \\ Z_c \end{bmatrix} = \begin{bmatrix} -f & 0 & p_x & 0 \\ 0 & -f & p_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{6.6}$$

which defines $\mathbf{p}_{c_l}$ on the left projection plane ($\mathbf{p}_p = \left[-h, 0\right]^T$), and $\mathbf{p}_{c_r}$ on the right projection plane ($\mathbf{p}_p = \left[h, 0\right]^T$) from the points $\mathbf{p}_l$ and $\mathbf{p}_r$, respectively (Figure 6.5(d)). The overall equation for the left camera, valid for both toed-in and parallel camera configurations, summarizing the transformations of this section (Equations (6.3) to (6.6)), is:

$$\begin{bmatrix} X_{c_l} Z_{c_l} \\ Y_{c_l} Z_{c_l} \\ Z_{c_l} \end{bmatrix} = \begin{bmatrix} -f & 0 & -h & 0 \\ 0 & -f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \cos\beta & 0 & \sin\beta & \frac{t}{2}\cos\beta \\ 0 & 1 & 0 & 0 \\ -\sin\beta & 0 & \cos\beta & -\frac{t}{2}\sin\beta \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \\ 1 \end{bmatrix}. \tag{6.7}$$

Similarly, one has, for the right camera,

$$\begin{bmatrix} X_{c_r} Z_{c_r} \\ Y_{c_r} Z_{c_r} \\ Z_{c_r} \end{bmatrix} = \begin{bmatrix} -f & 0 & h & 0 \\ 0 & -f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \cos\beta & 0 & -\sin\beta & -\frac{t}{2}\cos\beta \\ 0 & 1 & 0 & 0 \\ \sin\beta & 0 & \cos\beta & -\frac{t}{2}\sin\beta \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \\ 1 \end{bmatrix}. \tag{6.8}$$

6.3.3 From left and right projection planes coordinates to screen coordinates

This step is simply a linear scaling of the coordinates by the frame magnification M, which is the ratio of the projection screen width W_s to the sensor width W_c. In **Blender** 2.41, there is no sensor *per se*, but the projection plane has always a width of 32 "Blender units", so that

$$M = \frac{W_s}{32}. \tag{6.9}$$

We then have

$$\begin{bmatrix} X_{s_l} \\ Y_{s_l} \end{bmatrix} = M \begin{bmatrix} X_{c_l} \\ Y_{c_l} \end{bmatrix} \text{ and } \begin{bmatrix} X_{s_r} \\ Y_{s_r} \end{bmatrix} = M \begin{bmatrix} X_{c_r} \\ Y_{c_r} \end{bmatrix}, \tag{6.10}$$

for the left image and the right image, respectively.

6.3.4 From screen coordinates to physical space coordinates

This steps corresponds to the fusion of the s-3D stimulus. The display system (Figure 6.6) is geometrically defined by the viewing distance V, i.e. the distance between the spectator and the screen, the width of the screen W_s, and the interocular distance e (typically 65 mm).

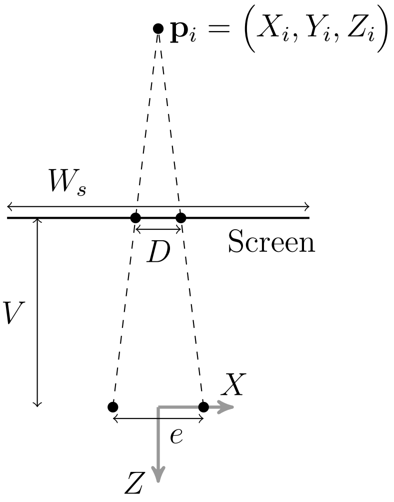


Figure 6.6: Viewing parameters of the SMART-I^2 display system. The spectator's eyes are symbolized by two dots, aligned along the X-axis, and separated by a distance e. The disparity D between the left and right images is defined as $D = X_{s_r} - X_{s_l}$. The fused stereoscopic image of the left and right images is $\mathbf{p}_i$.

Following the reasoning of Section 4.1.4, we construct two rays from each eye to the corresponding points on the screen $\mathbf{p}_{s_l}$ and $\mathbf{p}_{s_r}$. These two rays intersect at $\mathbf{p}_i$, the fused stereoscopic image of the two points.

We can write the equations of the two lines corresponding to the light rays. A point $\mathbf{p}$ from the line passing through the left eye and a point $\mathbf{p}_{s_l}$ in the left image is expressed as $\mathbf{e}_l\mathbf{p} = \lambda_l \mathbf{e}_l\mathbf{p}_{s_l}$, $\lambda_l \in \mathbb{R}$ (Section D.1). Similarly, a point $\mathbf{p}$ from the line passing through the right eye and a point $\mathbf{p}_{s_r}$ in the right image is expressed as $\mathbf{e}_r\mathbf{p} = \lambda_r \mathbf{e}_r\mathbf{p}_{s_r}$, $\lambda_r \in \mathbb{R}$. When the spectator's eyes are at the position depicted in Figure 6.6, that is, when

$$\mathbf{e}_l = \begin{bmatrix} -\frac{e}{2} \\ 0 \\ 0 \end{bmatrix} \text{ and } \mathbf{e}_r = \begin{bmatrix} \frac{e}{2} \\ 0 \\ 0 \end{bmatrix}, \tag{6.11}$$

the equations defining $\mathbf{p}_i$ are written as

$$\begin{cases} X_i = -\dfrac{e}{2} + \lambda_l(X_{s_l} + \dfrac{e}{2}) \\ Y_i = \lambda_l Y_{s_l} \\ Z_i = -\lambda_l V \end{cases} \tag{6.12}$$

and

$$\begin{cases} X_i = \dfrac{e}{2} + \lambda_r(X_{s_r} - \dfrac{e}{2}) \\ Y_i = \lambda_r Y_{s_r} \\ Z_i = -\lambda_r V \end{cases} \tag{6.13}$$

Solving these equations, we have

$$\begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix} = \begin{bmatrix} \frac{e}{e-D}\frac{X_{s_r}+X_{s_l}}{2} \\ \frac{e}{e-D}\frac{Y_{s_r}+Y_{s_l}}{2} \\ -\frac{e}{e-D}V \end{bmatrix} \tag{6.14}$$

where D is the disparity between the left and right images, defined as

$$D = X_{s_r} - X_{s_l}. \tag{6.15}$$

Since $\lambda_l = \lambda_r$ follows from Equations (6.12) and (6.13), there are in fact two different values of Y_i. The chosen value for Y_i is the average of those two values [Woods, 1993]. The difference between the two screen vertical coordinates is a measure of the easiness to fuse the stereoscopic image. Since the **Blender** s-3D camera is virtual, however, the sensor is perfect and there is no vertical disparity.

The equations given here are valid when the line joining the spectator's eyes is parallel to the screen. The Equations (6.11) can easily be generalized to the case of arbitrary rotations of the spectator's head. In this work, we assume that the spectator is looking towards the s-3D stimulus on the screen.

6.3.5 From physical space coordinates to rendering system coordinates

The final transformation step depends on the reproduction system.

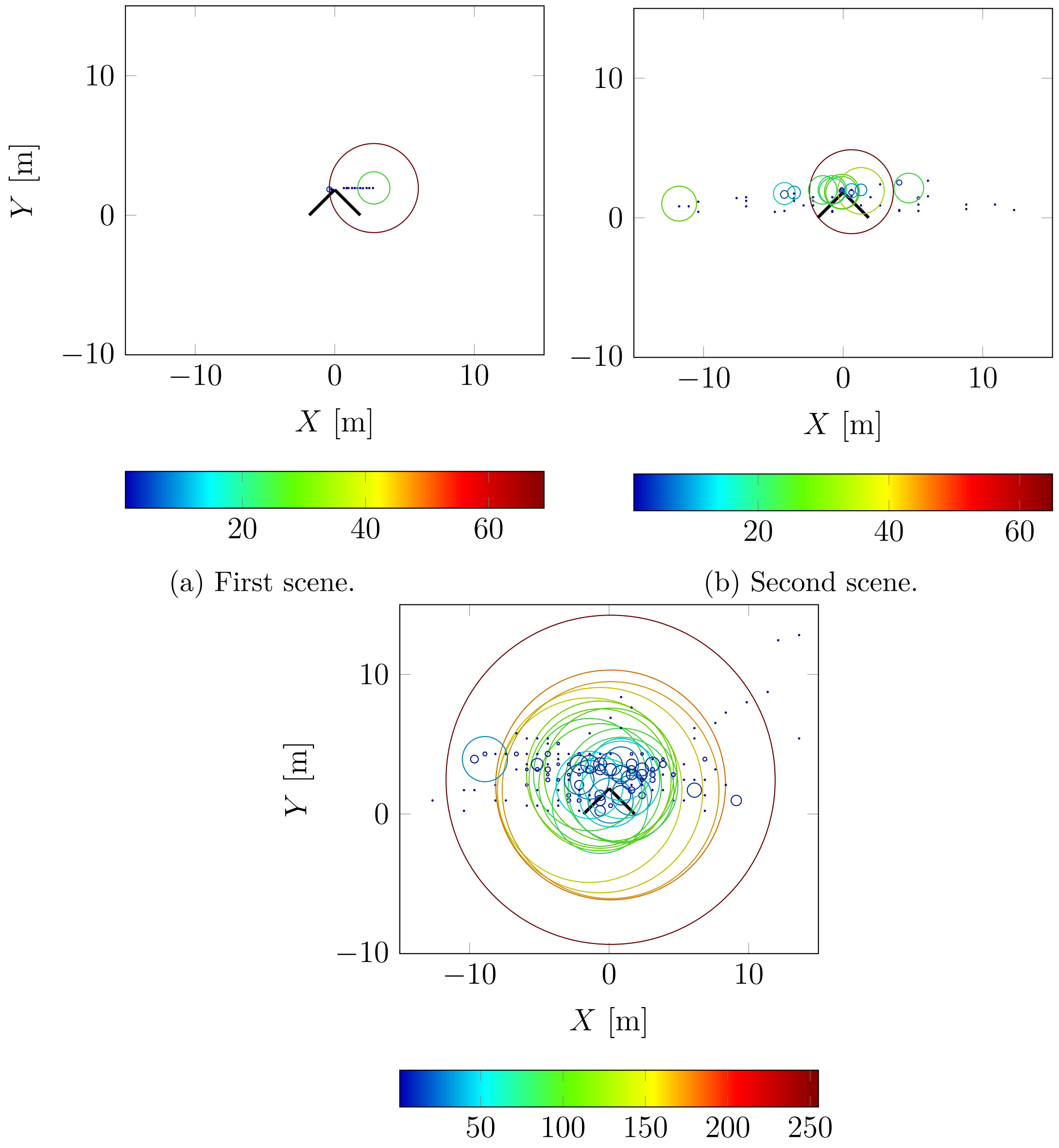
General 3D reproduction When the purpose of the transformation is the reproduction on a general 3D audio engine using head-centered spherical coordinates (Figure 2.2), one must simply permute the coordinates obtained in Section 6.3.4 to express the coordinates in a new Cartesian basis $(X_e, Y_e, Z_e) = (-Z, -X, Y)$. Then, the Cartesian coordinates can simply be transformed into spherical coordinates, to be used for example in a binaural engine.

Reproduction on the SMART-I^2 For the reproduction system used in this work, i.e. the SMART-I^2, the transformation consists in rotating the previously obtained axes by 180° about the X-axis, so that the new Z-axis faces towards the corner of the SMART-I^2 screens. The new position of the Z-axis defines the Y-axis of the SMART-I^2 (Figure 5.5).

Then, we transform these (X, Y) coordinates into the usual polar coordinates (r, θ). Finally, by subtracting 90° from the value of θ in the usual polar coordinates, we obtain the θ value used in the SMART-I^2 (Figure 5.5). The elevation information is discarded.

6.3.6 Coordinates transformation and the movie "Elephants Dream"

The proposed model is now used to obtain information about the positions of the visual sources present in the movie scenes. Our interest is in characterizing the positions of those visual sources which we might pair with a sound file. The density plots (also called bubbleplots) of these visual sources were calculated for the three scenes (Figure 6.7), indicating the positions where these sources are present. It can be observed that, for all three scenes, most of the sources were frontal and centered, located just behind the two screens (the panels of the SMART-I^2). The second scene exhibits many lateral sources. In general, few sources are found in front of the screen, with only the third scene exploiting depth variations. The paths of the cables in the second scene and the birds in the third scene are the farthest positioned sources.



(a) First scene.

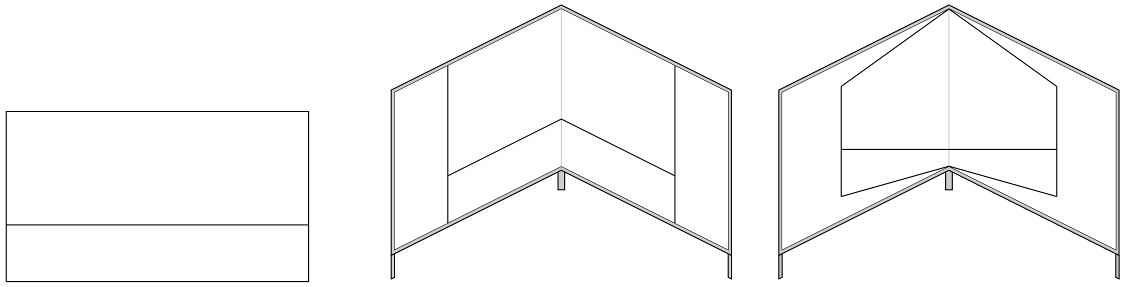
(b) Second scene.

(c) Third scene.

Figure 6.7: Bubbleplots of the sound source positions in the horizontal plane, taken every five frames for the: (a) first, (b) second, and (c) third scenes. At each couple (X, Y) where sources are present, a circle is drawn with a diameter proportional to the number of sources at that position. The line color of each circle also corresponds to the number of sources at the center of the circle, but this is mostly for readability. The diameters can be compared across the three figures. Some very distant source positions are not shown for clarity. The panels of the SMART-I^2 are represented by the Λ (inverted "V"), in black.

6.4 Experimental setup

To guarantee that the sound did not suffer from unwanted reflections, it was decided to have the spectator face the corner of the SMART-I^2. Since we wanted to approximate cinema conditions, it was necessary to compensate for this geometry, and project the video as if the screen was planar. The open source s-3D movie player **Bino**, which is compatible with the **Equalizer** library [Eilemann et al., 2009], was used to read the video stream. This allowed us to project onto the specific screen configuration and to obtain a result close to one that would be obtained on a regular planar screen, for a specifically defined viewing position. Figure 6.8 compares, for a given image, the result of the projection of this image on the corner of the SMART-I^2 with and without **Equalizer**. The difference was mainly seen at the top and bottom of the image, where trapezoidal (or *keystone*) distortion was visible (Figure 6.9). The subjective impact of this distortion is considered in the next chapter.



(a) Projected image.

(b) Without **Equalizer**.

(c) With **Equalizer**.

Figure 6.8: Illustration of the impact of **Equalizer** on the projection of images on the corner of the panels in the SMART-I^2. This figure compares the case where the image presented in (a) is projected on the corner of the SMART-I^2 (b) without **Equalizer** and (c) with **Equalizer**.

As the image does not fill the whole surface of the two panels, the audio window is capable of rendering objects, which are effectively outside the video window. For example, if the spectator is seated 1 m behind the SMART-I^2 origin (Figure 5.5), this gives a field of view of about 64°, and an audio field of about 122°. Therefore, sound sources can be effectively associated to visual objects that have left the visual scene or visual objects that are about to enter the visual scene.

The movie files resulting from the work in this chapter are distributed on two computers and consist of (1) the publicly released s-3D video (the images have not been re-rendered), (2) the control file (an audio metadata OSC sequence) sent

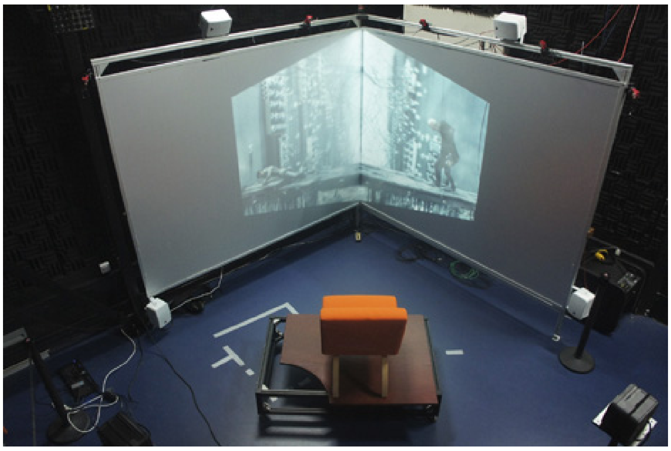


Figure 6.9: Photo of the SMART-I^2 installation for cinema projection.

from the movie player to the audio controller, and (3) the collection of audio files. Although the metadata are computed off-line, the audio engine in the SMART-I^2 performs the rendering in real-time. This approach allows for the audio engine to be changed to other spatial rendering systems, such as binaural or Ambisonics for future possible comparisons.

The audio metadata are sent at each new frame by the s-3D movie player **Bino**, which was modified to include an OSC client (Appendix C). This ensures that frame synchronization is maintained throughout the video playing.

A block-diagram of the global software and hardware architecture used in the experiment described in this chapter is shown in Figure 6.10.

6.5 Creation of the audio track

This section describes and explains the procedure followed in order to achieve the object-based sound re-mix.

The audio track of the movie was only available in the final downmix version (stereo and 5.1), with some additional rough mixes of most of the dialogs and music tracks. The original multitrack audio master was not available. It was therefore necessary to create a new audio master with each object corresponding to a separate track, in order to allow for position coherent rendering. The aim was to recreate an audio track similar to the original track. The available dialog and music dry tracks were retained. The rest of the audio elements were created from libraries and recorded sounds, with one audio file per object.

The result was an object oriented multitrack audio master that contained in-

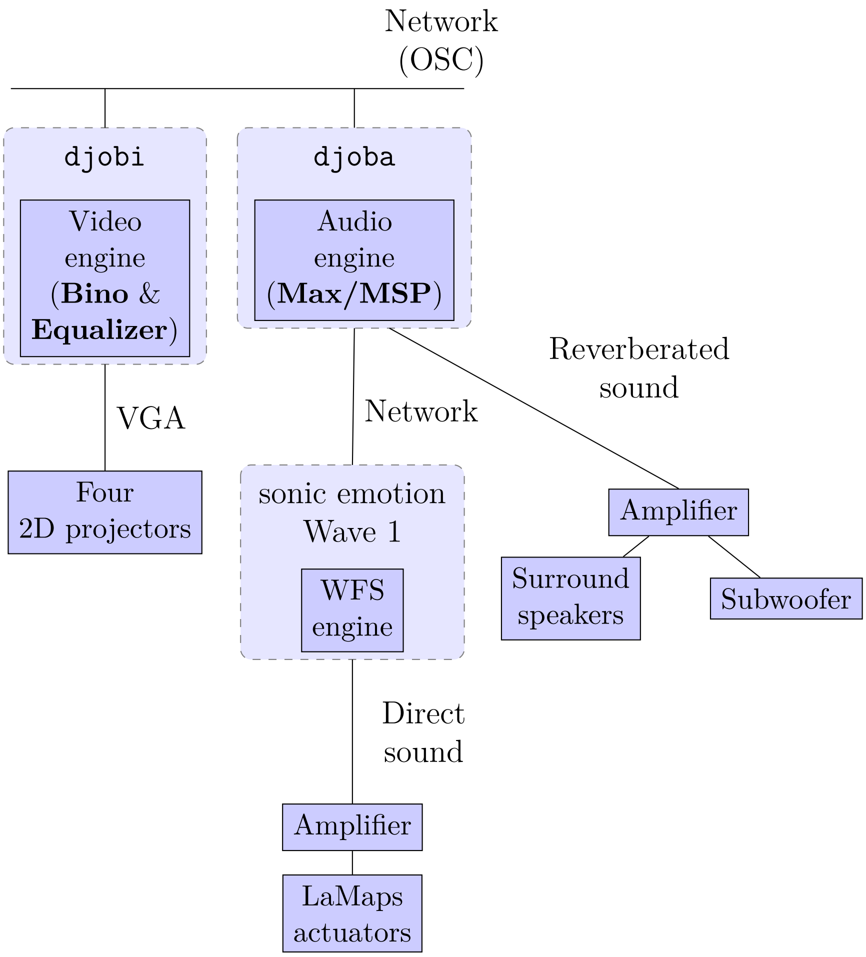


Figure 6.10: A block-diagram of the software and hardware architecture used in the experiment described in the present chapter. See the text and Section 5.2 for explanations.

dividual audio tracks for each individual audio object, allowing for individual rendering positions to be defined and controlled. The global workflow describing the creation of the new audio track is shown in Figure 6.11. This workflow is further discussed in the following sections.

6.5.1 The audio editing process

The aim was to recreate an audio track similar to the original track. The available dialog and music tracks were retained. The rest of the audio elements were created from libraries and recorded sounds. They were edited using a standard digital audio workstation (DAW) software, with each object being a discrete audio file.

The film audio track was first edited with the DAW, using a conventional stereo configuration. Each audio element was stored in the software database as a single

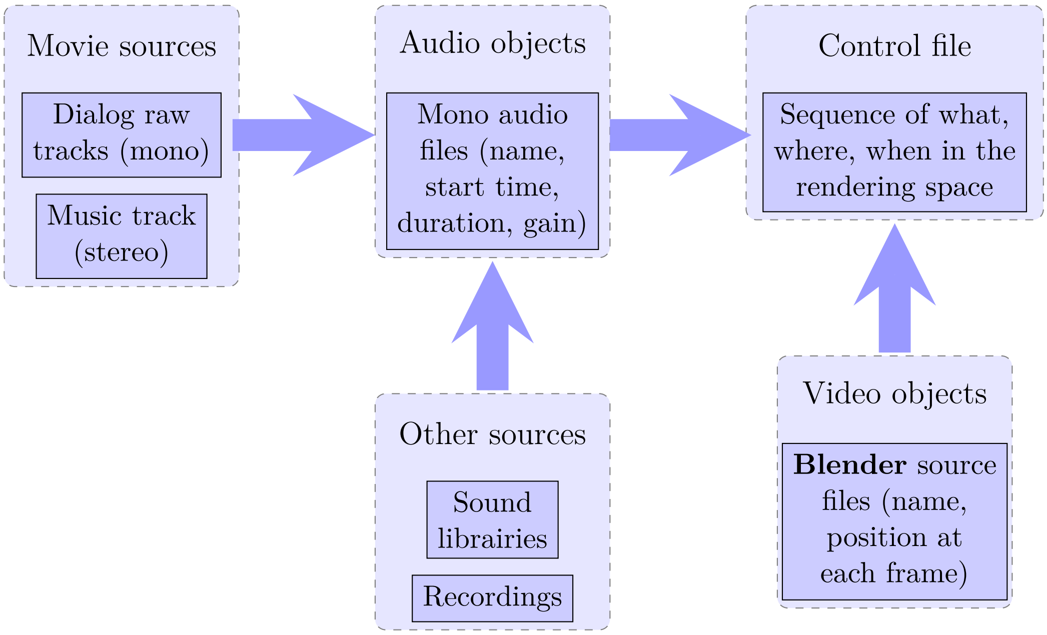

Movie sources
Dialog raw tracks (mono)
Music track (stereo)
Audio objects
Mono audio files (name, start time, duration, gain)
Control file
Sequence of what, where, when in the rendering space
Other sources
Sound librairies
Recordings
Video objects
Blender source files (name, position at each frame)

Figure 6.11: Global workflow describing the creation of the new audio track.

mono audio file. The audio file length was optimized to match the effective audio duration, and time-aligned within the mix. At the end of the process, a rough mix was generated to roughly adjust the sound levels for each track, but without any other form of processing. Since the number of available tracks was limited to 48 in the DAW, similar audio files that were not simultaneously read, were placed within the same track. Therefore, within a given track, a single level gain set in the DAW was applied to all the audio files. If the level gain of some of these audio files was to be adjusted independently of each other, this was accomplished by applying a destructive gain adjustment to them.

All these audio files were named using a formalized naming convention. For example, the filename `03_14_BodyImpact.wav` corresponds to scene number `03`, **Blender** "sub-scene" file number `14`, and the sound for the collision sound between the two characters ($t = 02$ min 16 s).

6.5.2 A spreadsheet as mixing table

In order to create the audio sequence, a list had to be made with all the required information. The solution chosen was to use a spreadsheet to create a sequencing table, in which all the various audio file names, their temporal information, and the **Blender** object reference corresponding to them were listed. The spreadsheet includes:

- an arbitrary and unique identifier of the audio file,
- the reading start time expressed in video frames,
- the duration in frames,
- the channel (1–16) through which the audio file will be read in the WFS processor, noting that no two files can be read simultaneously on the same channel,
- a two digit number defining the corresponding **Blender** object,
- the sound level gain expressed on a linear scale.

These values were manually extracted by reading the information directly off the DAW edit window, and were then entered in the spreadsheet. Since the audio metadata are sent by the s-3D movie player **Bino** (see Section 6.4), the timeline is frame-based in our application. The DAW version used did not provide a frame-based timeline, but rather a standard time-based scale. We thus had to include a time-to-frame conversion, with results rounded to the nearest integer frame number. This means that the largest time synchronization error is on the order of 21 ms, corresponding to half the duration of a frame: $1/24 \approx 42$ ms. While this value is not negligible, it is sufficiently small to be acceptable in terms of timing.

For future work, this export to a spreadsheet should be performed automatically, by choosing a software tool with such an embedded function. Another option would be to develop a customized script on an open source product.

6.5.3 The scripting process

This audio mix spreadsheet, as well as all object coordinates extracted from **Blender** were exported to two comma-separated value (CSV) documents, in order to simplify their processing.

Several scripts written in **Python** were then successively applied on these CSV files, to produce the final control file that were then read by the s-3D movie player.

The first script performs the first part of the coordinate transformations (as described in Sections 6.3.1 and 6.3.2). It is necessary to stop the transformation process at that point to be able to account for compositing. Several scenes in “Elephants Dream” are indeed composited, meaning that the **Blender** file contains different 3D scenes that are superimposed only later during the rendering process, after the camera projection from 3D to 2D. This implies that the same pre-rendered image position can produce different spatial positions as determined by each s-3D rig configuration (see Section 6.3). For example, the first scene is made of one 3D scene for the background, and of another for the characters, using two different s-3D rigs at different positions.

The objects corresponding to the different composited scenes are then united by a second script, which also gathers all the information from the audio mix, and computes the final basis transformation.

The resulting control file contains the sequence of OSC messages to be sent to the audio engine, as well as the frame numbers for timing, to ensure synchronization between the video and audio streams.

6.6 Discussion

During this work we came across many questions regarding some technical and artistic problems. We will review here some of the technical problems, and indicate how we solved them. For the artistic problems and their solutions, we refer the reader to [Évrard et al., 2011]. The artistic problems were related to the transitions between different scenes or shots, and to the Doppler effect in the second scene.

6.6.1 The sequencing table

One of the first technical question concerned the sequencing table. During the development of the project, we realized that a simple spreadsheet was not well suited due to its expanding size and complexity. The number of entries was increasing rapidly, although the project was limited to only three scenes of the original animation movie, and to sixteen concurrent sound sources. For example, at the end of the project, the number of sound objects to control reached 150.

Several spelling or numbering errors were not detected before the final reading stage with the s-3D player, causing a loss of time.

The use of a database would greatly simplify and provide a much more reliable solution. First, it is much easier to manage large amounts of data within a database, in comparison to a simple spreadsheet. Second, with a database, the exported file structure could easily be maintained, and kept constant despite changes made in the internal table structure. Finally, each element type could be entered through a dedicated form of the database, checked independently before proceeding to the required matchings in the relational tables, and finally exported to the CSV file format.

6.6.2 Incoherence in the perceived visual and audio distances

During the mixing of the film we realized that the perceived sound distances did not always match the perceived visual distances resulting from stereoscopic fusion of the images. There is a particular scene in the film where the characters stand on a bridge (Figure 6.12, t = 01 min 04 s), for which the effect was particularly noticeable.

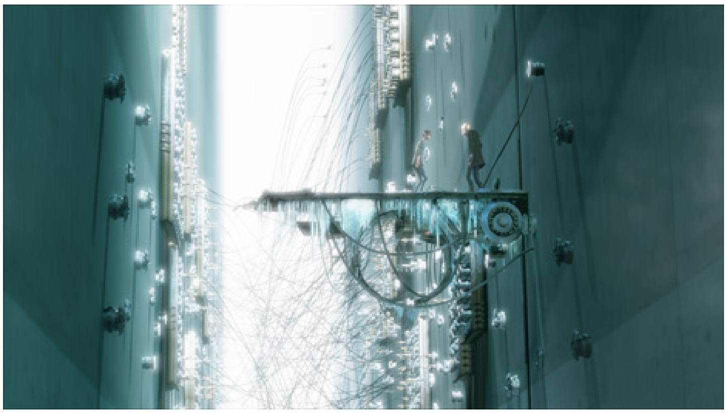


Figure 6.12: Image from the second scene (t = 01 min 04 s). In this scene, the camera is located far away from the action (the characters look small), but the stereoscopic capture parameters are set so that the characters appear close to the convergence plane, resulting in a conflict between the distance suggested by linear perspective and the distance suggested by stereoscopy.

In this particular scene, the camera is located far away from the characters (with respect to their size), but the stereoscopic capture parameters are set so that the characters appear close to the convergence plane. Since the convergence plane is located at a distance of about 2 m from the spectator in the SMART-I^2, the sound obviously appears to come from a close location. However, the characters are small on the screen, which, by our learned experience with perspective, is interpreted as the characters being very distant. This optic zoom effect is not taken into account in the developed geometrical basis transformations. The visual object's position is correctly determined, with the error being more a conflict between perception and cognition.

We searched for a comparable effect in audio. The best solution found that was both simple and functional was to take into account, for every object, two different distances, which were passed to the audio engine. The first distance is the actual distance computed from the model in Section 6.3. It is used as the distance between the spectator and the WFS virtual point source. The second distance is the distance from the camera to the object in "**Blender** Units" (BU). It is sent to the reverberation engine to modify the level of reverberant energy produced by the surround loudspeakers, for the given source. However, the reverberation engine interprets this value as expressed in m, although it was originally expressed in BU. This resulted in too much reverberation in certain scenes. It was determined through trial-and-error that using one-third the distance from the camera to the object allowed sufficient intelligibility, while providing suitable distance cues, thereby resolving this audiovisual distance conflict.

6.6.3 Manual coordinate definition for sound objects

For most objects, the coordinates we had automatically extracted were used. Here, we present some cases where coordinates needed to be manually created in **Blender**, in order to position sound objects as desired or required.

6.6.3.1 Large size objects

In the case where an object occupies a significant portion of the screen, the selection of which of its visual points will be matched to the sound object is crucial. As an example, in a close-up shot on a character's face (Figure 6.13, t = 00 min 26 s), choosing the lips, center of the head, or center of the character would result in a different sound localization perception.

However, these visual elements are not individually defined as objects *per se* in **Blender**, but as movable elements of the main object, i.e. the character, and are thus not accessible by the current script used to extract object coordinates. For instance, the tongue or any other part of the character are not actual objects; the entire character is represented as a single object in **Blender**. It was thus necessary to manually create invisible objects in the **Blender** scene that matched the selected audio element locations and motions, in order to be able to extract the coordinates of each element of interest, and transmit them to the audio rendering engine.

Figure 6.13: Image from the first scene, illustrating a large size visual object (t = 00 min 26 s).

6.6.3.2 Non-localized sound objects

For some objects such as general water flowing or wind noise, there was no real precise spatial position reference, as there was no associated visible object. This type of sound source is better suited to an immersive surround perception.

In this case, invisible virtual objects were manually created at fixed arbitrary coordinates, with the "reverberation distance" (as defined in Section 6.6.2) set to a large value. The aim was to increase the room effect, and thus the immersive perception of the sound, while minimizing any perception of the direct sound, and hence, any localization cues.

6.6.3.3 Music track

As the music track of the original stereo mix was available, it was retained in the new audio track. The two tracks of the stereo mix were manually positioned in the WFS array as virtual objects, slightly past the edges of the image (+50° and −50°), corresponding to virtual loudspeakers on the screen.

6.7 Conclusion

In this chapter, we presented the implementation of a true 3D audio track, for an s-3D animation movie using object-based sound mixing.

The context and the infrastructure were presented, as well as a detailed review of the mathematical development required to obtain the coordinates of the sound sources in the SMART-I^2 coordinate axes from the coordinates of the visual objects in the movie source files.

The creation of the audio track was also described in detail from the early editing stage, through the scripting process, and terminating with the sending of data to the rendering system.

In Chapter 7, the newly created 3D audiovisual content is used in an experiment examining the effect of the choice of audio rendering support. In particular, we compare traditional sound rendering, namely stereo, and the highly precise spatial sound rendering of the SMART-I^2, namely WFS.

Apart from our own particular use of this content, we believe that this work could also serve others as a good basis for their future research. While writing this chapter, our aim was to describe, in the most open way, the various steps taken during the course of this project, in order to allow future researchers, or artists to begin their project with some solid theoretical basis that will help them to start and progress faster. We refer the interested reader to [Évrard et al., 2011] for some aspects related to the creative side of the process.

Bibliography

Eilemann, S., Makhinya, M., Pajarola, R., 2009. Equalizer: A scalable parallel rendering framework. IEEE Trans. Vis. Comput. Graphics 15 (3), 436–452.
http://dx.doi.org/10.1109/TVCG.2008.104 141

Évrard, M., André, C. R., Verly, J. G., Embrechts, J.-J., Katz, B. F. G., 2011. Object-based sound re-mix for spatially coherent audio rendering of an existing stereoscopic-3D animation movie. In: Audio Eng. Soc. Conv. 131. New York, NY.
http://www.aes.org/e-lib/browse.cfm?elib=7628 126, 146, 150

Hartley, R., Zisserman, A., 2004. Multiple View Geometry in Computer Vision, 2nd Edition. Cambridge University Press. 135

Olaiz, N., Arumi, P., Mateos, T., Garcia, D., 2009. 3D-Audio with CLAM and Blender's Game Engine. In: Proc. Linux Audio Conf. 2009. Parma, Italy. 128

Woods, A. J., 1993. Image distortions in stereoscopic video systems. In: Proc. of SPIE 1915. San Jose, CA, pp. 36–48.
http://dx.doi.org/10.1117/12.157041 130, 132, 133, 138

Chapter 7

Impact of 3D sound on the reported sense of presence

Highlights

- ✓ We consider, in the s-3D cinema context, the cognitive differences between stereo sound and Wave Field Synthesis.
- ✓ The sense of presence is evaluated with a post-session questionnaire and heart rate monitoring.
- ✓ A between-subject experiment is designed with three different soundtracks.
- ✓ The sound condition does not affect the reported presence score directly for all subjects.
- ✓ The sound condition only impacts on the sense of presence of the group of participants who reported the highest level of presence.
- ✓ The analysis of the participants' heart rate variability shows that the frequency domain parameters correlate to the reported presence scores.

Contents

7.1 Introduction

As stereoscopic 3D (s-3D) cinema aims at providing the spectator with a strong impression of being part of the movie, there is a growing interest in the sense of presence induced by the media. Presence (or more accurately, telepresence) is a phenomenon in which spectators experience a sense of connection with real or fictional environments and with the objects and people in them [Lombard et al., 2009]. Previous research has shown that the addition of stereoscopic information to a movie increases the sense of presence reported by the spectators [Ijsselsteijn et al., 2001]. We hypothesize here that the spatial sound rendering quality of an s-3D movie impacts on the sense of presence as well.

The experiment reported in this chapter, previously published in [André et al., 2012], considers, in the cinema context, the cognitive differences between a traditional sound rendering (stereo, see Section 3.1.1), and a highly precise spatial sound rendering (Wave Field Synthesis or WFS, see Section 3.3.4). In particular, it will be examined whether a higher spatial coherence between sound and image leads to an increased sense of presence for the audience. The current chapter therefore presents the results of a perceptual experiment using a common video track and three different audio tracks. Using a post-stimuli questionnaire based on previous reports regarding the sense of presence, various cognitive effects are extracted and compared.

7.2 Soundtracks

The visual stimulus used for this experiment was the movie clip described in Chapter 6. The soundtrack of the movie clip was the independent variable in our experimental design. The original stereo soundtrack, the object-oriented soundtrack created in Chapter 6, and an additional "hybrid" soundtrack are used. These different soundtracks are now described and analyzed objectively.

7.2.1 Different spatial sound renderings

Three different soundtracks were used in this experiment. The first soundtrack was the original stereo soundtrack, termed `STEREO`. This soundtrack was rendered on the WFS system by creating two virtual point sources at ±30° in the (virtual) screen plane, roughly at the left/right edges of the image. The object-oriented soundtrack introduced in Chapter 6, termed `WFS`, was the spatially coherent rendering. This new audio track was created specifically as part of this thesis, but was inspired by the original `STEREO` audio track (Section 6.5). Because the `STEREO` and `WFS` soundtracks differ in content and spatial rendering quality, an ideal stereo mix was constructed using the same metadata and audio files as in the `WFS` version. The panning of each object in this mix was automatically determined according to a sine panning law relative to the object's actual position (the same as for the `WFS` version). One way to simulate a sense of distance in stereo is to adapt the intensity of the signal (Section 3.1.1). Therefore, a r^{-2} distance attenuation factor was applied to the intensity of each object's signal to simulate a sound originating at a distance r from the spectator. This hybrid soundtrack, termed `HYBRID`, thus had the same content as the `WFS` track, but was limited in its spatial rendering quality. The HYBRID track was rendered over the same two virtual point sources as the STEREO track. A graphical comparison between the three soundtracks is given in Figure 7.1.

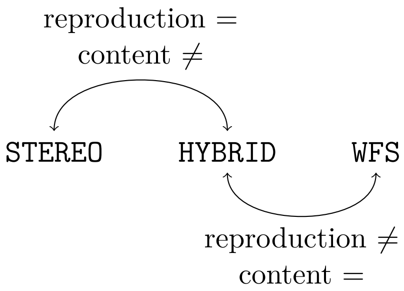


Figure 7.1: Comparison of content and spatial reproduction of the soundtracks.

Due to differences between the soundtracks (Figure 7.2), a global equalization across the entire movie was inappropriate, and resulted in distinctly different perceived levels. Therefore, it was decided to equalize for an element that was common to all conditions. One character line, at t = 00 min 22 s, duration 4 s, was chosen as it was common to all three soundtracks (dialog tracks were identical) and background sounds were minimal at that moment. This audio calibration segment was adjusted to 61 dBA, measured at the viewer's head (ambient noise level of 33 dBA).

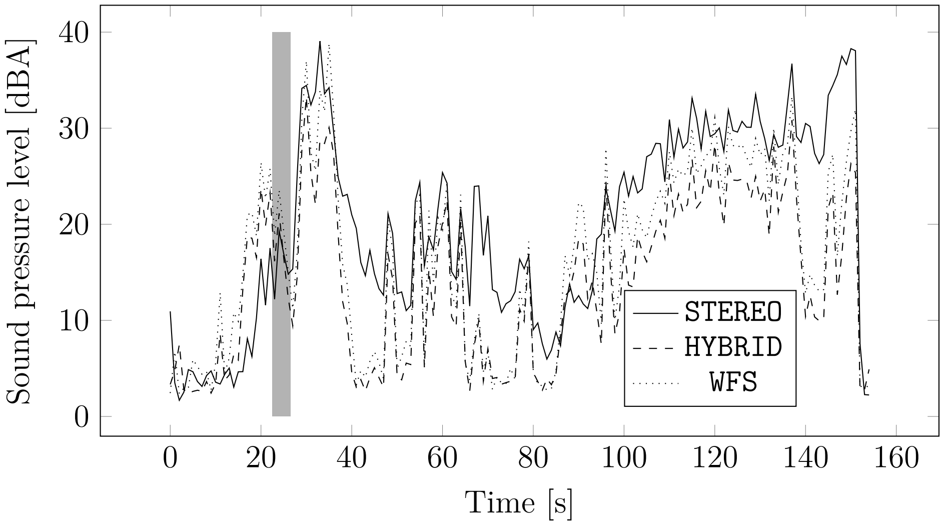


Figure 7.2: Evolution of the sound pressure level measured at the listener's head position against time for the three sound modalities. The gray region corresponds to the time interval used for equalization.

7.2.2 Audio sweet-spot effect

It should be noted that all participants in this study were located at the sweet spot of the rendering system, and they could thus enjoy the best sound reproduction. The impact of an off-axis seating would certainly be more pronounced for the `HYBRID` soundtrack than it would be for the `WFS` soundtrack as the process of stereo panning relies on the proper positioning of the listener in the sweet-spot. Indeed, taking into account the geometry of the reproduction system, the sweet-spot of the stereo reproduction has a width of merely 10 cm according to [Theile, 1990]. When the listener is outside the sweet-spot, sources tend to be attracted to the closer speaker position. On the other hand, the ability of WFS to reproduce a sound position independently from the listener position [Theile et al., 2003], combined with the ventriloquism effect [Thurlow and Jack, 1973], would result in a larger sweet-spot because the sound location is preserved when the listener is off-axis but can still be perceived as coming from the visual object. The congruence in that case is limited by the difference in audio and video perspectives that can be detected by the spectator [de Bruijn and Boone, 2002]. The question of the congruence between s-3D images and WFS virtual sound sources is addressed in Chapter 8.

7.2.3 Objective analysis

An objective analysis of the rendered audio was performed. A binaural recording of each condition was made with an artificial head placed at the sweet-spot, equivalent to the spectator position during the subsequent experiment. The evolution of the relative sound level at the listener position for the three conditions was measured using a 1 s sliding window and averaged over both ears.

Outside of the region used to calibrate the three conditions, the `STEREO` soundtrack has a higher level at several moments. This is due to the difference in audio content, as the original track contained a richer audio mix. Some differences are observed between the `WFS` and `HYBRID` conditions. The different spatialization processes lead to slight differences in sound level that cannot be compensated exactly using only a main volume equalization (Figure 7.2).

The perceived distribution of the sound sources is of interest. The interaural level differences (ILDs) and the interaural time differences (ITDs) are thus computed from the binaural signals (Section 2.1.1). Binaural signals are subdivided into 1 s segments and analyzed in third-octave bands to obtain ILD and

ITD values, using the **Binaural Cue Selection** toolbox for **Matlab** [Faller and Merimaa, 2004]. These values are then averaged across pertinent frequency bands (< 1.5 kHz for ITD, > 1.5 kHz for ILD [Blauert, 1997]). The threshold value of 1.5 kHz also corresponds to the SMART-I^2 WFS aliasing frequency, meaning that the ITD should be reliable, but the ILD might suffer from bias.

Table 7.1 presents the means, standard deviations, and skewness of the obtained values. In both cases, the mean decreases from STEREO to WFS to HYBRID. All means are statistically different from each other, except when comparing the HYBRID and WFS ITD means (one-sided Wilcoxon rank sum test, at the 0.05 level). One would also expect that the cues are more spread out for WFS than for HYBRID. This is the case since the standard deviation increases from STEREO to HYBRID to WFS for both ITDs and ILDs.

Table 7.1: Means, standard deviations, and skewness of the computed ILDs and ITDs as a function of SOUND CONDITION.

ITD	Mean [ms]	Std dev. [ms]	γ [/]
STEREO	-0.0012	0.0445	-0.52
HYBRID	-0.0112	0.0543	-0.84
WFS	-0.0106	0.0596	0.71

(a) ITDs [ms].

ILD	Mean [dB]	Std dev. [dB]	γ [/]
STEREO	-0.16	0.21	1.50
HYBRID	-0.45	0.22	0.26
WFS	-0.27	0.29	-0.19

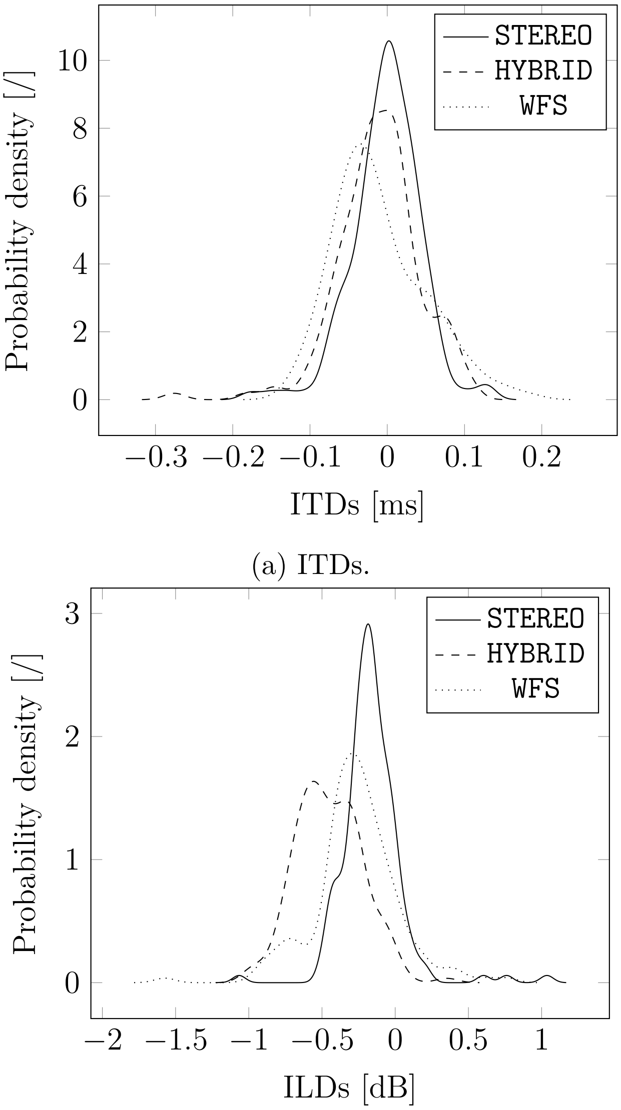
(b) ILDs [dB].

Distributions of mean ILDs and ITDs are shown in Figure 7.3. In both cases, the peak of the probability density function is higher for STEREO than it is for HYBRID and WFS. This confirms that the HYBRID and WFS localization cues are more distributed or spread out than those for the STEREO condition.

7.3 Method

Thirty-three (33) subjects took part in the experiment (26 men, 7 women, age 16 to 58 years, mean = 30.66, stdev = 10.77). They answered to a call for participants describing a “3D cinema experiment”. Each was compensated with a soft drink and a cookie while filling out the post-session questionnaire.

To determine whether or not the sound modality impacts on the reported sense



(a) ITDs.

(b) ILDs.

Figure 7.3: Estimates of the probability density functions of the mean interaural time differences (ITDs) and interaural level differences (ILDs) obtained for each soundtrack.

of presence, a between-subjects experiment was designed. The three different soundtrack conditions, `STEREO`, `HYBRID`, and `WFS` (Section 7.2) were used as an independent variable. Each participant was assigned randomly to one particular condition, with 11 participants presented with each soundtrack.

In order to assess the sense of presence as a dependent variable, two methods were used. A post-session questionnaire was developed, providing a subjective assessment. In addition, an oxymeter was used to continuously measure the heart rate of the participants. The goal was to compare this objective measure, which will be further described in Section 7.3.4, with the presence score obtained with the questionnaire. The heart rate was measured at 60 Hz using a finger mounted pulse oxymeter (CMS50E, Contec Medical Systems Co.). A picture of the oxymeter is shown in Figure 7.4.

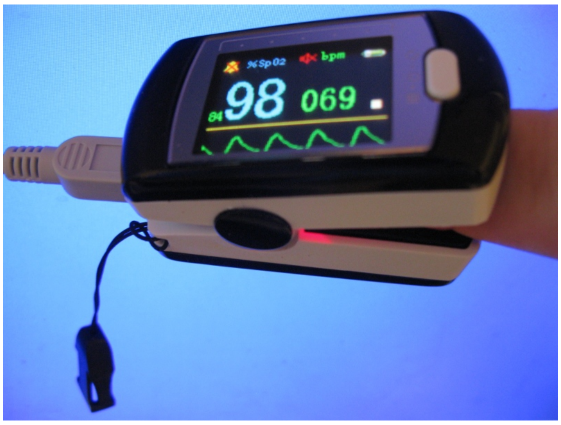


Figure 7.4: The CMS50E finger mounted pulse oxymeter.

We hypothesize that the spatial rendering quality of sound will impact on the reported sense of presence, as measured by the questionnaire. It is also hypothesized that measures extracted from the heart rate signal will reflect a change from baseline due to the movie presentation and that this change in value is linked to the spatial rendering quality of sound.

7.3.1 Experimental setup

A block-diagram of the software and hardware architecture used in this chapter is presented in Figure 7.5. The architecture of the last chapter is reused to read the 3D audiovisual content created in Chapter 6. Two additional laptops are used in this experiment, one running the online survey software **LimeSurvey** to collect the participants' answers to the presence questionnaire, and the other receiving the data from the oxymeter on a serial connection over USB.

We wrote the custom software running on the laptop 1 in **Python** by gathering different programs into one, tailored to our needs. It used Greg Pinero's software oscilloscope written for the Arduino electronic chip[*] based on Eli Bendersky's demo of plotting in **Python** using **matplotlib**. The serial communication was established thanks to code found on the internet[†], and then fed as input to the

[*] https://github.com/gregpinero/ArduinoPlot, last accessed 24/09/2013.

[†] http://stackoverflow.com/questions/1093598/pyserial-how-to-read-last-line-sent-from-serial-device, last accessed 24/09/2013.

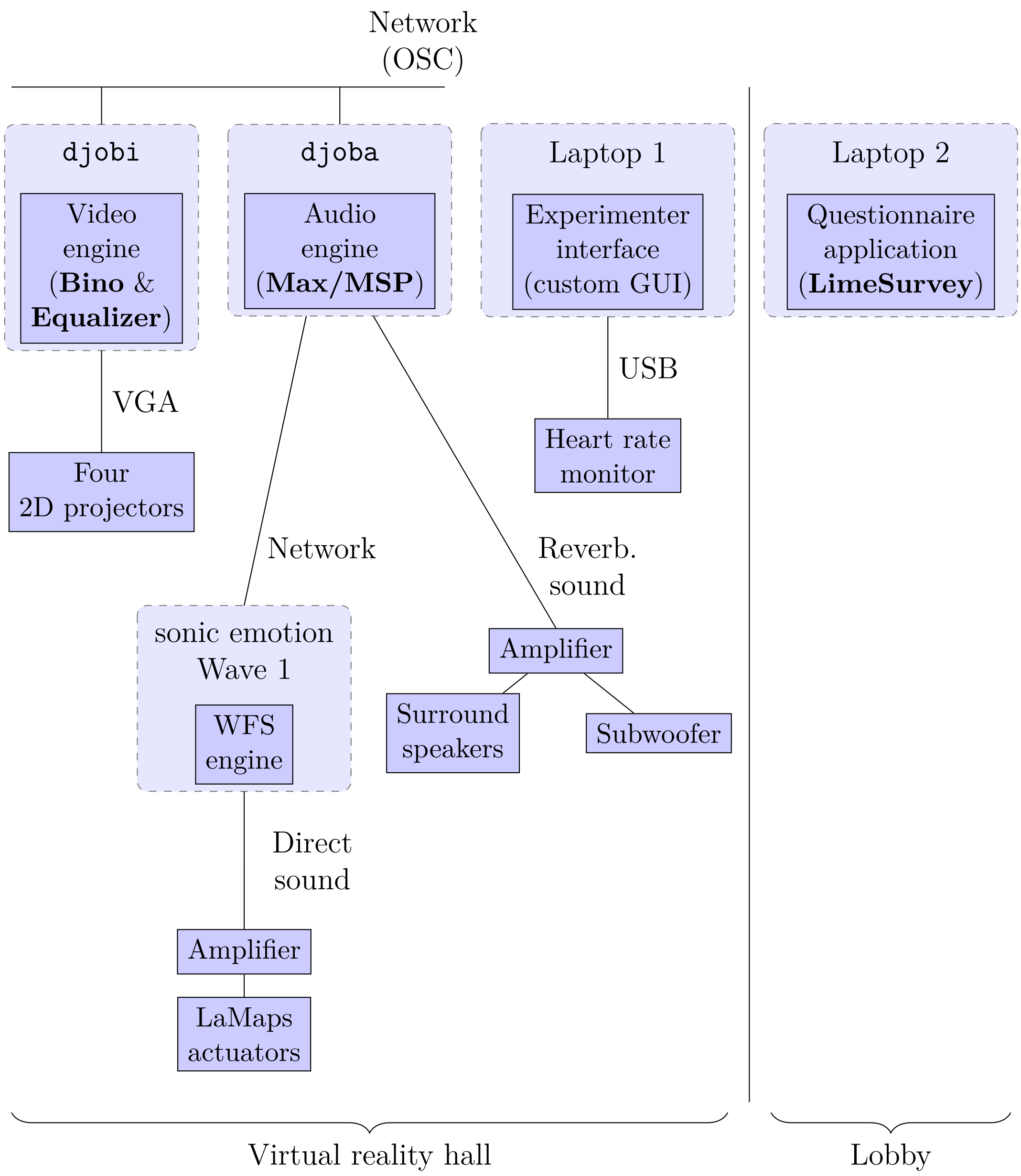


Figure 7.5: A block-diagram of the software and hardware used in the experiment described in the current chapter.

oscilloscope.

Thanks to information on the oxymeter available on the internet*, it is quite easy to read the data it sends to the serial port. The serial port was checked ten times per second. The graphical user interface (GUI) plotted the different values sent by the oxymeter in real time, which allowed for a quick visual control on the data, in particular to check whether the oxymeter was correctly placed. The heart rate signal values were saved, along with a time-code, in a comma-separated value (CSV) file for each subject.

7.3.2 Procedure

Each participant was seated in a comfortable chair (see Figure 6.9) in front of the SMART-I^2 and was provided with written instructions regarding the experiment. The oxymeter was placed at the tip of the middle-finger of his/her left hand. The participant was left alone in the experimental room. The room was then completely darkened for a period of 30 s after which the movie was started from a remote control room. This allowed the participant to accommodate him/herself to the darkened environment, and to approach a “cinema” experience. At the end of the movie, the participant was directly taken to the lobby of the virtual reality hall to complete a questionnaire.

7.3.3 Post-session questionnaires

We created a presence questionnaire using three groups of questions gathered from different sources previously reported. The complete list of questions, as well as an example of the presentation, is given in Appendix E.

The first group came from the Temple Presence Inventory (TPI) [Lombard et al., 2009], a 42-item cross-media presence questionnaire. The TPI is subdivided into eight groups of questions that measure different aspects of presence. These subgroups, or components, are given in Table 7.2 with the associated number of questions. The sensitivity of the TPI to both the media form and the media content has been previously confirmed [Lombard et al., 2009]. The second group of questions was taken from the short version of the Swedish Viewer-User Presence (SVUP-short) questionnaire [Larsson et al., 2007]. We selected three questions

*http://sourceforge.net/apps/mediawiki/sleepyhead/index.php?title=CMS50X, last accessed 24/09/2013.

regarding the sound rendering. Finally, the third group of questions, which measured negative effects, were from Bouvier's PhD thesis [Bouvier, 2009].

Table 7.2: The eight components in the Temple Presence Inventory, and the associated number of questions. From [Lombard et al., 2009].

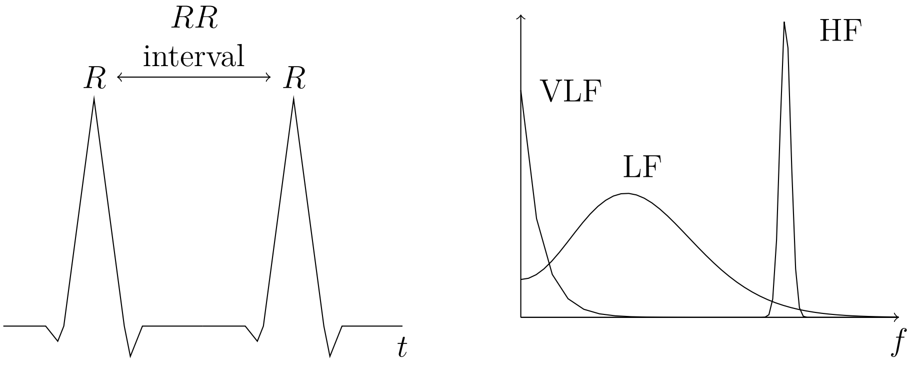
Factors	Numbers of questions
Spatial presence	7
Social presence – actor within medium	7
Social presence – passive interpersonal	4
Social presence – active interpersonal	3
Engagement (mental immersion)	6
Social richness	7
Social realism	3
Perceptual realism	5

The resulting questionnaire was translated into French. Each question was presented using a 7-point radio button scale, with two opposite anchors at the extreme values, resulting in a score between 1 and 7. Composite scores were calculated as the mean results for all items in each group.

The main score of interest is the global score obtained with the TPI, termed `TEMPLE`. Of all the components in the TPI, the scores "Spatial presence" (`SPATIAL`) and "Presence as perceptual realism" (`PERCEPTUAL_REALISM`) are expected to be significantly varying with the media form [Lombard et al., 2009]. The `SWEDISH` score, from the SVUP-short, gives additional information on the perception of each sound condition. The `NEGATIVE` score, from Bouvier's PhD thesis, allows one to discard participants who experienced discomfort.

7.3.4 Heart Rate Variability

Heart Rate Variability (HRV) describes the changes in heart rate over time. Several studies have used HRV as a physiological measure in experiments involving virtual reality [Huang et al., 2008; Slater et al., 2006]. Standards exist [HRV, 1996] describing the different measures that can be extracted from an electrocardiographic (ECG) record. Although HRV is calculated from time intervals between two heart contractions (RR intervals) in an ECG signal, it has been shown that it is possible to obtain the same results from peak-to-peak intervals given by a finger-tip photoplethysmograph (PPG) [Selvaraj et al., 2008]. This small device monitors the relative blood volume changes using an infrared light source and a phototransistor at the fingertip. The heart rate is derived from the PPG pulsatile waveform. Since the signal is captured at only one point on the body, the PPG is less intrusive than the ECG. Analysis of the resulting HRV data was performed in both the time domain and the frequency domain (Figure 7.6).



(a) Time domain.

(b) Frequency domain

Figure 7.6: Analysis of the heart rate signal. (a) A peak detection algorithm determines the time series of RR interval which yields the time domain parameters. (b) Power spectral density estimation of the time series yields the frequency domain parameters. See the definitions of VLF, LF and HF in the text.

The majority of time domain HRV measures require recordings longer than 5 min, which are not possible due to the duration of the film excerpt used. Only the following measures were calculated:

- MeanRR - mean RR interval [ms]
- MinRR - shortest RR interval [ms]
- MaxRR - longest RR interval [ms]
- ΔRR - difference between MaxRR and MinRR [ms].

Frequency domain measures obtained through power spectral density estimation of the RR time series are of particular interest, since their evolution has been correlated with positive or negative emotions when presenting movie clips [Vianna and Tranel, 2006].

In the case of short-term recordings (from 2 to 5 min), three main spectral components are distinguished [HRV, 1996]: the very low frequency (VLF) component between 0.003 Hz and 0.04 Hz, the low frequency (LF) component between 0.04 Hz and 0.15 Hz, and the high frequency (HF) component between 0.15 Hz and 0.4 Hz. Instead of the absolute values of VLF, LF, and HF power components in ms^2, the values are expressed as LFnorm and HFnorm in normalized units (n.u.), which represent the relative value of each component in proportion to the total power minus the VLF component:

$$\text{LFnorm} = \frac{\text{LF}}{\text{Total power} - \text{VLF}}, \tag{7.1}$$

and similarly for HFnorm.

The parasympathetic activity, which governs the HF power [Stein and Kleiger, 1999], aims at counterbalancing the sympathetic activity, which is related to the preparation of the body for stressful situations, by restoring the body to a resting state. It is believed that LF power reflects a complex mixture of sympathetic and parasympathetic modulation of heart rate [Stein and Kleiger, 1999]. Emotions such as anger, anxiety, and fear, which correspond to the emotions elicited by our movie clip, would be associated to a decreased HF power [Kreibig, 2010].

7.4 Results from post-session questionnaires

This section presents the results of the statistical analysis results carried out on the questionnaire data using the statistical software **R**, as well as several packages listed in the appropriate sections below.

7.4.1 Treatment of missing values

There were 10 answers (out of 2785) left blank in the questionnaire results. To avoid discarding the corresponding participants, multiple imputations of the incomplete dataset were used to treat these missing values. This was done using the package **Amelia II** [King et al., 2001]. Multiple imputation builds m (here five) complete datasets in which each previously missing value is replaced by a new imputed value estimated using the rest of the data. Each imputed value is predicted according to a slightly different model and reflects sampling variability.

In the subsequent analysis, the analyses of variance (ANOVAs) were carried out with the package **Zelig** [Imai et al., 2008]. F-statistics and their associated p-value were estimated according to the method given in [Raghunathan and Dong, 2011], resulting in ANOVAs with degrees of freedom which are no longer integers. More information on the statistical methods of this section is given in Appendix F.

7.4.2 Negative effects

It is necessary to verify that no participant suffered physically from the experiment. The initial analysis of the results considers the `NEGATIVE` group of questions, measuring negative effects induced by the system, such as nausea, eye strain, or headache.

We carried out this analysis with a bivariate boxplot, or *bagplot* [Rousseeuw et al., 1999]. The bagplot (see Figure 7.7) is a bivariate version of the traditional boxplot. Its main components are a *bag*, which comprises 50% of the data points, and a *fence*, which separates inliers from outliers. The region outside the bag but inside the fence is called the *loop*. It is usual to present a mark at the median in a boxplot. The equivalent of this mark in the bagplot is represented by the white region in our plots. A bagplot of the `NEGATIVE` score versus the `TEMPLE` score (Figure 7.7), indicated that participant 23 was an outlier, reporting feeling much worse than the other participants. This participant was therefore discarded from the study. All others obtained a `NEGATIVE` score less than 2.17 (minimum possible value = 1), which can be considered as having experienced little or no negative effects during the experiment. Also, the presence score seems relatively independent of the experienced negative effects. Suffering from (little) negative effects does not seem to reduce the feeling of presence. Conversely, a low presence score is not necessarily associated to the experience of negative effects.

7.4.3 Impact of sound rendering condition on presence

The mean scores in each presence category of interest, obtained for each `SOUND CONDITION`, are given in Table 7.3. Following an ANOVA analysis, all scores failed to achieve the 0.05 significance level. Hence, no significant effect was observed for the sound condition over all subjects.

7.4.4 A model for the perceived presence

The probability density function of each of `SPATIAL`, `PERCEPTUAL_REALISM`, and `TEMPLE` scores (see Figure 7.8 for the probability density function of `TEMPLE`) suggest that there are in fact two groups with a normal distribution centered around different means. The data can thus be modeled as a special form of a Gaussian mixture model (GMM). The package **Mclust** [Fraley and Raftery, 2003] allows one to find coefficients of a Gaussian mixture from the data by selecting the optimal model according to the Bayesian information criterion (BIC) applied to

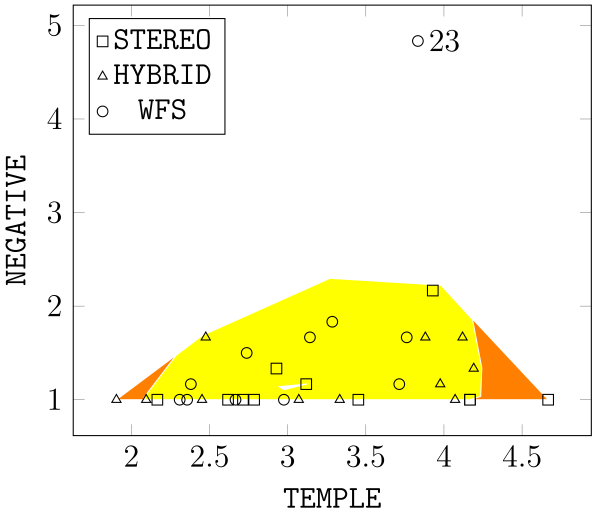


Figure 7.7: Outliers in the NEGATIVE vs. TEMPLE relationship. The bivariate boxplot (or bagplot) generalizes the univariate boxplot. The graph includes two colored regions, called the bag (yellow) and the loop (orange). The interior of the bag includes 50% of the data, and the loop includes all the points outside the bag but inside the fence, thereby outlining potential outliers [Rousseeuw et al., 1999].

Table 7.3: Presence questionnaire mean scores for each category by SOUND CONDITION (and standard deviations in parentheses). Note: the degrees of freedom associated to the ANOVAs are variable.

	STEREO	HYBRID	WFS	F	p-value
SPATIAL	2.90 (1.2)	2.83 (1.12)	2.47 (0.73)	0.50	0.900
PERCEPTUAL_ REALISM	2.71 (1.1)	2.71 (0.87)	2.66 (0.74)	0.01	0.991
TEMPLE	3.34 (0.79)	3.23 (0.87)	2.94 (0.54)	0.79	0.487
SWEDISH	5.15 (0.78)	4.97 (1.34)	4.57 (0.83)	0.89	0.773

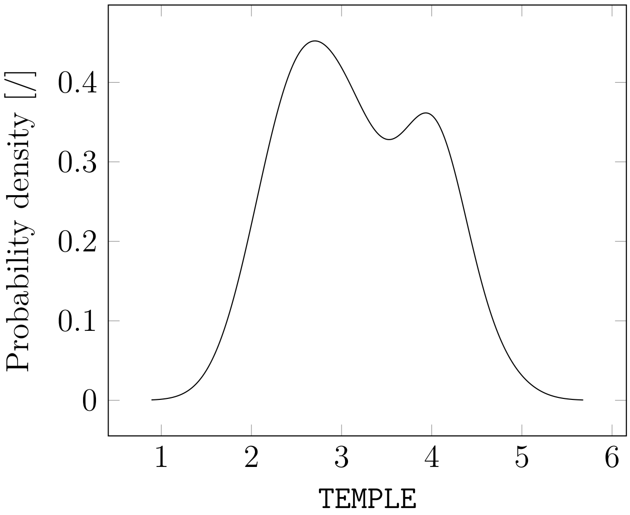


Figure 7.8: The probability density function of TEMPLE scores.

an expectation-maximization (EM) algorithm initialized by hierarchical clustering for parameterized Gaussian mixture models.

The algorithm was run on the data defining the TEMPLE score and the resulting optimal model contains four Gaussian components. The probability that a given participant is not correctly classified using this model ranged from 0 to 5.8×10^{-3} (mean $= 1.2 \times 10^{-3}$). This demonstrates the good quality of the classification. The four groups, referred to by the factor CLASS, are given by the algorithm in descending order of the number of participants they contain: 15, 10, 5, and 2. The mean presence scores for each CLASS category are given in Table 7.4.

Table 7.4: Presence questionnaire mean scores for each category by CLASS (and standard deviations in parentheses). Note: the degrees of freedom associated to the ANOVAs are variable.

	1	2	3	4	F	p-value
SPATIAL	2.25 (0.57)	3.81 (0.78)	1.71 (0.29)	3.64 (0.51)	19.49	$< 10^{-5}$
PERCEPTUAL_ REALISM	2.08 (0.46)	3.66 (0.49)	2.40 (0.91)	3.20 (0.57)	17.07	$< 10^{-5}$
TEMPLE	2.82 (0.39)	4.05 (0.34)	2.28 (0.22)	3.74 (0.03)	39.88	$< 10^{-5}$
SWEDISH	4.78 (0.97)	5.73 (0.73)	4.00 (0.53)	4.00 (0.00)	6.24	0.107

Figure 7.9 shows an analysis of the TEMPLE score depending on the classification CLASS. Groups 1 and 3 tend to have a lower presence score than the groups 2 and 4.

An analysis of variance was carried out on the TEMPLE score with the fixed

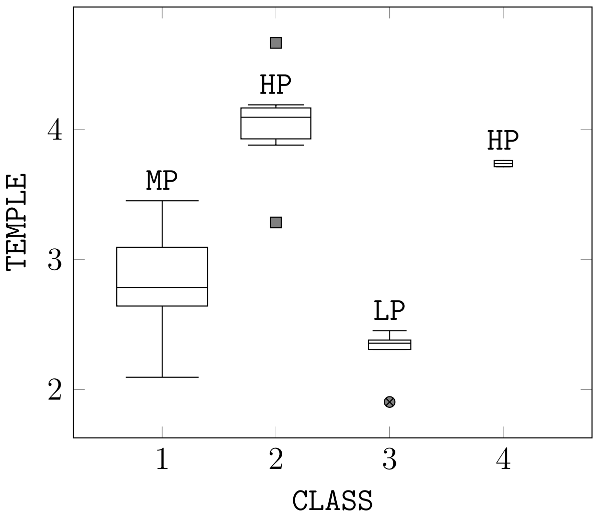


Figure 7.9: Boxplots of the `TEMPLE` score vs. the GMM classification `CLASS`. The widths of the boxplots are proportional to the sample size: 15, 10, 5, and 2 from left to right. The labels `LP`, `MP`, and `HP` are introduced in the text.

factor `CLASS` (four levels). The factor showed a significant effect ($F_{2.72,27.88} = 39.88$, $p < 10^{-5}$). Subsequent post-hoc comparisons (Tukey's HSD test), with an α level of 0.05, showed that groups 2 and 4 do not differ significantly (they form a homogeneous subset), while groups 1 and 3 are significantly different and both differ from the aforementioned set of groups 2 and 4. In the following sections, group 3 will be referred to as `LP` (low presence, 5 subjects), group 1 as `MP` (medium presence, 15 subjects), and the combination of groups 2 and 4 as `HP` (high presence, 12 subjects).

7.4.5 Further analysis in each group

An analysis of variance was carried out on the `TEMPLE` score with the fixed factor `SOUND CONDITION` (three levels) for each presence group defined in the previous section. The factor showed a significant effect on group `HP` ($F_{1.95,8.99} = 6.85$, $p = 0.016$). However, `SOUND CONDITION` failed to reach statistical significance for group `MP` ($F_{2.00,11.90} = 0.11$, $p = 0.896$) and for group `LP` ($F_{1.00,3.00} = 0.69$, $p = 0.468$).

Subsequent post-hoc comparisons (Tukey's HSD test, $\alpha = 0.05$) on the group `HP` showed that conditions `STEREO` (4 participants) and `HYBRID` (5 participants) do not differ significantly (they form a homogeneous subset), while condition `WFS` (3 participants) differs significantly from each of the conditions `STEREO` and `HYBRID`.

Figure 7.10 shows an analysis of the TEMPLE score for each sound condition in group HP. Presence scores are (statistically) lower in the WFS group than in the two other groups. A similar analysis was performed on the SPATIAL, PERCEPTUAL_REALISM, and SWEDISH scores. These failed to achieve the 0.05 significance level. This result, combined with the result on the TEMPLE score, indicates that the impact of sound reproduction is spread across different components of presence rather than confined to the components "Spatial presence" and "Perceptual realism".

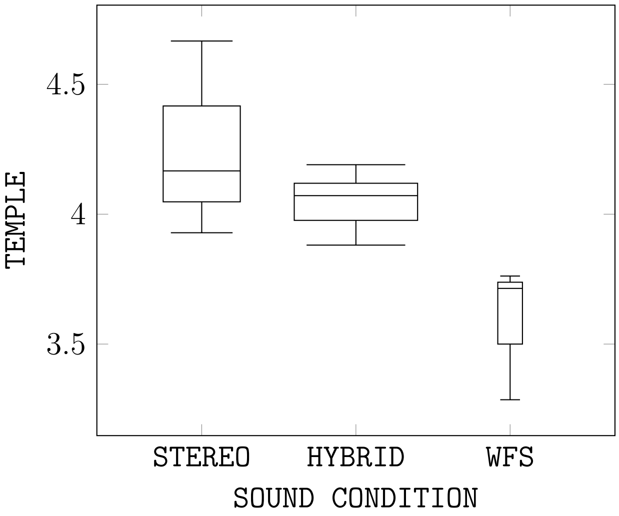


Figure 7.10: Boxplots of TEMPLE score vs. SOUND CONDITION for participants in group HP. The widths of the boxplots are proportional to the sample size: 4, 5, and 3 from left to right.

7.4.6 Discussion

SOUND CONDITION as an independent variable fails to predict the obtained presence score for all participants. Rather, the participants are classified in three groups according to their presence score. The first group has a low presence score (LP), the second has a somewhat higher presence score but also a higher variability (MP), and the third has a high presence score (HP).

SOUND CONDITION has a statistically significant impact for the group HP. In this group, the HYBRID soundtrack is not statistically different from the original STEREO version, which means that the slight difference in content between the two soundtracks did not impact on the reported sense of presence.

When comparing the results for the HYBRID and the WFS soundtracks, one can see that there is a statistical difference in reported sense of presence which is to the advantage of HYBRID. In this condition, sound objects were limited to the space between the virtual speakers, and since the participants were at the sweet-spot, objects between the two virtual sources were fairly well localized in azimuth. Therefore, one could hypothesize that presence is lessened when the auditory objects extend beyond the screen boundaries. Indeed, the virtual loudspeakers in the HYBRID condition were located near the screen borders, and Figure 7.3 shows the spread of the mean ITDs increasing with SOUND CONDITION, from STEREO to WFS. Further studies with different source material would be required to substantiate this hypothesis.

7.5 Results from Heart Rate Variability

The statistical analysis presented in the previous section for the questionnaire answers is repeated here on the recorded heart rate, using the same statistical software. Due to a technical glitch, however, the heart rate could not be recorded for one of the participants, who is thus not included.

7.5.1 Overall comparison of the baseline phase and the experimental phase

Table 7.5 shows the HRV time domain parameters (see Section 7.3.4) averaged over all subjects for the two phases: *baseline*, when the participant is in the dark, and *experiment*, when the participant watches the movie. Since the data does not meet the normality assumption, a non-parametric test, the Wilcoxon signed rank test, was applied between the parameters of the baseline and the experiment. The values of the four parameters are statistically different, at the 0.05 level, between the two phases.

Table 7.5: HRV time domain parameters, averaged over all participants.

HRV	Baseline	Experiment	p-value
MeanRR [ms]	835.8	849.7	0.026
MinRR [ms]	648.9	627.4	0.044
MaxRR [ms]	1024.2	1135.5	0.007
ΔRR [ms]	375.3	508.1	0.006

Table 7.6 shows the changes of HRV frequency domain parameters averaged over all subjects for the two same phases. The Wilcoxon signed rank test was

applied between the parameters of the baseline and the experiment. The last column of the table gives the corresponding p-values. All the HRV frequency domain parameters are statistically different at the 0.05 level.

Table 7.6: HRV frequency domain parameters, averaged over all participants.

HRV	baseline	experiment	p-value
LFnorm [n.u.]	42	55	0.0164
HFnorm [n.u.]	58	45	0.0164
LF/HF [/]	1.08	1.75	0.0335

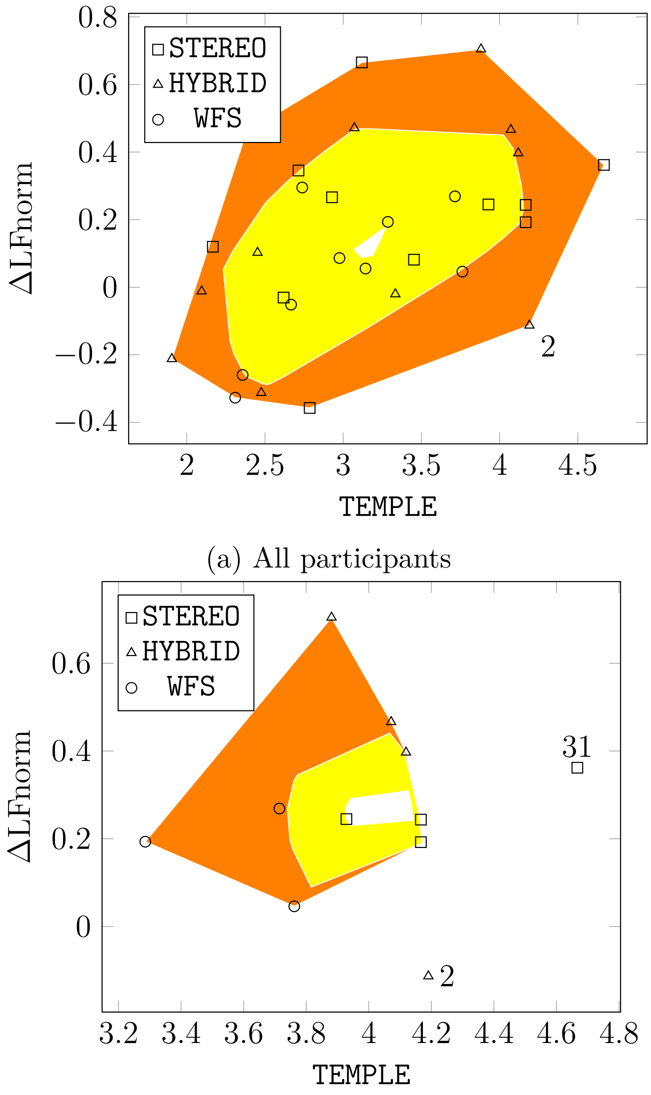
In agreement with the literature [Nickel and Nachreiner, 2003], HRV allows one to discriminate between rest and "work" (the movie presentation). The decreasing HF component is similar to that observed in [Vianna and Tranel, 2006] where different positive and negative emotions are expressed through different movie clips.

7.5.2 Heart Rate Variability

To investigate the effect of `SOUND CONDITION` on HRV, an analysis of variance was carried out on the difference between LFnorm during experiment and baseline (ΔLFnorm) with the fixed factor `SOUND CONDITION` (three levels) for each presence group `LP`, `MP`, and `HP`, defined in Section 7.4.4. The factor showed no significant effect on any group at the 0.05 level. However, the factor showed a significant effect on the group `HP` ($F_{2.00,7.00} = 7.68$, $p = 0.017$) if participant 2 was removed from the analysis. According to the bivariate analysis [Rousseeuw et al., 1999] in Figure 7.11(a), participant 2 would not be classified as an outlier, though he is near the limit. Still, this subject was the only one to exhibit a negative ΔLFnorm (decrease relative to the baseline) in group `HP`. In addition, it can be seen in Figure 7.11(b) that this participant is an outlier in the bivariate analysis of the participants in group `HP`. As such, further results were calculated both with and without subject 2 included.

Subsequent post-hoc comparisons (Tukey's HSD test, $\alpha = 0.05$) on the group `HP` showed that conditions `STEREO` (4 participants) and `WFS` (3 participants) do not differ significantly (they form a homogeneous subset), while condition `HYBRID` (3 participants) differs significantly from this set of conditions.

Figure 7.12 shows the analysis of ΔLFnorm values for each sound condition in group `HP`. ΔLFnorm values are (statistically) higher in the `HYBRID` group than



(a) All participants

(b) Group `HP`

Figure 7.11: Outliers in the ΔLFnorm vs. `TEMPLE` relationship. Bivariate boxplots (a) for all participants, and (b) for group `HP`. See Figure 7.7 for a description of the bivariate boxplot.

in the two other groups. Similar results can be obtained for ΔHFnorm, since, by definition (Equation (7.1)), it is linearly dependent on ΔLFnorm. The results obtained with ΔLF/HF fail to reach the 0.05 significance level just as the results obtained with the time domain parameters do.

In summary, the ideal stereo soundtrack (`HYBRID`) is significantly different from the two other soundtracks in the group of subjects that reported the highest sense of presence. The `HYBRID` soundtrack leads to an increased low frequency component of the HRV.

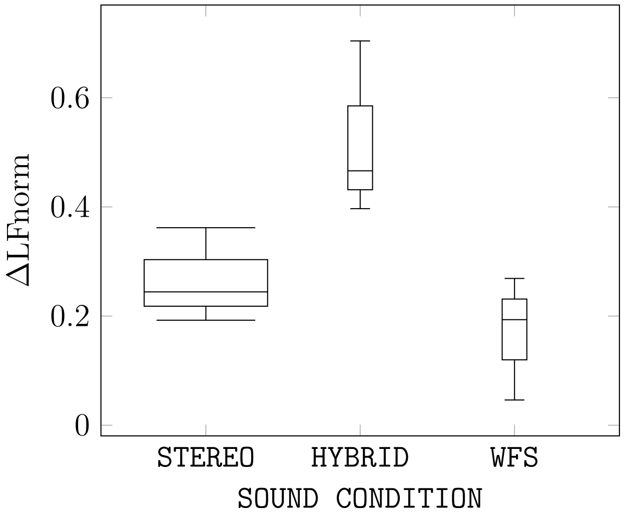


Figure 7.12: Boxplots of the ΔLFnorm value vs. SOUND CONDITION for participants in group HP. The width of the boxplots is proportional to the sample size: 4, 3, and 3 from left to right.

7.5.3 Relationship between HRV and questionnaire scores

In order to evaluate the correlation between the questionnaire scores and the evolution of the frequency domain HRV parameters for all participants, Pearson's product-moment correlation was computed. The results, including the 95% confidence interval, are given in Table 7.7. Naturally, the opposite values are found for ΔHFnorm.

Table 7.7: Pearson's product-moment correlation and 95% confidence interval between ΔLFnorm and the presence scores (in the first imputed dataset). In parentheses, the values obtained when participant 2 is discarded.

	Sample estimate	t_{29} (t_{28})	p-value	Lower bound	Upper bound
SPATIAL	0.41 (0.49)	2.41 (2.95)	0.022 (0.006)	0.06 (0.15)	0.67 (0.72)
PERCEPTUAL_ REALISM	0.42 (0.44)	2.47 (2.56)	0.020 (0.016)	0.07 (0.09)	0.67 (0.69)
TEMPLE	0.45 (0.52)	2.73 (3.23)	0.011 (0.003)	0.12 (0.20)	0.70 (0.74)
SWEDISH	0.52 (0.56)	3.24 (3.54)	0.003 (0.001)	0.20 (0.24)	0.74 (0.76)

The correlation is significantly different from 0 (at the 0.05 level) for every presence score of interest. The highest value is obtained with the SWEDISH score, which pertains only to the sound rendering. When participant 2 is discarded from the analysis, the values are improved. This is indicated in parentheses in Table 7.7.

7.5.4 Discussion

The presentation of the movie to the participants had an impact on several Heart Rate Variability (HRV) statistics in both the time domain and the frequency domain. For all participants, a relation is found between the reported presence score TEMPLE and the evolutions of both LFnorm and HFnorm between the baseline and the experiment.

SOUND CONDITION as an independent variable fails to predict the obtained evolutions of HRV parameters for all participants. The analysis according to each presence group shows that SOUND CONDITION has a statistically significant impact on ΔLFnorm and ΔHFnorm for the group HP (with participant 2 discarded). In that case, the HYBRID soundtrack is statistically different from both the original STEREO soundtrack and the WFS soundtrack.

When comparing the HYBRID and the WFS soundtracks, one can see that there is a statistical difference in the evolutions of LFnorm and HFnorm, which is higher in the HYBRID case. Participants in the HYBRID condition therefore experienced a higher increase in LFnorm than the others. Since ΔLFnorm correlates positively with TEMPLE for all participants, this supports our previous findings (Section 7.4.6) that the participants experienced a stronger sense of presence with the HYBRID soundtrack than with the WFS soundtrack.

7.6 Results from the participants' feedback

Among the comments the participants made about the experiment, a few recurring ones can be highlighted. Nine participants indicated that they were disappointed by the (visual) 3D. Maybe they expected to see more depth in the movie than they actually saw. As can be seen in Figure 6.7, the range of depth of the sources is rather narrow (roughly from 0.5 m to 5 m). The length of the movie was also a problem for seven participants who reported that it was too short. They needed more time to forget they were in an experiment. Five participants found the end of the movie excerpt too abrupt; they would have appreciated to know more about the story. Regarding the setup, four participants were distracted by the visibility of the corner of the panels in the SMART-I^2 and three complained about the passive polarized glasses (two of which wore prescription glasses).

It is therefore possible that the results found in this study could vary, or be more representative, if a longer film was shown, and if the projection was made on a traditional flat format screen. These comments will be taken into consideration in future studies.

7.7 Conclusions

Different sound spatialization techniques were combined with an s-3D movie. The impact of these techniques on the sense of presence was investigated using a post-session questionnaire and heart rate recordings.

The sound condition did not affect the reported presence score directly for all subjects. Rather, participants could be classified according to their presence score independently of the sound condition. In the group that reported the highest sense of presence, for which sound rendering condition was influential, the spatially coherent soundtrack (`WFS`) was significantly different from the two other stereo soundtracks (`STEREO` and `HYBRID`). The `WFS` soundtrack led to a decreased reported sense of presence. Analysis of the participants' Heart Rate Variability (HRV) revealed that, in the group that reported the highest sense of presence, the ideal stereo version (`HYBRID`) was significantly different from the two other soundtracks. The `HYBRID` soundtrack led to an increased low frequency (LF) component of the HRV.

The increase in the HRV LF component between the baseline and the movie presentation was also shown to be positively correlated with the overall presence score for all participants. Both the subjective (questionnaire) and objective (HRV) measures showed that the `HYBRID` soundtrack led to a higher sense of presence than the `WFS` soundtrack for participants that reported the highest sense of presence.

The comments made by the participants underline the limitations of this experiment. Most were related to the content, rather than the setup. Some participants found that the movie did not present much depth, and that the movie was too short to allow some of them to forget they were taking part in an experiment. Several participants were disappointed with the end of the story, or even did not like the movie at all.

The results found here constitute a basis for future research. The impact of an off-axis seating position needs further investigation, since the s-3D image is egocentric. This is the topic of Chapter 8. Apart from the reverberation, all the sound in this experiment came from the front. Therefore, there is also a need to investigate the effect of 360° sound reproduction. Finally, one could investigate other types of 3D sound rendering, such as Ambisonics, binaural, or possible hybrid combinations of multiple systems.

Bibliography

André, C. R., Rébillat, M., Embrechts, J.-J., Verly, J. G., Katz, B. F. G., 2012. Sound for 3D cinema and the sense of presence. In: Proc. of the 18th Int. Conf. on Auditory Display (ICAD 2012). Atlanta, GA, pp. 14–21.
http://hdl.handle.net/2268/127803 153

Blauert, J., 1997. Spatial hearing: the psychophysics of human sound localization. MIT Press. 156

Bouvier, P., Dec. 2009. La présence en réalité virtuelle, une approche centrée utilisateur. Ph.D. thesis, Université Paris-Est, Paris, France. 161

de Bruijn, W. P. J., Boone, M. M., 2002. Subjective experiments on the effects of combining spatialized audio and 2D video projection in audio-visual systems. In: Audio Eng. Soc. Conv. 112.
http://www.aes.org/e-lib/browse.cfm?elib=11350 155

Faller, C., Merimaa, J., Nov. 2004. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. J. Acoust. Soc. Am. 116 (5), 3075–3089.
http://dx.doi.org/10.1121/1.1791872 156

Fraley, C., Raftery, A. E., Sep. 2003. Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. J. Classif. 20, 263–286.
http://dx.doi.org/10.1007/s00357-003-0015-3 164

Huang, S.-F., Tsai, P.-Y., Sung, W.-H., Lin, C.-Y., Chuang, T.-Y., Oct. 2008. The comparisons of heart rate variability and perceived exertion during simulated cycling with various viewing devices. Presence-Teleop. Virt. 17 (6), 575–583.
http://dx.doi.org/10.1162/pres.17.6.575 161

Ijsselsteijn, W., de Ridder, H., Freeman, J., Avons, S. E., Bouwhuis, D., Jun. 2001. Effects of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence. Presence-Teleop. Virt. 10 (3),

298–311.
http://dx.doi.org/10.1162/105474601300343621 152

Imai, K., King, G., Lau, O., 2008. Toward a common framework for statistical analysis and development. Journal of Computational and Graphical Statistics 17 (4), 892–913.
http://dx.doi.org/10.1198/106186008X384898 163

King, G., Honaker, J., Joseph, A., Scheve, K., 2001. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. Am. Polit. Sci. Rev. 95, 49–69. 163

Kreibig, S. D., Jul. 2010. Autonomic nervous system activity in emotion: A review. Biol. Psychol. 84 (3), 394–421.
http://dx.doi.org/10.1016/j.biopsycho.2010.03.010 163

Larsson, P., Västfjäll, D., Olsson, P., Kleiner, M., Oct. 2007. When what you hear is what you see: Presence and auditory-visual integration in virtual environments. In: Proc. 10th Annu. Int. Workshop Presence. Barcelona, Spain, pp. 11–18. 160

Lombard, M., Ditton, T., Weinstein, L., Nov. 2009. Measuring (tele)presence: The Temple Presence Inventory. In: Proc. 12th Annu. Int. Workshop Presence. Los Angeles, CA, pp. 1–15. 152, 160, 161, 261

Nickel, P., Nachreiner, F., 2003. Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload. Hum. Factors 45 (4), 575 –590.
http://dx.doi.org/10.1518/hfes.45.4.575.27094 170

Raghunathan, T., Dong, Q., 2011. Analysis of variance from multiply imputed data sets. Tech. rep., Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI. 163

Rousseeuw, P. J., Ruts, I., Tukey, J. W., Nov. 1999. The bagplot: A bivariate boxplot. The American Statistician 53 (4), 382–387.
http://dx.doi.org/10.1080/00031305.1999.10474494 164, 165, 170

Selvaraj, N., Jaryal, A., Santhosh, J., Deepak, K. K., Anand, S., Jan. 2008. Assessment of heart rate variability derived from finger-tip photoplethysmography

as compared to electrocardiography. J. Med. Eng. Technol. 32 (6), 479–484.
http://dx.doi.org/10.1080/03091900701781317 161

Slater, M., Guger, C., Edlinger, G., Leeb, R., Pfurtscheller, G., Antley, A., Garau, M., Brogni, A., Friedman, D., Oct. 2006. Analysis of physiological responses to a social situation in an immersive virtual environment. Presence-Teleop. Virt. 15 (5), 553–569.
http://dx.doi.org/10.1162/pres.15.5.553 161

Stein, P., Kleiger, R., Feb. 1999. Insights from the study of heart rate variability. Annu. Rev. Med. 50 (1), 249–261.
http://dx.doi.org/10.1146/annurev.med.50.1.249 163

Task Force of The European Soc. of Cardiology and The North Am. Soc. of Pacing and Electrophysiology, Mar. 1996. Heart Rate Variability: Standards of measurement, physiological interpretation, and clinical use. Circulation 93 (5), 1043–1065.
http://dx.doi.org/10.1161/01.CIR.93.5.1043 161, 162

Theile, G., 1990. On the performance of two-channel and multi-channel stereophony. In: Audio Eng. Soc. Conv. 88.
http://www.aes.org/e-lib/browse.cfm?elib=5807 155

Theile, G., Wittek, H., Reisinger, M., 2003. Potential Wavefield Synthesis applications in the multichannel stereophonic world. In: Audio Eng. Soc. 24th Int. Conf.: Multichannel Audio, The New Reality.
http://www.aes.org/e-lib/browse.cfm?elib=12280 155

Thurlow, W. R., Jack, C. E., Jun. 1973. Certain determinants of the “ventriloquism effect”. Percept. Motor Skill 36, 1171–1184.
http://dx.doi.org/10.2466/pms.1973.36.3c.1171 155

Vianna, E., Tranel, D., Jul. 2006. Gastric myoelectrical activity as an index of emotional arousal. Int. J. Psychophysiol. 61 (1), 70–76.
http://dx.doi.org/10.1016/j.ijpsycho.2005.10.019 162, 170

This page intentionally left blank.

Chapter
8

Subjective Evaluation of the Audiovisual Spatial Congruence in the Case of Stereoscopic-3D Video and Wave Field Synthesis

Highlights

- ✓ The congruence between Wave Field Synthesis and s-3D video is evaluated.
- ✓ We consider the angular error when the spectators are seated at different locations.
- ✓ The ambient noise level was varied as an independent variable with two levels.
- ✓ The point of subjective equivalence, where participants were equally likely to detect the angular error, was 18.3° (SNR = 19 dB) and 19.4° (SNR = 4 dB).
- ✓ Precise spatial sound reproduction is theoretically possible in a movie theater with a realistic installation complexity.

Contents

8.1 Introduction

The present chapter addresses the question of the perceptual congruence between the sound and the image when the spectator in a cinema is presented with a 3D sound scene spatially coherent with the stereoscopic 3D (s-3D) scene. The material in this chapter has been published in [André et al., 2013] and [André et al., 2014].

In essence, the depth perception in s-3D is created by presenting a different image to the two eyes. Both images in an s-3D pair are displayed on the cinema screen and all spectators thus look at the same pair of images. When one compares the visual perception of two spectators seated at different locations in the room, one finds, both geometrically and experimentally, that the objects of the scene displayed on the screen are rendered at different locations in the room (Section 5.1.2).

The present study considers the potential error in the angle between the sound and the image when presenting precise spatial sound through Wave Field Synthesis (WFS, see Section 3.3.4) in combination with s-3D video to spectators seated at different locations. The spectators evaluate the spatial coherence between a displayed virtual character and a reproduced speech sound. The psychometric function, which relates the physical stimulus to the participants' responses, is obtained in presence or absence of additional background noise.

8.1.1 Subjective evaluation of the audiovisual congruence

Following the discussion on the auditory-visual spatial ventriloquism in azimuth, given in Section 2.3.2, the audiovisual apparatus of several recent experiments conducted on the association of image and sound are compared in Table 8.1. The studies concerned with 3D sound all used WFS for the sound reproduction.

None of these previous studies addressed the problem of the s-3D video projection of a natural scene to multiple users. In addition, the effect of the auditory ambience noise level is investigated here. Because WFS reproduces a sound position independent from the listener position [Theile et al., 2003], the limit of audiovisual integration found here is also a measure of the sweet-spot for accurate auditory-visual reproduction mentionned in Section 7.2.2.

8.1.2 Audiovisual spatial coherence in s-3D video

As previously stated, the illusion of depth perception in s-3D cinema is created by presenting a different image to each eye, and the perceived location of displayed objects varies with the seating location of the spectator (Section 5.1.2). In fact, only visual objects with a zero parallax, such that they are perceived as located at the depth of the screen plane, are consistently perceived among spectators. All other positions are not consistently perceived within the room. However, any spectator's line of sight crosses the screen at the position of the s-3D stimulus on the screen (position I in Figure 5.3 and Figure 8.1).

In combination with 3D sound, this property of s-3D images can lead to an audiovisual error for spectators seated off-axis (Section 5.1.2).

8.1.3 Improving the spatial coherence

We introduce a method to reduce the angular error mentioned in Section 5.1.2. This method already exists in the 2D case [de Bruijn and Boone, 2003]. Here, we

Table 8.1: Comparison between audiovisual apparatus used in published papers and in the present chapter.

Study	Visual system	Image	Size (in.)	Sound	Content
Komiyama [1989]	HDTV (16:9)	2D	72	mono	Real (person)
de Bruijn and Boone [2003]	Projector (4:3)	2D	100	WFS	Real (person)
Melchior et al. [2003]	Projector (16:9)	2D	163	WFS	Real (person)
Melchior et al. [2006]	HMD[a] (N/A[b])	s-3D	NA[b]	WFS	CGI[c] (abstract)
Present chapter	Projector (4:3)	s-3D	129	WFS	CGI (character)

[a] HMD stands for head-mounted display.

[b] the diagonal field of view of their HMD is 31.5°.

[c] CGI stands for computer-generated images.

extend it to s-3D. This constitutes one contribution of the present chapter.

First, we describe the method which was developed to combine spatially accurate sound rendering, by means of WFS with regular 2D video to build a teleconferencing system [de Bruijn and Boone, 2003]. The researchers faced a problem related to linear perspective (Section 4.2.4). A user of their system, not sitting at the viewpoint, would experience a discrepancy between the sound of the voice and the image of the face of his or her interlocutor. Indeed, as Goldstein [1987] has shown, when a picture is viewed at an angle from the ideal viewpoint, judgments of the spatial layout are relatively constant over a large range of angles, but not judgments on the orientation of lines in space. In the latter case, the result varies systematically with the viewing angle. Directions that point to the sides of the pictures remain constant up to about |20°| away from the viewpoint. On the contrary, directions that point outside the pictures seem to "follow" the viewer. A famous example of this is Uncle Sam's finger in the "Uncle Sam wants you" poster. This is also the case for the gaze in a portrait, such as that of Mona Lisa's in La Joconde.

This paradox makes it difficult to compute the location of the viewer's visual percept when watching a 2D picture from an off-axis location. The researchers placed the sound sources at the exact positions specified by the true 3D layout. Participants then graded the perceived discrepancy between the sound and the image according to the ITU 5-point impairment scale defined as follows: (1) imperceptible, (2) perceptible, but not annoying, (3) slightly annoying, (4) annoying, and (5) very annoying. The experiment revealed that annoying effects did occur when viewers shifted away laterally from the ideal viewpoint. A shift in depth seemed less concerning.

As de Bruijn and Boone [2003] further suggested, it is possible to limit the angular discrepancy between the sound and the image by pulling the audio sources towards the screen along the line between the visual object and the ideal viewpoint. At the same time, the audio gain is adjusted to produce the same sound level as the original sound source at the ideal viewpoint.

Second, the proposed method is adapted to s-3D video. We consider again in Figure 8.1 the geometry of Figure 5.3. Given the positions of the visual object V_1, the ideal viewpoint S_1, and the screen, one can compute the positions of the two points in the left and right images on the screen corresponding to the visual object, according to the geometrical model described in Section 6.3. The sound can be placed at a point A' anywhere along the line defined by S_1 and V_1, say according to a real parameter ρ defined by

$$A' - V_1 = \rho(I - V_1). \tag{8.1}$$

where I is the intersection of the line S_1V_1 and the screen. Therefore, $\rho = 0$ yields $A' = V_1$ and $\rho = 1$ yields $A' = I$.

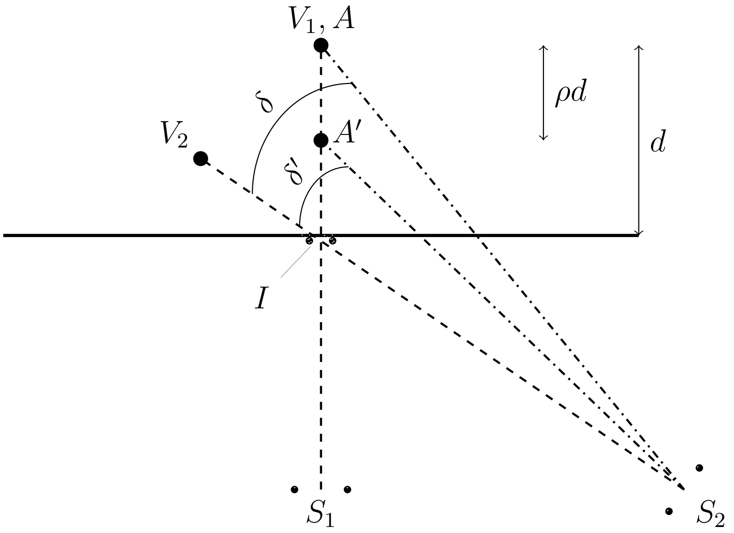


Figure 8.1: Illustration of the method of reduction of the angular error between sound and image as a function of seating position. The spectator at S_2 watches the same point-like s-3D object as the spectator at S_1, the ideal viewpoint. The compression of the audio depth (A' instead of A) allows one to reduce the angular error between the sound and the image ($\delta' < \delta$). The distance d is $|I - V_1|$.

For a spectator seated at S_2, the visual object appears at V_2, resulting in an angular error δ between the sound and the image if the sound is positioned at $A = V_1$. When the sound is pulled closer to the screen, say at A', the angular error decreases for the spectator at S_2, i.e. $\delta' < \delta$. Note that this remains true when the line S_1V_1 is not perpendicular to the screen, as will be the case in this experiment. Provided that the sound level at A' is adjusted to match the volume it would have produced from A, the audiovisual congruence should be maintained at S_1. It should however be noted that a single adjustment will not be correct for all seating positions, as the acoustic attenuation is a function of the distance squared, such that the error in level adjustment will be greater for seating positions closer to the screen than S_1. These positions are not considered in the present study.

8.1.4 Objectives

In the present study, an experiment is conducted with naive spectators to evaluate the threshold of bimodal integration associated with the angular error between audio and video in the case of Wave Field Synthesis (WFS) and stereoscopic 3D (s-3D) video. A virtual scene consisting of a character in an apartment is chosen to simulate a cinema context. The impact of the presence of additional ambient noise is also investigated. The value of the thresholds with and without ambient noise is obtained through the measurement of the associated psychometric functions [Klein, 2001]. According to the maximum-likelihood theory of sensory integration [Ernst and Banks, 2002], it is expected that the psychometric function will have a slower decay associated with a higher threshold in presence of ambient noise. Audiovisual rendering is provided via the SMART-I^2 platform (Section 5.2) using passive s-3D video and WFS audio. This virtual reality system provides its users with stable auditory and visual cues in a large rendering area. The psychometric functions are obtained in a yes/no experiment with the method of constant stimuli with and without ambient noise. The first objective is to study to what extent naive subjects perceive the inconsistency between the sound and the image when viewing s-3D contents. The second objective is to verify that ambient noise allows the subjects to maintain the auditory-visual stimulus integration at higher angles of error. The third objective is to verify that the compression of the audio space towards the screen reduces the perception of the inconsistency between the sound and the image when viewing s-3D contents combined with spatially accurate sound.

8.2 Method

8.2.1 Experimental design

In each session, three participants faced the right panel of the SMART-I^2, used as the screen, and were seated at 2 m from it (Figure 8.2). The first participant was seated at S_1, facing the middle of the panel. The other two were seated at S_2 and S_3, at 0.6 m and 1.2 m to the right of S_1, respectively. One virtual character, the visual stimulus, was rendered 1.5 m behind the screen, at 0.8 m to the left of S_1. A speech signal, the auditory stimulus, was rendered at five different positions along the line joining S_1 and the virtual character position, V_1. These positions are labelled $A_{①}$ (closest to the screen) to $A_{⑤}$ (farthest from the screen). $A_{③}$ corresponds to the position of the virtual character, i.e. there is no audiovisual discrepancy for this sound position if the spectator is at S_1. In addition, a control position A_c is defined as the mirror image of $A_{③}$ with respect to the perpendicular to the screen passing through S_1. The different subscripts used to denote the audio and visual object positions underline that these are independent. An ambience sound signal was reproduced at the positions Bg, behind the speech source position.

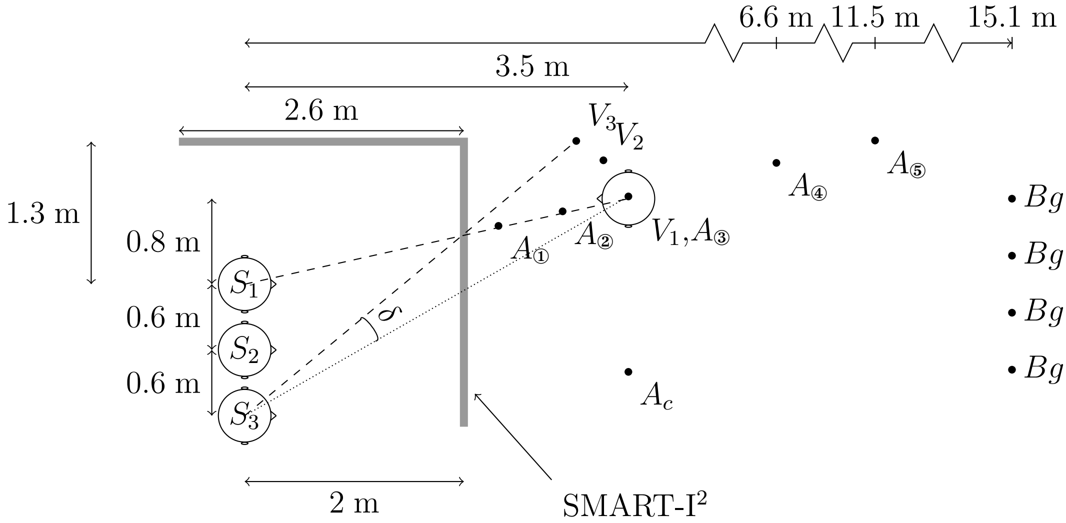


Figure 8.2: Layout of the experimental setup with respect to the SMART-I^2 panels (thick gray segments). The S_i's are the positions of the subjects. V_i is the perceived position of the virtual character seen from S_i. The $A_{①}$'s are the audio positions of the rendered speech, and the Bg's are the audio positions of the rendered background ambience sound. The angle δ illustrates the angular separation between the perceived location of the character and a position of the rendered speech.

A sample of 17 subjects took part in the experiment (14 men, 3 women, age 19 to 30 years old, mean = 23.5, stdev = 3.2). They all worked at the LIMSI. They were naive as to the experiment and they were not financially compensated. All but one participant had already seen at least one s-3D movie in a cinema. Twelve participants played 3D video games (but not necessarily in s-3D) at most once a month. Only five participants used spatialized audio systems more than once a month, and three of them were the only ones to use virtual reality systems. The subjects can therefore be considered as being naive with respect to the combination of audio and video technologies used here.

The chosen experimental design was a within-subjects design with three factors: the sound position (six levels), the presence of background noise (two levels), and the repetition (four levels) (see Section 8.2.4). Together, the seat position and the sound position define the angular error `AVangle` (14 levels) between the sound and the image, in degrees [deg]. The presence of background noise `background` is coded as a binary variable, with the values `BG` and `NOBG` indicating presence and absence, respectively.

In the case of a yes/no experiment, Lam et al. [1999] shows that four points are enough to accurately estimate the psychometric function while keeping a low standard deviation on the parameters. These points are those where the positive answer rate is expected to be 12, 31, 69, and 84%. Thirty to fifty trials at each point yield accurate estimates of the threshold and slope parameters. In a pilot experiment for this study, the stimulus values corresponding to these optimal sampling points were first estimated using the curve A in Figure 4 of [Komiyama, 1989]. The four values that optimally sample this curve are 1.5, 6.6, 12.7, and 16.9°. The pilot experiment with six subjects showed that slightly larger values were needed to ensure that every subject could perceive an audiovisual discrepancy. The chosen values of ρ and their corresponding angles of error are given in Table 8.2. These values were chosen as a compromise between being close to the optimal sampling values and ensuring that the SMART-I^2 was able to reproduce exactly the sound source at the chosen location. The values of the angular error corresponding to each value of ρ and each position S_i can be obtained from the geometry in Figure 8.2. Given the coordinates S_1 and V_1 and the coordinates of the eyes of the viewer at S_1, the projections I_l and I_r of V_1 in the left and right images can be obtained. Then, the coordinates of V_2 and V_3 can be computed. It is assumed that each viewer is facing the direction of the midpoint between I_l and I_r.

Table 8.2: Chosen values of ρ and their corresponding angles of error `AVangle` [deg] for each position S_i in the layout of Figure 8.2.

	$A_{①}$	$A_{②}$	$A_{③}$	$A_{④}$	$A_{⑤}$	A_c
ρ	0.79	0.40	0.01	-2.07	-5.32	control
S_1	0.0	0.0	0.0	0.0	0.0	26
S_2	1.9	4.3	6.0	10.2	12.2	31
S_3	2.9	6.9	9.9	17.4	21.2	34

8.2.2 Experimental setup

A block-diagram of the software and hardware architecture used in this chapter is presented in Figure 8.3. Mostly, we used the originally available architecture described in Figure 5.6. One additional laptop is used in this experiment to collect the participants' answers, given via WiiMotes connected to the laptop via Bluetooth, and to pilot the audio and video engine via OSC.

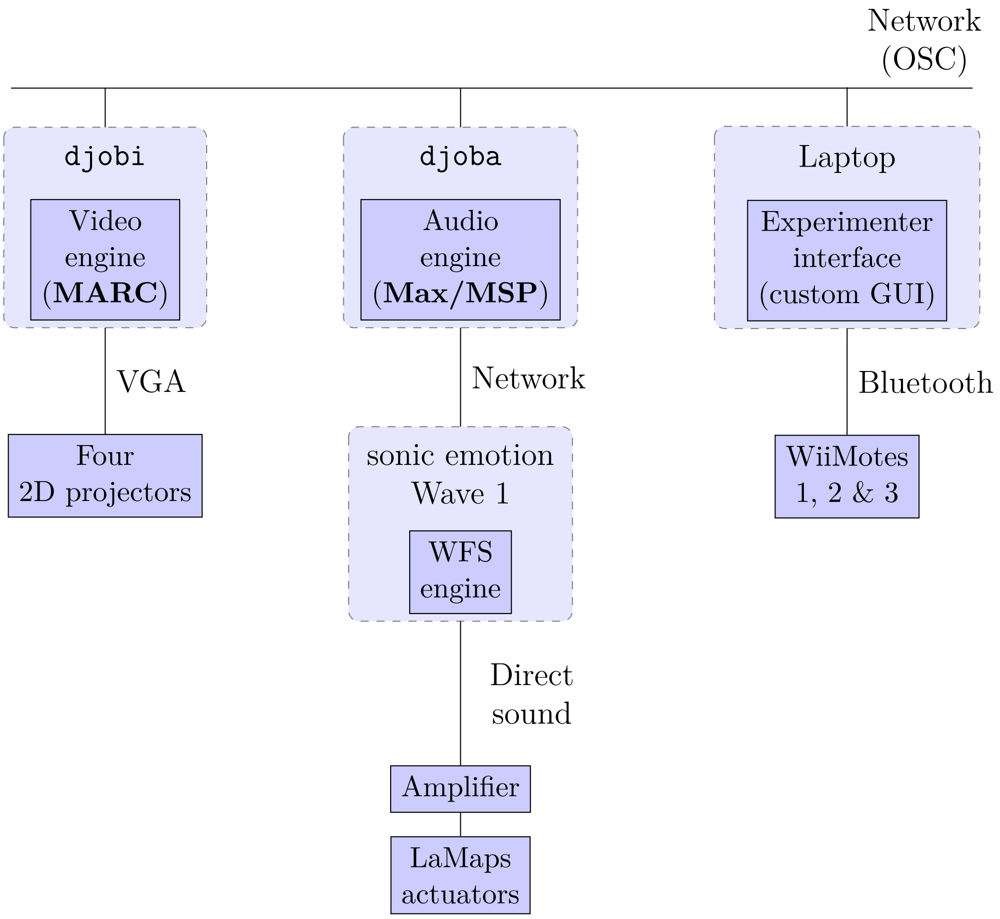


Figure 8.3: A block-diagram of the software and hardware used in the experiment described in the present chapter.

The software used to render the visual part of the experiment is **MARC** (Multimodal Affective and Reactive Characters), a framework for real-time affective interaction with multiple characters [Courgeon and Clavel, 2013]. **MARC** features three main modules: facial expressions edition, body gesture edition, and real-time interactive rendering. **MARC** relies on GPU programming (OpenGL/GLSL) to render in real-time detailed models and realistic skin lighting (shadow casting, simulation of light diffusion through skin). The integration of **MARC** in the SMART-I^2 is described in [Courgeon et al., 2010].

8.2.3 Audiovisual material

The visual material consisted of one **MARC** character (Simon) in a scene depicting an apartment (Figure 8.4). The point of view was chosen so that the character's mouth was at the height of the SMART-I^2's loudspeakers, to avoid any vertical discrepancy. The scene was rendered at a 1:1 scale, i.e. life-size.

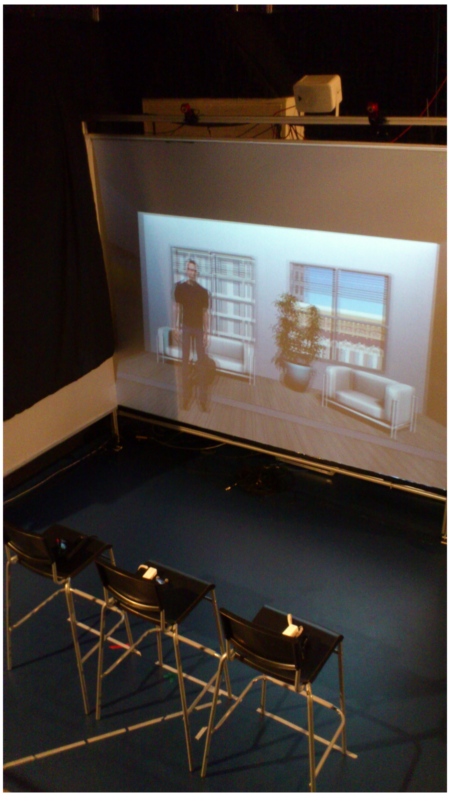


Figure 8.4: Photo of the experimental setup showing the three bar stools and a projected s-3D image.

The audio material contained two signals. The first signal was the speech pronounced by the virtual character. There were two different five-second long sentences from two tales selected from a corpus [Doukhan et al., 2011]. The level of the stimuli was fixed at 52 dB(A) RMS at S_1.

The second signal was that of the background ambience. This signal was made up of several uncorrelated recordings made on a street of New York City. The positions Bg of the virtual sound sources and the uncorrelation of the signals prevented the subjects from localizing the position of the ambience. The level of the ambience was fixed at 48 dB(A) RMS at S_1. The speech and background ambience were loud enough relative to the ambient noise in the room (33 dB(A) RMS). The SNR was therefore either 19 dB(A) (for `NOBG`) or 4 dB(A) (for `BG`).

The visual content was played continuously throughout the trial sessions, and the background ambience level was adjusted as a stimulus value.

8.2.4 Experimental task

In order to make efficient use of the installation and minimize total experimental time, up to three participants took part in each experimental session. Each participant sat successively at the three positions S_1, S_2, and S_3 (not necessarily in this order). The participants were first provided with written instructions regarding the experiment. They wore passive linear polarizing s-3D glasses and received a WiiMote controller. There was no restriction on their head movement.

Each experimental session consisted in three consecutive blocks to allow for each participant to sit at the three different positions. Each block consisted in 48 trials for data collection, corresponding to six sound positions, two values of background level, and four repetitions of each combination of sound position and background level. The first block started with a training session to make sure that the participants understood the task. This training alternated between the correct audiovisual combination ($A_{③}$) and the control position for the sound (A_c). The order of the stimuli was randomized in each block. Each value of the repetition factor was associated with one of two different speech sentences, alternating between the two. This was done to avoid monotony during the experiment. Each trial started with a five-second stimulus followed by a five-second period during which subjects answered the question “Is the voice coherent with the character position?” by pressing a button of the WiiMote. The number of repetitions was chosen to keep the experiment short (about 15 minutes per block, and 60 minutes in total) and the number of subjects needed low. The stimuli in each block were played in an automated way, with the subjects being observed remotely. A five-minute rest was granted between two successive blocks.

8.2.5 Modelling of the psychometric function

When dealing with psychometric data, it is customary to use the following expression to relate the value of the stimulus x to the value of the psychometric function $\psi(x)$ [Wichmann and Hill, 2001a]

$$\psi(x) = \gamma + (1 - \gamma - \lambda)F(x; \alpha, \beta). \tag{8.2}$$

In the case of a yes/no paradigm such as in the method used here, the response ψ is the proportion of answers "yes", and the parameters α and β determines the shape of the sigmoid curve F, which takes values in $[0, 1]$. γ is the guessing rate, which is a free parameter in a yes/no task, and λ is the lapse rate, which is the fraction of recordings where subjects respond independently of the stimulus level. The values of γ and λ are of secondary interest because they characterize the stimulus-independent behavior of the subjects.

Several expressions can be used for the sigmoid F. The logistic function is the default in `psignifit`, a maximum-likelihood estimation software tool [Wichmann and Hill, 2001a]. This function is defined as

$$F(x; \alpha, \beta) = \frac{1}{1 + e^{-\frac{x-\alpha}{\beta}}}. \tag{8.3}$$

It is assumed that the sigmoid curve F accounts for the psychological process of interest. However, the performance of the subjects is better understood in terms of the threshold at a certain performance level and the slope of the curve at the same point. With a yes/no paradigm, the threshold is taken as the point of subjective equivalence (PSE), which is the stimulus value x_t for which $F(x_t) = 0.5$ [Treutwein, 1995]. The threshold is therefore the stimulus value at which the subjects answer "yes" half of the time. The slope determines how strongly this judgment varies with the stimulus value. A large slope (in absolute value) means that the threshold separates clearly the stimuli into two categories. In this work, the stimulus value is the angular error between the sound and the image. As such, the design of the experiment results in a curve with a negative slope (see Figures 8.5 and 8.6 for some examples of the sigmoid curve). Therefore, subjects are more likely to answer "yes" for stimulus values below the threshold, and "no" for values above it. Note that we will always report the slope at the PSE.

8.3 Results

This section presents the results of the statistical analysis carried out on the answers from the participants. The answers "yes" and "no" are coded as 1 and 0, respectively. The mean score averaged over all participants is called `congruence`.

8.3.1 Panel performance and outliers detection

Our goal here is to analyze the participants' responses and to evaluate their performance. As a starting point, the psychometric function was fitted using `psignifit` version 2.5.6*, a software tool that implements the maximum-likelihood method described by Wichmann and Hill [2001a]. Since this tool allows fitting with respect to only one variable, the data was split according to each value of `background` (`NOBG` and `BG`).

By default, the parameters γ and λ (see Equation (8.2)) were each constrained to be in the interval $[0, 0.05]$. But the algorithm gave results on the upper extremity of the interval. So this interval was increased to $[0, 0.2]$ in order to obtain correct estimated values. The results are shown in Figure 8.5.

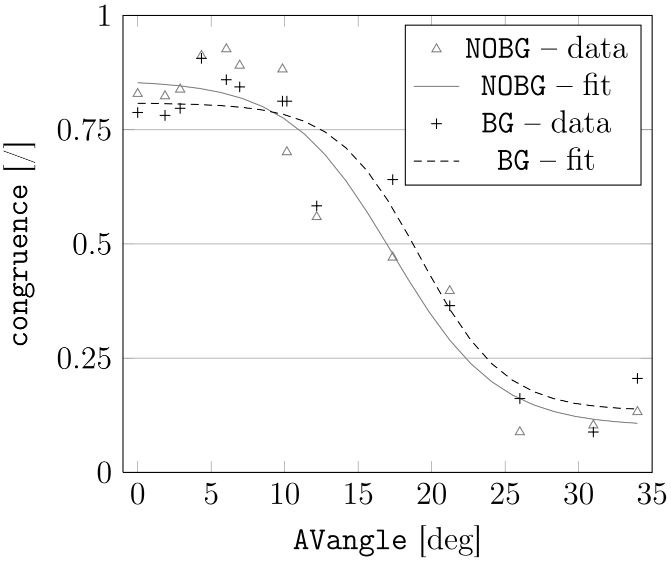


Figure 8.5: Mean responses over all participants for each condition and associated psychometric functions.

An unexpectedly high value was obtained for the largest value of `AVangle` (i.e. 34 deg). This value correspond to the control case at position S_3. The proximity of both the source and the listener to the extremity of the panel is judged to have played a role in this result. The data corresponding to this configuration is therefore discarded from the rest of the analysis, to avoid an underestimation of the slopes.

With the highest value of `AVangle` discarded, the performance of each participant was evaluated. The data for each participant was analyzed separately,

*http://bootstrap-software.org/psignifit/, last accessed 17/12/2013.

Table 8.3: Normal parameter estimates of the distribution of β over participants for each level of `background`.

Condition	Mean	Median	Std dev.
`NOBG`	1.53	-0.89	11.61
`NOBG` (w/o 9)	-1.36	-1.02	1.24
`BG`	-52.40	-2.58	168.61
`BG` (w/o 4 & 9)	-2.63	-1.06	3.45

and the corresponding fits were obtained using the same constraints on the curve parameters. In Table 8.3, the normal parameter estimates of the distribution of β are given. This parameter is related to the slope of the psychometric curve at the PSE, i.e. `AVangle` $= \alpha$. The large standard deviation, as well as the difference between the means and the medians, suggested the presence of outliers.

A criterion based on the Median Absolute Deviation (MAD), less sensitive to the presence of outliers than the standard deviation, was used as an alternative to the more traditional standard deviation criterion [Leys et al., 2013]. The MAD is defined as

$$\mathrm{MAD} = b \operatorname*{median}_i \left\{ \left| \beta_i - \tilde{\beta} \right| \right\}, \tag{8.4}$$

where $\tilde{\beta}$ is the median of the dataset, and $b = 1.4826$ is a coefficient linked to the assumed normality of the data. The threshold for rejecting a measurement was set at 3 (a very conservative value). A non-outlier value should therefore lie in the range

$$\tilde{\beta} - 3\mathrm{MAD} < \beta_i < \tilde{\beta} + 3\mathrm{MAD}. \tag{8.5}$$

In condition `NOBG` (MAD = 1.26), participant 9 was clearly an outlier with a β value of 44.83. In condition `BG` (MAD = 3.34), participant 9 was again an outlier ($\beta = -162.06$), and this was also the case for participant 4 ($\beta = -689.21$). In both cases, the associated threshold was above the considered range of `AVangle`. The data from participant 9 for `BG` and `NOBG` and the data from participant 4 for `BG` were not retained in the subsequent analysis.

In summary, the following data was discarded: the results corresponding to the largest value of `AVangle`, the results from participant 4 for `BG`, and all the results from participant 9. The abnormal slopes obtained indicates that these two participants did not perform the task correctly.

8.3.2 Main analysis: psychometric functions

Psychometric functions were fitted using `psignifit` after removing the outliers from the data. Confidence intervals were found by the BC$_a$ bootstrap method implemented in `psignifit`, based on 10000 simulations [Wichmann and Hill, 2001b]. The two fits corresponding to the two values of `background`, along with the data points, are plotted in Figure 8.6. The four parameters α, β, γ, and λ of the fits are given in Table 8.4 along with the deviance on each fit. The deviance is the sum of the absolute differences between the predicted response and the data. The deviance is a measure of goodness-of-fit, a smaller value indicating a better fit for a given dataset [Agresti, 2007].

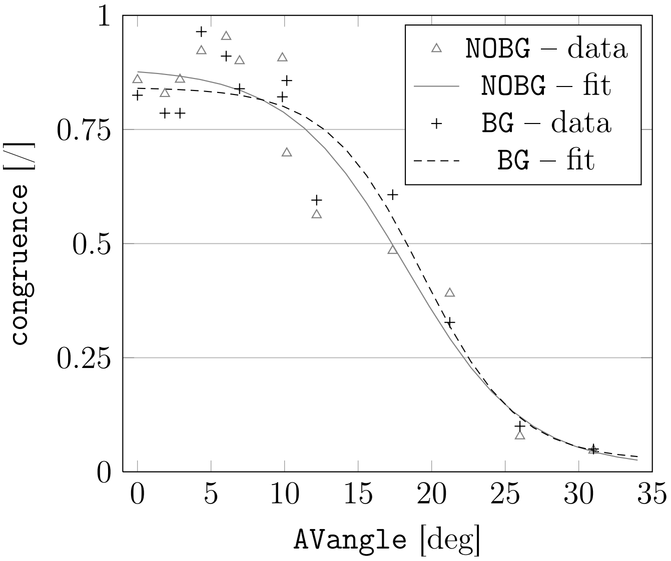


Figure 8.6: Mean responses over all participants for each condition, and associated psychometric functions, after discarding the outliers.

Table 8.4: Parameter estimates from `psignifit` corresponding to the curves in Figure 8.6.

Parameter	NOBG	BG
α [deg]	18.3	19.4
β [deg^{-1}]	-4.0	-3.2
γ [/]	0.009	0.025
λ [/]	0.115	0.158
Deviance	34.7	24.5

The corresponding threshold and slope estimates (as well as their 95% confidence interval) are given in Table 8.5. There is an overlap in the confidence intervals of both estimates. Hence, the presence of background noise did not yield a statistical difference between the estimates. Still, the mean threshold and slope estimates are larger in absolute value in presence of background noise.

Table 8.5: Estimates of the point of subjective equality (PSE) and the slope of the psychometric curve at that point.

Quantity	Condition	Estimate	Confidence interval
PSE [deg]	`NOBG`	18.3	$[16.8, 19.9]$
PSE [deg]	`BG`	19.4	$[17.8, 21.0]$
Slope [deg^{-1}]	`NOBG`	-0.062	$[-0.083, -0.046]$
Slope [deg^{-1}]	`BG`	-0.077	$[-0.120, -0.054]$

8.3.3 Compression of the audio space

To understand how the mean score values relate to the positions of the sound and the positions of the participants, the number of yes and no answers are reported in Table 8.6, summed over all participants. In this section, only the `NOBG` case is considered. Similar results are obtained in the `BG` case.

Table 8.6: Counts of yes/no answers summed over all participants, discarding the outliers, for each sound position ($A_{\text{①}}$ to $A_{\text{⑤}}$ and A_c) and each participant position (S_1 to S_3). Due to minor technical glitches, 17 values were not recorded (out of 1152 trials).

	$A_{\text{①}}$	$A_{\text{②}}$	$A_{\text{③}}$	$A_{\text{④}}$	$A_{\text{⑤}}$	A_c
S_1	56/8	56/8	54/10	59/4	49/15	5/59
S_2	53/11	58/5	61/3	44/19	36/27	3/59
S_3	53/9	54/6	58/6	31/32	25/37	NA

In order to determine the values of `AVangle` at which the perception of the `congruence` is statistically different from that at `AVangle` $= 0$, a χ^2 test was performed for each sound position (see Appendix G for information on the statistical procedure). At each sound position, except at the control position A_c, one sample corresponds to S_1 and serves as a reference for the congruence (`AVangle` $= 0$). Since there is no reference sample at A_c (all the samples are incongruent by design), we include in this particular χ^2 test the values obtained at S_1 with the sound source at $A_{\text{③}}$. Therefore, each χ^2 test was performed on three populations (df $= 2$), corresponding to three different values of `AVangle`, or four populations (df $= 3$) at the control position A_c.

Table 8.7: Results of the χ^2 test comparing the data collected at each participant position for each sound position (first line), and the corresponding p-values (second line).

	$A_{\textcircled{1}}$	$A_{\textcircled{2}}$	$A_{\textcircled{3}}$	$A_{\textcircled{4}}$	$A_{\textcircled{5}}$	A_c
χ^2	0.56	0.73	4.3	30.2	17.0	117.7
p	0.75	0.70	0.12	$< 10^{-6}$	0.0002	$< 10^{-6}$

The results of the tests are given in Table 8.7. The table shows that the value of the χ^2 statistic increases almost monotonically with the distance from S_1 to the sound position (decreasing value of ρ in Equation (8.1)). The lower value of the χ^2 statistic when the sound is located at $A_{\textcircled{5}}$ is a result of the lower proportion of yes answers at S_1 (the reference sample) for this sound position. When the sound is placed too far away, the assumption that only adjusting the sound level is enough to maintain the congruence at S_1 ceases to be valid.

At the three closest sound positions (`AVangle` $< 10°$), the proportions obtained in each test are statistically identical, irrespective of `AVangle`. An overall estimate $\bar{p}$ can be computed for each group of tested samples by collapsing all the corresponding counts in Table 8.6. The resulting mean proportion $\bar{p}$ is 0.85, when the sound is at $A_{\textcircled{1}}$, and 0.90 when the sound is at $A_{\textcircled{2}}$ and $A_{\textcircled{3}}$.

At $A_{\textcircled{4}}$, $A_{\textcircled{5}}$, and the control position A_c, the χ^2 test reaches significance at the 0.05 level, and therefore at least one proportion is different from the others. The Marascuilo procedure is applied to compare all pairs of proportions [Marascuilo, 1966]. At each sound position, all comparisons between the reference sample and the two other samples are significant, except for the comparison between the proportions at S_1 and S_2 when the sound is at $A_{\textcircled{5}}$. Note, however, that the significance is obtained if the sample at S_1 is replaced by another reference sample, indicating once again that this result is obtained because of the lower count of yes at S_1 when the sound is at $A_{\textcircled{5}}$.

8.4 Discussion

8.4.1 General discussion

An increasing angular error `AVangle` between the sound position and the perceived character position decreased the reported `congruence`, i.e. the proportion of "yes" answers to the judgment of the spatial congruence between the sound and the image of the character. The presence of background noise increased both the point of subjective equivalence (PSE) and the absolute value of the slope of the psychometric curve. This means that the congruence was maintained at slightly higher angular separations and that the stimuli were separated more clearly into two categories. This effect, however, was not statistically significant.

When the angle of error between the sound and the image was greater than 10°, the congruence statistically significantly decreased. The reported `congruence` continued to decrease with increasing angular discrepancy.

When the angle of error was smaller than 10°, the reported feeling of congruence was statistically independent of the angle of error, and the congruence score was maximal, between 0.85 and 0.9. These values of the angle of error (below 10°) also correspond to the cases where the sound was located nearest to the screen. This indicates that the method consisting in pulling the audio sources close to the screen with respect to an "ideal" viewer (Section 8.1.3) helps to improve the audio-visual congruence when accurate spatial sound is used in combination with s-3D images. For memory, the experiment was carried out for sound sources located in the horizontal plane.

The window of bimodal integration obtained in this experiment is far larger than those obtained with arbitrary stimuli. In Figure 8.6, the integration is close to maximum up to about 10°. In particular laboratory conditions, however, humans are able to discriminate auditory-visual stimuli discrepant by only 1° [Perrott, 1993]. To the best of our knowledge, no experiment available so far measured the bimodal minimum angle using WFS. However, data on the minimum audible angle obtained with WFS is available. Start [1997] measured the sound field produced by a single loudspeaker and a virtual source on a WFS loudspeaker array (24 loudspeakers each separated by 11 cm) with a KEMAR dummy head. The stimuli were broadband and band-limited white noises from 100 Hz to 8000 Hz, and from 100 Hz to 1500 Hz, respectively. Then, the recorded (binaural) signals were played to participants through headphones. A 2-AFC paradigm was used to evaluate the minimum audible angle (MAA) for each stimulus. No difference was found between the real and the virtual source, both for the broadband stimulus (MAA = 0.8°) and the band-limited stimulus (MAA = 1.1°). Our experimental method differs from the experiments measuring a minimum discriminable angle in two ways: the stimuli is such that the unity assumption holds, and the participant is asked specifically to focus on the spatial coherence of the stimuli, without any previous training. The tasks measuring a minimum discriminable angle rather involves making a left/right judgment on the stimuli after extensive practice.

For angular errors above 10°, the curve decreases rather sharply, crossing the PSE at 18.3° (for NOBG) and 19.4° (for BG). Our results are in agreement with those found in the literature on multimedia perception. Combining an HDTV with ten monaural loudspeakers, Komiyama [1989] showed that non-expert listeners found an angular error of 20° acceptable. De Bruijn and Boone [2003] combined standard 2D video and WFS to build a videoconferencing system. The participants seated off-axis rated as annoying discrepancies of 14° and 15° between the sound and the image. The thresholds obtained here fall within the range defined by these two references. With a setup similar to that of de Bruijn and Boone, Melchior et al. [2003] used the ITU-R 5 grade-impairment scale [ITU, 2003], which they scaled between 0 and 100. The threshold for a slightly disturbing angular error, measured as the 50% crossing on the psychometric curve, was between 5 and 7° for various audiovisual source positions. However, the participants to this experiment were trained to detect small audiovisual discrepancies. Melchior et al. [2006] considered an augmented reality system consisting of a Head-Mounted Display (HMD) with s-3D video and WFS. The ITU-R 5 grade-impairment scale was used again. It was not indicated in the study whether, or how, the subjects were trained. The reported threshold was approximately 4 to 6° (on-axis) and 6 to 8° (off-axis). We argue that the results reported in the last two studies were obtained with participants that were not completely naive with respect to the localization task, which explains the lower thresholds. Nonetheless, it is interesting to note that the average performance of the participants in these experiments was around 90% in the reference case, when the sound matched the image. This is only slightly better than the value obtained here, i.e. around 85%.

8.4.2 Impact of the ambient noise level on the bimodal integration

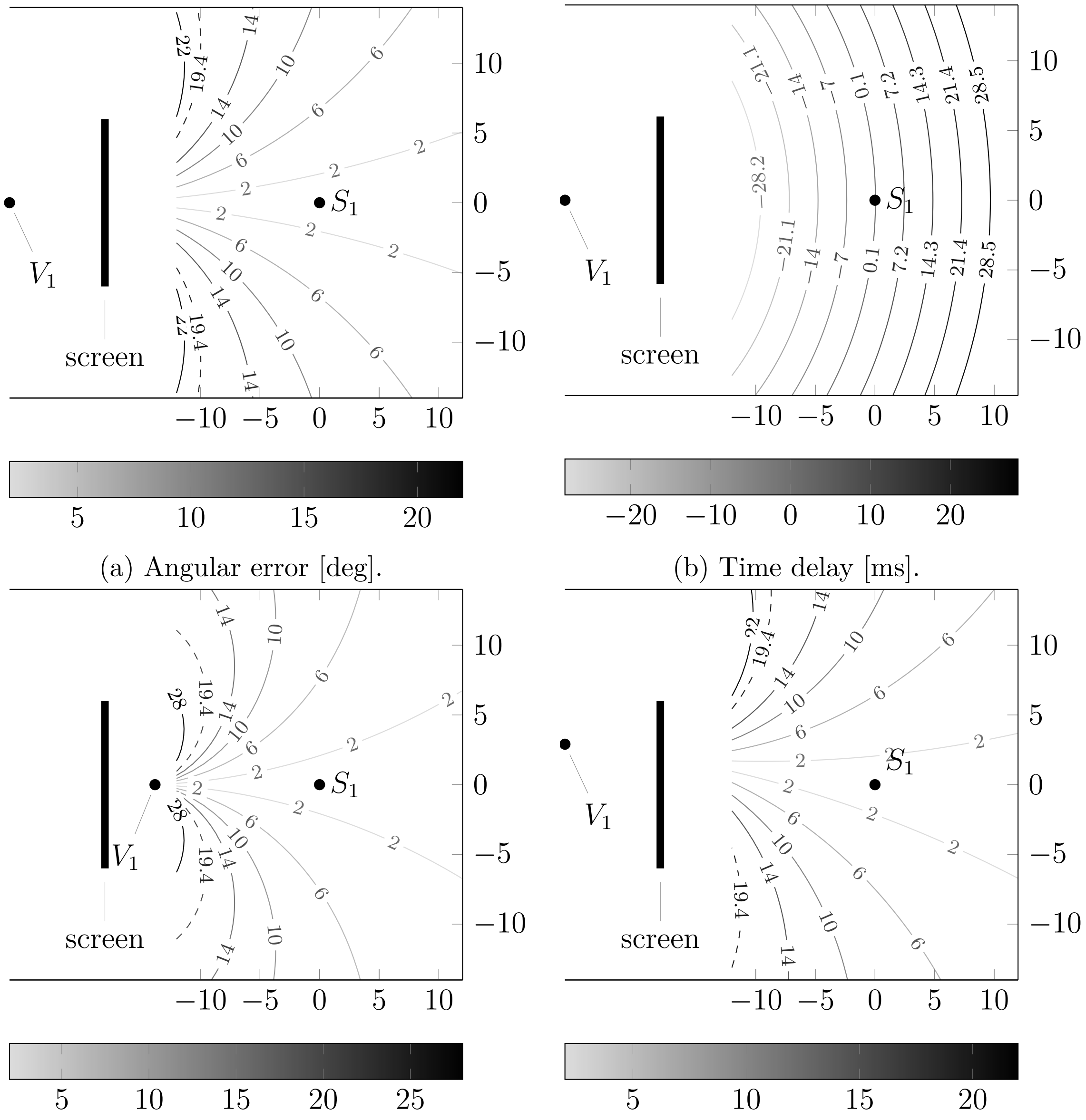
The non-significant effect of the background ambient noise level can be explained by the relatively high signal-to-noise ratio (SNR) used in our experiment. Motivated by the cinema context, we chose to favor speech intelligibility over a high noise level. This resulted in an SNR that was too high to observe any significant degradation in sound localization. Lorenzi et al. [1999] found that the subject localization accuracy remained unaffected by noise at positive SNRs, with a signal level at 70 dB SPL. A similar conclusion was drawn by Good and Gilkey [1996], where the SNRs were relative to the subject detection threshold when both the signal and the masker were coming from the same speaker. The error in the left/right dimension increased only when the SNRs were negative. In an experiment using headphones, Braasch and Hartung [2002] found no difference in localization performance with and without a distractor in the frontal direction with 0 dB SNR (signal at 70 dB SPL). This conclusion was shown both in an anechoic environment and in a reverberant environment. All three studies showed that the accuracy of localization judgments decreased only when the SNR became negative.

The results from Lorenzi et al. [1999] also provide some insight on the localization performance which can be expected with more than one source present at a time. Indeed, the researchers tested three different locations of the masker with respect to the listener: to the left, in front, and to the right. At positive SNRs, the localization accuracy was independent of the masker location. Therefore, it seems that the level of the most important sound sources in the scene should always be superior to the level of less important sources in the background. This would guarantee the best localization performance of the important sources.

8.4.3 Audiovisual spatial and time coherence in a movie theater

In this section, the results are discussed in a cinema context. We consider a movie theater with the following dimensions: distance of 28 m between the side walls, distance of 30 m between the screen and the back wall, screen width of 12 m, first row of seats at 6 m from the screen, ideal viewpoint (the origin of axes) at 18 m from the screen, equidistant from the side walls. Sound sources are reproduced at the intended position of the visual object through WFS and are therefore perceived all over the room at the intended position.

The audiovisual source, V_1 in our example, is obtained with a visual stimulus on the screen that has a parallax of 2 cm, so that the geometrical model developed in Section 6.3 is applicable. In Figure 8.7(a), the value of the angular error between the sound and the image is shown for all possible location in this theater when the parallax is positive (the source appears behind the screen). The angular error is the largest near the side walls and close to the screen. Because of the geometry, the values are always symmetric with respect to the line joining S_1 and V_1. In Figure 8.7(c), the value of the angular error resulting from a negative parallax is shown (the source appears in front of the screen). The angular error is increased with respect to the previous case. The largest angular errors are still found at the same location. In Figure 8.7(d), the value of the angular error is shown when



(a) Angular error [deg].

(b) Time delay [ms].

(c) Angular error [deg].

(d) Angular error [deg].

Figure 8.7: Layout of a movie theater with overlayed curves of (a), (c), (d) constant angular error (in degree) and (b) time delay (in ms) at each location compared to the ideal viewpoint S_1, the origin of axes (dimensions in m). The audiovisual source V_1 is obtained with a visual stimulus on the screen that has a parallax of (a), (b) 2 cm, (c) −2 cm, and (d) 2 cm, combined with a shift of the stimulus 2 m to the right of the screen. The dashed line in each of (a), (c), and (d) corresponds to the threshold of 19.4° in presence of background noise.

a positive parallax is combined with a 2 m shift to the right on the screen (the source still appears behind the screen). Compared to the first case, the value of the angular error still increases closer to the screen. Because the angle values are symmetric with respect to the S_1V_1 line and the rows of seats are vertical lines in the figure, the largest values are found at the top of the figure. The magnitude of the error is comparable to that of Figure 8.7(a).

In Figure 8.7(b), the time delay perceived at each location with respect to the ideal viewpoint is also shown. These values fall in the temporal integration window found in the literature [Lewald and Guski, 2003; van Wassenhove et al., 2007].

In Figure 8.7(a), about 95% of the possible seating area of the cinema theater is below the threshold of 19.4°. Similarly, 94% and 96% of the possible seating area of the cinema theater are below the threshold in Figure 8.7(c) and Figure 8.7(d), respectively. The regions with an angular error above the threshold are those closest to both the screen and the side walls. Additionally, the impact of the time delay found in Figure 8.7(b), which is always below 50 ms in absolute value, should be minimal according to Slutsky and Recanzone [2001]. By comparison, the largest time delay in our experiment was 3.4 ms.

If still desired, one way to increase the sweet-spot for correct reproduction is to shrink the audio space towards the screen. The three congruence measures at the lowest stimulus values each correspond to a different seat and were obtained with the sound located at $A_{①}$, i.e. nearest to the screen. The reported feeling of `congruence` at these angles of error was maximal and relatively independent of the participant position. This indicates that the method illustrated in Figure 8.1 is adequate to improve the audiovisual congruence when accurate spatial sound is used in combination with s-3D images.

8.5 Conclusion

A study of the auditory-visual spatial integration of 3D multimedia content by naive subjects is presented. The audiovisual rendering was provided by a combination of passive s-3D imaging and acoustic WFS.

A subjective experiment was carried out where an angular error between an s-3D video and a spatially accurate sound reproduced through WFS was presented to naive subjects. Motivated by a cinema application, we chose a stimulus consisting of a talking character in an apartment scene. The psychometric curve was obtained with the method of constant stimuli, and the threshold of bimodal integration was estimated. After a five-second speech stimulus, subjects gave their answer to the question "Is the voice coherent with the character's position?" The ambient noise level was varied as an independent variable with two levels. The point of subjective equivalence (PSE), where the subjects answered "yes" half of the time, was 18.3° when only the speech signal was present, and 19.4° when additional ambient noise was present, with an SNR of 4 dBA. These values are much higher than the minimum audible angle obtained in particular laboratory conditions and reported in the literature. Differences in experimental design that lead to this result are discussed.

In addition to the slight threshold increase with ambient noise, an increase in the absolute value of the slope was observed. This means that the feeling of congruence was maintained at higher separation angles and that the stimuli were separated more clearly into two categories. These effects, however, were not statistically significant. We argue that this was because the SNR was too high to observe any significant degradation in sound localization.

The WFS employed in the experiments offers optimum localization accuracy but employed a large, impractical number of loudspeakers (one every 20 cm) for a large installation. The results indicated that audiovisual spatial congruency was obtained with rather large angular disparities ($\simeq 19°$). Practical 2D or 3D WFS applications consisting of sparse loudspeaker arrays, as described by Corteel et al. [2012], offer good, though non-ideal, localization accuracy over an extensive listening area. The limited localization error provided by such practical systems should therefore offer a similar level of audiovisual spatial congruency maintaining the benefits of WFS (consistency of spatial impression over a large listening area, limited perception of individual loudspeakers, power efficient rendering through the use of multiple loudspeakers for each source position) with a realistic installation complexity (once or twice the number of loudspeakers of today's typical theater installations). Nonetheless, the method consisting of compressing the audio towards the screen was proven to be adequate when accurate spatial sound was used in combination with s-3D video.

Further studies should consider spatial congruence for audiovisual disparities in the vertical plane using either 3D WFS and/or other 3D audio rendering techniques by simulating different characters at different heights. The audio-visual coherence evaluation of several concurrently active sources should also be studied.

Bibliography

Agresti, A., 2007. An introduction to categorical data analysis, 2nd Edition. Wiley Series in Probability and Mathematical Statistics. Wiley-Interscience, Hoboken, NJ. 194

André, C. R., Corteel, É., Embrechts, J.-J., Verly, J. G., Katz, B. F., Jan. 2014. Subjective evaluation of the audiovisual spatial congruence in the case of stereoscopic-3D video and Wave Field Synthesis. International Journal of Human-Computer Studies 72 (1), 23–32.
http://dx.doi.org/10.1016/j.ijhcs.2013.09.004 180

André, C. R., Corteel, É., Embrechts, J.-J., Verly, J. G., Katz, B. F. G., 2013. A new valited method for improving the audiovisual spatial congruence in the case of stereoscopic-3D video and Wave Field Synthesis. In: 2013 International Conference on 3D Imaging (IC3D). pp. 1–8. 180

Braasch, J., Hartung, K., 2002. Localization in the Presence of a Distracter and Reverberation in the Frontal Horizontal Plane. I. Psychoacoustical Data. Acta Acust. united with Acust. 88 (6), 942–955. 199

Corteel, É., Rohr, L., Falourd, X., NGuyen, K.-V., Lissek, H., Apr. 2012. Practical 3-dimensional sound reproduction using Wave Field Synthesis, theory and perceptual validation. In: Proceedings of the 11th French Congress of Acoustics and 2012 Annual IOA Meeting. Nantes, France, pp. 895–900. 202

Courgeon, M., Clavel, C., 2013. MARC: a framework that features emotion models for facial animation during human–computer interaction. J. Multimodal User Interfaces , 1–9.
http://dx.doi.org/10.1007/s12193-013-0124-1 188

Courgeon, M., Rébillat, M., Katz, B. F., Clavel, C., Martin, J.-C., Dec. 2010. Life-sized audiovisual spatial social scenes with multiple characters: MARC & SMART-I^2. In: Proceedings of the 5$^{\text{èmes}}$ Journées de l'AFRV. Orsay, France. 188

de Bruijn, W. P. J., Boone, M. M., 2003. Application of Wave Field Synthesis in life-size videoconferencing. In: Audio Eng. Soc. Conv. 114. 181, 182, 183, 198

Doukhan, D., Rilliard, A., Rosset, S., Adda-Decker, M., d'Alessandro, C., 2011. Prosodic analysis of a corpus of tales. In: INTERSPEECH. pp. 3129–3132. 189

Ernst, M. O., Banks, M. S., Jan. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415 (6870), 429–433.
http://dx.doi.org/10.1038/415429a 185

Goldstein, E. B., May 1987. Spatial layout, orientation relative to the observer, and perceived projection in pictures viewed at an angle. J. Exp. Psychol. Human 13 (2), 256–266.
http://dx.doi.org/10.1037/0096-1523.13.2.256 183

Good, M. D., Gilkey, R. H., 1996. Sound localization in noise: The effect of signal-to-noise ratio. J. Acoust. Soc. Am. 99 (2), 1108–1117.
http://dx.doi.org/10.1121/1.415233 198

ITU, Dec. 2003. Recommendation BS.1284. General methods for the subjective assessment of sound quality. ITU-R. 198

Klein, S. A., Nov. 2001. Measuring, estimating, and understanding the psychometric function: a commentary. Perception & psychophysics 63 (8), 1421–1455. 185

Komiyama, S., 1989. Subjective evaluation of angular displacement between picture and sound directions for HDTV sound systems. J. Audio Eng. Soc. 37 (4), 210–214.
http://www.aes.org/e-lib/browse.cfm?elib=6094 182, 187, 198

Lam, C. F., Dubno, J. R., Mills, J. H., 1999. Determination of optimal data placement for psychometric function estimation: A computer simulation. J. Acoust. Soc. Am. 106 (4), 1969.
http://dx.doi.org/10.1121/1.427944 187

Lewald, J., Guski, R., May 2003. Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. Cognitive Brain Research 16 (3), 468–478.
http://dx.doi.org/10.1016/S0926-6410(03)00074-0 201

Leys, C., Ley, C., Klein, O., Bernard, P., Licata, L., Jul. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. J. Exp. Soc. Psychol. 49 (4), 764–766. 193

Lorenzi, C., Gatehouse, S., Lever, C., 1999. Sound localization in noise in normal-hearing listeners. J. Acoust. Soc. Am. 105 (3), 1810–1820.
10.1121/1.426719 198, 199

Marascuilo, L. A., 1966. Large-sample multiple comparisons. Psychol. Bull. 65 (5), 280–290. 196

Melchior, F., Brix, S., Sporer, T., Roder, T., Klehs, B., 2003. Wave Field Synthesis in Combination with 2D Video Projection. In: AES 24th International Conference.
http://www.aes.org/e-lib/browse.cfm?elib=12268 182, 198

Melchior, F., Fischer, J., de Vries, D., 2006. Audiovisual perception using Wave Field Synthesis in combination with augmented reality systems: Horizontal positioning. In: AES 28th International Conference.
http://www.aes.org/e-lib/browse.cfm?elib=13832 182, 198

Perrott, D. R., Jun. 1993. Auditory and visual localization: Two modalities, one world. In: AES 12th International Conference. pp. 221–231. 197

Slutsky, D. A., Recanzone, G. H., Jan. 2001. Temporal and spatial dependency of the ventriloquism effect. Neuroreport 12 (1), 7–10. 201

Start, E. W., Jun. 1997. Direct sound enhancement by Wave Field Synthesis. Ph.D. thesis, TU Delft, The Nederlands. 197

Theile, G., Wittek, H., Reisinger, M., 2003. Potential Wavefield Synthesis applications in the multichannel stereophonic world. In: Audio Eng. Soc. 24th Int. Conf.: Multichannel Audio, The New Reality.
http://www.aes.org/e-lib/browse.cfm?elib=12280 181

Treutwein, B., Sep. 1995. Adaptive psychophysical procedures. Vision research 35 (17), 2503–2522. 191

van Wassenhove, V., Grant, K. W., Poeppel, D., Jan. 2007. Temporal window of integration in auditory-visual speech perception. Neuropsychologia 45 (3), 598–607.
http://dx.doi.org/10.1016/j.neuropsychologia.2006.01.001 201

Wichmann, F. A., Hill, N. J., Nov. 2001a. The psychometric function: I. Fitting, sampling, and goodness of fit. Perception & Psychophysics 63 (8), 1293–1313. 190, 191, 192

Wichmann, F. A., Hill, N. J., Nov. 2001b. The psychometric function: II. Bootstrap-based confidence intervals and sampling. Perception & Psychophysics 63 (8), 1314–1329. 194

Chapter

9

Conclusions and perspectives

Highlights

✓ Conclusions of this thesis are drawn.

✓ Perspectives are suggested for future work.

Contents

9.1 Conclusions

In this thesis, we study the addition of spatially accurate sound rendering to regular stereoscopic-3D (s-3D) images in a cinema theater. Our goal is to provide a perceptually matching sound source at the perceived position of every object producing sound in the scene. This thesis examines, and contributes to, the questions of usefulness and practical feasibility of this goal. In particular, the following questions are considered, in the s-3D cinema context:

Usefulness Does a higher spatial coherence between sound and image leads to an increased sense of presence for the audience, compared to traditional (stereophonic) sound rendering?

Feasibility In which conditions can the angular error between the sound and the image be detected, when presenting precise spatial sound through Wave Field Synthesis (WFS) in combination with s-3D images to spectators seated at different locations?

Before investigating the "usefulness" issue, we needed a way to obtain a content combining 3D audio with s-3D video. Therefore, we defined and implemented a strategy for producing a true, experimental 3D audiovisual content.

With respect to this experimental content, the main contributions are:

- the re-expression in matrix form of a mathematical model which transforms the 3D coordinates of a visual object captured by an s-3D camera into the 3D coordinates where the object appears when the movie is played;
- a process describing the creation of a 3D audio track for an existing animation movie, from the early editing stage, through the scripting process, to the sending of data to the rendering system.

To investigate the "usefulness" issue, a between-subject experiment was designed with three different soundtracks. The sense of presence was evaluated with a post-session questionnaire and heart rate monitoring. The main findings of this experiment are as follows.

- The soundtrack condition does not affect the reported presence score directly for all subjects.
- The soundtrack condition only impacts the sense of presence of the group of participants who reported the highest level of presence.
- In the group that reported the highest sense of presence, for which sound rendering condition was influential, the spatially coherent soundtrack (WFS rendering) is statistically significantly different from the two other stereo soundtracks. The WFS soundtrack leads to a decreased reported sense of presence.
- The change in heart rate over time was quantified by using the Heart Rate Variability (HRV) measures. The analysis of HRV measures shows that HRV

allows one to discriminate between a baseline state and the movie presentation. In addition, all the HRV frequency domain parameters correlate to the reported presence scores.

A potential explanation for the decreased reported sense of presence when the WFS-rendered soundtrack was presented is that presence is lessened when the auditory objects extend beyond the screen boundaries. Further studies with different source material would be required to substantiate this hypothesis.

To investigate the "feasibility" issue, an experiment was conducted with naive spectators, seated at different locations, to evaluate the threshold of bimodal integration associated with the angular error between WFS audio and s-3D video. In addition, the ambient noise level was varied as an independent variable with two levels. The main findings of this experiment are:

- The point of subjective equivalence (PSE), where participants were equally likely to detect the angular error, is 18.3° (SNR = 19 dB) and 19.4° (SNR = 4 dB). The difference in PSE with and without additional ambient noise is not statistically significant.
- The method consisting in pulling the audio sources close to the screen with respect to an "ideal" viewer helps to improve the audiovisual congruence when accurate spatial sound is used in combination with s-3D images.
- Precise spatial sound reproduction is theoretically possible in a movie theater with a realistic installation complexity (once or twice the number of loudspeakers of today's installations).

9.2 Perspectives

9.2.1 Incoherence in the perceived auditory-visual distance

This section is a follow-up on the discussion in Section 6.6.2, where we find that the perceived auditory distance does not always match the perceived visual distance resulting from stereoscopic fusion of the images. This incoherence is a direct consequence of the mathematical model developed in Section 6.3. In this model, any point in the **Blender** space which does not produce any horizontal disparity in the left and right projection planes, i.e. for which $X_{c_l} = X_{c_r}$, is localized in the physical world at the depth of the screen plane, because the disparity D defined in Equation (6.15) is zero in this case. In the case of a parallel camera configuration,

$X_{c_l} = X_{c_r}$ holds for points in the **Blender** space in the plane parallel to the interaxial line which includes the intersection of the cameras optical axes, that is the plane with $Z_0 = -C$. In the physical space coordinates (the physical world), these points always appear at the depth of the screen plane, which is located at a distance V from the (ideal) spectator. However, the stereoscopic parameter C can be varied for artistic purposes. In the particular scene described in Section 6.6.2, small characters appear at the depth of the screen plane. From this information, the spectator can infer that C in this shot is larger than C in the previous close-up shot. Obviously, the distance between the spectator and the screen, V, has not changed between the two shots, and an incoherence is perceived.

In our experiment, we chose to introduce a different distance value to the reverberation engine to mitigate this perceptual conflict. In future work, a mathematical model of stereopsis which also takes into account the effect of linear perspective is needed. We now discuss two different strategies to obtain this new model.

9.2.1.1 Visual integration of perspective cues and disparity cues

The integration of sensory cues by human subjects has been shown several times in the literature to be nearly statistically optimal [Ernst and Bülthoff, 2004]. This has already been discussed in this thesis in the case of an auditory-visual stimulus (Section 2.3.2).

Therefore, we might hypothesize that the depth defined by monocular cues may be integrated with the depth defined by horizontal disparity in a similar optimal fashion. However, Figure 6.12 contains numerous monocular depth cues: relative size and relative density are combined to produce linear perspective and texture gradients. Determining whether an optimal model could indeed model the sensory integration here would require to consider each of these depth cue (including horizontal disparity) separately to determine the weights associated to each one. Then the results predicted by the statistically optimal model should be compared to the results of an experiment combining all the considered cues.

This procedure might not prove successful, as the literature comparing one perspective cue to horizontal disparity yields contradictory results. On the one hand, Hillis et al. [2004] showed that texture gradient and horizontal disparity were optimally integrated when estimating slant. One should note that slant estimation is not strictly equivalent to depth estimation. On the other hand, Zalevski et al. [2007] could not accurately describe their data on the combination of perspective cues (relative size and relative density) and horizontal disparity cues in discriminating depth (stereoacuity).

Researchers interested in vision science may consider this research question. In the next section, we describe a potential experiment closer to our own work.

9.2.1.2 Impact of the camera lens on the perceived auditory-visual distance

Here, we consider the solution given in Section 6.6.2 to the perceived incoherence, that is, one distance for the rendering of direct sound and another, termed `reverb`, for the rendering of reverberant sound. We suggest an experiment in two steps, much like the experiment reported in [Kruszielski et al., 2011].

First, we evaluate the impact of the `reverb` distance on the perceived audiovisual distance for several values of the field-of-view (FOV) of the camera, denoted by α. The FOV α is related to the stereoscopic camera parameters (see Figure 6.4) through

$$\alpha = \text{atan}\,(\frac{W_c + h}{f}) + \text{atan}\,(\frac{W_c - h}{f}). \tag{9.1}$$

In [Kruszielski et al., 2011], three different (2D) images of the same scene are captured: (1) with a 70° FOV, at about 3 m from the object of interest in the scene, (2) with a 70° FOV, at about 1.5 m from the object of interest in the scene, and (3) with a 20° FOV, at about 3 m from the object of interest in the scene. The configuration in (3) is calibrated so that the object of interest, in the foreground, appears to be the same size as in (2), but the background is different. More can be seen of the background in (2) than of the background in (3). Similar captures of a scene could be done in s-3D.

Since linear perspective is involved, the participants have to be able to evaluate the size in the physical world of the object of interest in the image. We suggest a mobile phone. Participants can be shown a real mobile phone before the experiment in order to provide an anchor for the size of the mobile phone in the virtual world. In addition, localizing a mobile phone by its ringtone is a usual task.

In the experiment, participants could be asked to adjust the level of reverberant sound, through a direct manipulation of the `reverb` distance, to the level they feel is right for the presented image. The outcome of this experiment is a reference value of the `reverb` distance for each presented s-3D image.

Second, we evaluate the subjective impact of modifying this reference value of the `reverb` distance on the perceived audiovisual distance. Particularly, we should evaluate whether modifying this value can trick the participants into perceiving the object closer or farther in depth than it actually is. One interesting potential result of this evaluation is related to the depth budget, an important s-3D measure, which is equal to the total amount of depth in front and behind the screen plane which can be comfortably perceived. If, by modifying the `reverb` distance, an audiovisual object appears further behind the screen plane, or closer in front of the screen plane, then 3D sound can be used as a mean to increase the (visual) depth budget, without modifying the display size. This would confirm, using virtual sound sources, the results obtained by Turner et al. [2011] with real sound sources.

9.2.2 Spatially accurate sound rendering all around the audience

When adding spatially accurate sound rendering to s-3D movies, the aim is to provide the moviegoer with more coherent auditory-visual cues. Therefore, one would want to position sound sources in a volume at least corresponding to the audience's field of view. The question of whether a movie theater sound system should be able to render spatially accurate sound sources in the whole 3D sphere needs to be addressed by the people involved in the process of 3D moviemaking.

If moviemakers prefer to limit 3D sound to objects inside the field of view, one can imagine a hybrid solution combining a frontal system capable of precise localization and one or several ambience channels. In that case, 3D audio would be used only to render the material that is currently sent to the three front channels, including the dialogs that are currently never spatialized for lip-sync reasons. The ambience channels can then be rendered by the existing loudspeaker arrays.

9.2.3 Personalized 3D soundscape in s-3D cinema

The results in Chapter 8 suggest that the personalization of the sound rendering is not needed when combining spatially accurate sound rendering to s-3D movies, because of the visual capture of sound. Nevertheless, the impact of a binaural rendering of the soundtrack of an s-3D movie* should be evaluated [André et al., 2010], if only because mobile applications (mobile phones, tablets, ...) are a growing market and already require the use of headphones.

*There is equivalent interest for the addition of 3D soundtracks to regular 2D movies.

In cinemas, such a sound rendering raises questions that go beyond mere technical considerations. Several consequences, positive or negative, follow. In terms of accessibility, for example, the user could choose the language of the soundtrack. In terms of social interactions, wearing headphones prevent the moviegoers from interacting with each other, thereby removing an important part of the cinema experience for some of them. Finally, practical constraints for the cinema operator should be expected.

9.2.4 Impact of 3D sound on the perception of s-3D stimuli

As previously mentioned, Turner et al. [2011] have shown that 3D sound can have an influence on the perceived depth of auditory-visual stimulus. It would be interesting to investigate the potential influence of 3D sound on other aspects of s-3D perception. We suggest to investigate the influence of 3D sound on s-3D image impairments [Meesters et al., 2003]. These include (1) the puppet-theater effect [Yamanoue et al., 2006], a miniaturization effect which makes people in the scene look like tiny animated puppets, (2) the cardboard effect [Yamanoue et al., 2000], which makes the objects in the scene appear flat but still separated in depth, and (3) the window violation [Devernay and Beardsley, 2010], when an object appears in one image of the s-3D pair, but not in the other.

9.2.5 Other contexts

All the work in this thesis focused on the s-3D cinema context. Other applications might require different technological choices. Nowadays, movies are consumed (as a product) not only in cinema theaters, but also on television (home-cinema), on computers, and on mobile devices. Furthermore, the distinction between these differing modes of consumption tends to blur. One might, for example, buy a movie on a web-based video-on-demand service, start watching the movie on a television, and later catch the rest of the movie on a mobile device. There is a need for digital formats that cater to all these different needs. The same requirement for interoperability is therefore applicable to new 3D sound formats.

Bibliography

André, C. R., Embrechts, J.-J., Verly, J. G., Nov. 2010. Adding 3D sound to 3D cinema: Identification and evaluation of different reproduction techniques. In:

Proc. 2nd Int. Conf. on Audio Language and Image Processing (ICALIP 2010). Shanghai, China, pp. 130–137.
http://hdl.handle.net/2268/73310 212

Devernay, F., Beardsley, P., Jan. 2010. Stereoscopic cinema. In: Ronfard, R., Taubin, G. (Eds.), Image and Geometry Processing for 3-D Cinematography. No. 5 in Geometry and Computing. Springer, pp. 11–51. 213

Ernst, M. O., Bülthoff, H. H., Apr. 2004. Merging the senses into a robust percept. Trends in Cognitive Sciences 8 (4), 162–169.
http://dx.doi.org/10.1016/j.tics.2004.02.002 210

Hillis, J. M., Watt, S. J., Landy, M. S., Banks, M. S., Jan. 2004. Slant from texture and disparity cues: Optimal cue combination. Journal of Vision 4 (12), 1.
http://dx.doi.org/10.1167/4.12.1 210

Kruszielski, L. F., Kamekawa, T., Marui, A., 2011. The influence of camera focal length in the direct-to-reverb ratio suitability and its effect in the perception of distance for a motion picture. In: Audio Engineering Society Convention 131.
http://www.aes.org/e-lib/browse.cfm?elib=16105 211

Meesters, L., IJsselsteijn, W., Seuntiëns, P. J., 2003. A survey of perceptual quality issues in three-dimensional television systems. Vol. 5006. SPIE, pp. 313–326.
http://dx.doi.org/10.1117/12.474132 213

Turner, A., Berry, J., Holliman, N., Feb. 2011. Can the perception of depth in stereoscopic images be influenced by 3D sound? Proceedings of SPIE 7863, 786307.
http://dx.doi.org/10.1117/12.871960 212, 213

Yamanoue, H., Okui, M., Okano, F., Jun. 2006. Geometrical analysis of puppet-theater and cardboard effects in stereoscopic HDTV images. IEEE Transactions on Circuits and Systems for Video Technology 16 (6), 744– 752.
http://dx.doi.org/10.1109/TCSVT.2006.875213 213

Yamanoue, H., Okui, M., Yuyama, I., Apr. 2000. A study on the relationship between shooting conditions and cardboard effect of stereoscopic images. IEEE Transactions on Circuits and Systems for Video Technology 10 (3), 411–416.
http://dx.doi.org/10.1109/76.836285 213

Zalevski, A. M., Henning, G. B., Hill, N. J., Jan. 2007. Cue combination and the effect of horizontal disparity and perspective on stereoacuity. Spatial Vision 20 (1/2), 107–138.
http://dx.doi.org/10.1163/156856807779369706 210

This page intentionally left blank.

Part III

Appendices

Appendix

A

List of utilized software tools

Highlights

✓ The list of software tools used to carry out this thesis is given.

✓ The list of open source software tools is given in Section A.1.

✓ The list of proprietary software tools is given in Section A.2.

✓ Visit the software website by clicking its name.

Contents

A.1 Open source software tools

The following open source software tools were used to carry out this thesis:

- The operating system **Ubuntu** was on virtually all the computers used throughout this thesis, mostly in its versions 10.04 and 12.04 LTS.
- The free numerical computing environment **GNU Octave** and its graphical user interface (GUI) **QtOctave**, as a free alternative to **Matlab**.
- The programming language of the GUIs for the experiments was always **Python**, for its simplicity, its readability, and its cross-platform nature. Along with several standard libraries included within **Python**, the following additional libraries were used:

  - **Numpy** which includes numerous matrix calculation capabilities (among others).
  - **wxPython** to create GUIs.
  - **pyOSC** which implements the OpenSound Control (OSC) communication protocol in pure **Python**.
  - **pywii** which allows to use WiiMotes (the Nintendo Wii Controller) with computers through Bluetooth.

- The statistical analysis of the experiment results were carried out using the software **R**, as well as several freely available packages:
  - **Amelia II**, which performs multiple implementation of datasets with missing values.
  - **Zelig**, which performs analyses of variance (ANOVAs) on multiply imputed datasets.
  - **Mclust**, which performs normal mixture modelling on a dataset.
  - **aplpack**, which contains the `bagplot` function which plots bivariate boxplots.

- This document, as well as all the related publications, were composed using **LaTeX**. The figures were realized thanks to the power of **PGF** and **Ti*k*Z**. The documents were composed first using **Kile** and later using **LaTeXila**. Backup copies were saved on **Subversion** (svn) repositories.

- Although not a computer program, the open source movie "Elephants Dream" was a very important ingredient of this thesis. The 3D computer graphics software **Blender** is strongly associated with this movie.

- The stereoscopic video player **Bino** was used in combination with the **Equalizer** library, which allows to support multi-projector setups. Actually, **Bino** was modified in the course of this work to support OSC communication through **oscpack**. It is written in **C++**.

- Participants' answers to a questionnaire were collected using the online survey application **LimeSurvey**.

A.2 Proprietary software tools

The following proprietary software tools were used to carry out this thesis:

- **Max/MSP**, a visual programming language widely used in the audio community.
- **MARC**, a framework developed at LIMSI-CNRS for addressing the design and the real-time interaction with visually realistic expressive virtual agents.
- **ProTools**, a digital audio workstation.
- The numerical computing environment **Matlab** was necessary to run the **Binaural Cue Selection** toolbox.

Max/MSP was used on **Windows XP** and **ProTools** was used on **Mac OS X**.

This page intentionally left blank.

Appendix

B

Other limitations of s-3D imaging

Highlights

✓ Two artifacts of s-3D imaging are discussed.

✓ Crosstalk is discussed in Section B.1.

✓ Flicker is discussed in Section B.2.

Contents

B.1 Crosstalk

Since the multiplexing technologies described in Section 4.1.3 are not perfect, part of the signal in the channel intended for one eye may find its way into the channel corresponding to the other eye, a phenomenon called *crosstalk*. The perceptual consequence of crosstalk is called *ghosting*. It is the unwanted perception of double contours of objects, resulting from the disparity between the two images and the crosstalk phenomenon. Woods [2010] gave a comprehensive list of mechanisms producing the crosstalk for different s-3D reproduction systems. We give here the mechanisms we are mainly interested in, which correspond to the systems described in Section 4.1.3.

RealD 3D and MasterImage are time-sequential polarized s-3D projection systems. In addition to potential crosstalk from the polarization process, timing considerations must also be taken into account. In summary, the following list gives the potential sources of crosstalk:

- the optical quality of the polarizers,
- the optical quality of the screen,
- the optical quality of the modulator,
- the potentially incorrect orientation of the polarizers, at the projector or in the glasses,
- the time delay between the modulator and the projector.

XpanD uses systems combining DLP projection with active liquid crystal shutter glasses. In essence, DLP projection does not introduce any crosstalk. Therefore, only the shutter glasses are responsible for the crosstalk in these systems. The following characteristics have to be considered:

- the optical performance of the liquid crystal cells in the glasses. This includes the amount of transmission in the “closed” state, the rise time, the fall time, and the amount of transmission in the “open” state,
- the synchronization between the glasses and the display,
- the angle of view through the glasses, because the performance of the liquid crystal cells is only maximum when seen perpendicularly.

Dolby 3D is similar in its concept to anaglyph s-3D. The sources of crosstalk are therefore the same:

- the spectral quality of the color wheel,
- the spectral quality of the glasses, and how well it matches the spectral quality of the color wheel.

One additional source of crosstalk present in anaglyph s-3D is related to the anaglyph generation matrix. There is no generation matrix in Dolby 3D, however, so that this is not a potential source of crosstalk.

A simple mathematical measure of the level of crosstalk is given by the fraction of the RGB value of a pixel in a view that is present in the RGB value for the pixel at the same position in the other view. For example, the level of crosstalk,

p, in the left view is defined as:

$$R'_l(x,y) = \min\left(R_l(x,y) + \frac{p}{100}R_r(x,y), 255\right), \tag{B.1}$$

$$G'_l(x,y) = \min\left(G_l(x,y) + \frac{p}{100}G_r(x,y), 255\right), \tag{B.2}$$

$$B'_l(x,y) = \min\left(B_l(x,y) + \frac{p}{100}B_r(x,y), 255\right), \tag{B.3}$$

where $\left[R'_l, G'_l, B'_l\right]$ is the RGB triplet of the left view with crosstalk, $\left[R_l, G_l, B_l\right]$ is the correct RGB triplet of the left view, and $\left[R_r, G_r, B_r\right]$ is the correct RGB triplet of the right view.

In order to study the perceptual impact of crosstalk, Seuntiëns et al. [2005] used a mirror stereoscope in front of a computer screen. The apparatus shows, by optical means, one half of the screen to each eye. This apparatus is by design free from crosstalk. The researchers added levels of 5, 10, and 15% of crosstalk to two different natural scenes. Although the subjects could effectively detect the increasing levels of crosstalk, their ratings of visual strain and perceived depth remained constant. This result is not quite in agreement with that of Kooi and Toet [2004], who found that 5% of crosstalk reduced the visual comfort "a bit", 15% reduced it "a lot", and 25% reduced it "extremely". Crosstalk also impairs the quality of depth rendering [Tsirlin et al., 2011]. All studies confirm that the amount of distortion is more easily detected with increasing interaxial distance. The general consensus seems to be that, at least for natural scenes with few hard edges, 2% of crosstalk remain acceptable.

B.2 Flicker

Flicker is a perceived fluctuation in the brightness of the visual stimulus. When time multiplexing is used to reproduce the s-3D stimulus to the spectator, the glasses alternatively shows an image to an eye, and then obstruct the light to the same eye. If the interval where the eye receives no light is too long, the spectator may perceive the image with a reduced brightness.

Flicker may cause unwanted depth perception: a flickering stimulus is perceived at a larger depth than a non-flickering one [Miller and Patterson, 1995].

According to Hoffman et al. [2011], the visibility of flicker is mainly influenced by the reproduction rate at the eye. This is the reason why triple-flash presentation is used (see Section 4.1.3). Flicker may appear in regular 2D video reproduction. This is why standards have issued a minimum reproduction rate of 50 Hz. When triple flash is used, each eye receives the image at 72 Hz. Therefore, the reproduction system may be considered flicker-free.

Bibliography

Hoffman, D. M., Karasev, V. I., Banks, M. S., Mar. 2011. Temporal presentation protocols in stereoscopic displays: Flicker visibility, perceived motion, and perceived depth. Journal of the Society for Information Display 19 (3), 271–297. http://dx.doi.org/10.1889/JSID19.3.271 225

Kooi, F. L., Toet, A., Aug. 2004. Visual comfort of binocular and 3D displays. Displays 25 (2–3), 99–108. http://dx.doi.org/10.1016/j.displa.2004.07.004 225

Miller, R. J., Patterson, R., Jul. 1995. Influence of flicker on perceived size and depth. Perception & Psychophysics 57 (5), 604–613. http://dx.doi.org/10.3758/BF03213266 225

Seuntiëns, P., Meesters, L., IJsselsteijn, W., Oct. 2005. Perceptual attributes of crosstalk in 3D images. Displays 26 (4–5), 177–183. http://dx.doi.org/10.1016/j.displa.2005.06.005 225

Tsirlin, I., Wilcox, L., Allison, R., 2011. The effect of crosstalk on the perceived depth from disparity and monocular occlusions. IEEE Transactions on Broadcasting 57 (2), 445–453. http://dx.doi.org/10.1109/TBC.2011.2105630 225

Woods, A., May 2010. Understanding crosstalk in stereoscopic displays. In: Proceedings of the 3DSA conference. Tokyo, Japan. 223

Appendix
C

An overview of the OSC protocol

Highlights

✓ A user-level overview of the OSC communication protocol is given.

✓ Our use of OSC in this thesis is described.

Contents

OpenSound Control (OSC) is a communication protocol widely spread in the audio and computer-assisted music fields. OSC allows to communicate digitally between computers, synthesizers, human interfaces, ... In Section C.1, we give a user-level overview of the OSC communication protocol, as found in [Wright et al., 2003]. Then, in Section C.2, we describe the use we made of OSC in this work.

C.1 Basics

OSC establishes a communication between a client, who sends messages, and a server, who receives them. The server has its own address space, which is a tree structure with named nodes. Each node in the tree is the potential target of a message, the elementary piece of OSC communication. The message is itself composed of an address and arguments.

The address is a string describing the path from the root of the tree to the intended node, with the character / delimiting each node, much like a URL. For example, the address "/root/node1/node2" refers to a node "node2", which is the child of "node1", itself the child of the top-level node "root". The tree structure is entirely arbitrary, which makes it possible to name the nodes in a self-explanatory way.

OSC supports arguments of several types. Mostly ASCII strings, 32 bit floating point and integer numbers were used in this work. For the sake of completeness, it should be noted that the message also contains a string before the arguments, detailing the data type of each argument. However, this string is seamlessly constructed by the libraries we used, based on the type of the argument passed to the message data structure.

When several messages are sent at once, one can group them in a bundle, which is treated as if all messages arrived at the same time.

C.2 Implementations

OSC was already in use in the SMART-I^2 before this work. We describe in this section our own additions to the existing work.

We used two implementations of OSC in the course of this work, namely **pyOSC**, written in **Python**, and **oscpack**, written in **C++**. These libraries send OSC data packets as UDP packets, although this is not specifically required.

We integrated the **C++** library **oscpack** into the sources of the s-3D video player **Bino**. This feature was used in Chapters 6 and 7 to send data from the video engine (`djobi`) to the audio engine (`djoba`). At each video frame, the player would send an OSC message containing all the audio metadata related to sound sources at that frame.

We integrated the python library **pyOSC** into the programs we wrote to control our experiments. We used OSC to communicate between three computers on a local network, namely `djobi`, `djoba`, and the laptop controlling the experiments in Chapter 8. In addition, OSC was used to gather the answers participants gave on WiiMotes.

Bibliography

Wright, M., Freed, A., Momeni, A., 2003. OpenSound Control: State of the art 2003. In: Proc. 2003 Conf. on new interfaces for musical expression (NIME-03). Montreal, Canada, pp. 153–159. 227

Appendix

D

Affine geometry, homogeneous coordinates, and rigid transformations

Highlights

✓ A background in affine geometry is given. In particular, the concept of homogeneous coordinates is introduced.

✓ The manipulation of homogeneous coordinates, namely translation and rotation, is described.

Contents

The mathematical background for affine geometry is given in Section D.1. In particular, the homogeneous coordinates are introduced in Section D.1.5. Then, the manipulation of homogeneous coordinates, namely translation and rotation, is described in Section D.2.

D.1 A brief introduction to affine geometry

D.1.1 Affine spaces

We use the definition of the affine space given in Chapter 1 of Gallier [2011]. This particular definition stresses the physical interpretation of *points* as particles in space and *vectors* as forces acting on these particles. The motivation in defining affine spaces is to define points and properties of points which are independent of the frame which defines the coordinates of these points.

Definition D.1. *An* affine space *is either the degenerate space reduced to the empty set, or a triple* $\langle E, \overrightarrow{E}, +\rangle$ *consisting of a non-empty set* E *(of* points*), a vector space* $\overrightarrow{E}$ *(of* translations*, or* free vectors*), and an action* $+ : E \times \overrightarrow{E} \to E$*, satisfying the following conditions:*

1. $a + \mathbf{0} = a$
2. $(a + \mathbf{u}) + \mathbf{v} = a + (\mathbf{u} + \mathbf{v})$*, for every* $a \in E$ *and every* $\mathbf{u}, \mathbf{v} \in \overrightarrow{E}$*.*
3. *For any two points* $a, b \in E$*, there is a unique* $\mathbf{u} \in \overrightarrow{E}$ *such that* $a + \mathbf{u} = b$*.*

This allows us to denote the unique vector $\mathbf{u} \in \overrightarrow{E}$ such that $a+\mathbf{u} = b$ as either $\overrightarrow{ab}$, or $\mathbf{ab}$, or $b - a$.

It is trivial to show that $\langle \mathbb{R}^3, \mathbb{R}^3, +\rangle$, where $+$ is the symbol of the usual addition, is an affine space. Indeed, let $a = \left[A_1, A_2, A_3\right]^T$, $b = \left[B_1, B_2, B_3\right]^T$ be two points in $\mathbb{R}^3$, and $\mathbf{u} = \left[U_1, U_2, U_3\right]^T$, $\mathbf{v} = \left[V_1, V_2, V_3\right]^T$ be two vectors in $\mathbb{R}^3$. Then we have

$$a + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = a, \tag{D.1}$$

$$(a + \mathbf{u}) + \mathbf{v} = a + (\mathbf{u} + \mathbf{v}) \tag{D.2}$$

and for any a, b, there is a unique free vector $\overrightarrow{ab} = \left[B_1 - A_1, B_2 - A_2, B_3 - A_3\right]^T$ such that

$$b = a + \overrightarrow{ab}. \tag{D.3}$$

D.1.2 Chasles's identity

Given any three points $a, b, c \in E$, since $c = a + \overrightarrow{ac}$, $b = a + \overrightarrow{ab}$, and $c = b + \overrightarrow{bc}$, we have

$$c = b + \overrightarrow{bc} = (a + \overrightarrow{ab}) + \overrightarrow{bc} = a + (\overrightarrow{ab} + \overrightarrow{bc}), \tag{D.4}$$

and thus

$$\overrightarrow{ab} + \overrightarrow{bc} = \overrightarrow{ac}, \tag{D.5}$$

which is known as Chasles's identity.

D.1.3 Affine combinations

Care must be taken when writing linear combinations of points such as

$$\alpha_1 a_1 + \cdots + \alpha_n a_n, \tag{D.6}$$

where $\alpha_1, \ldots, \alpha_n$ are numbers and $a_1, \ldots, a_n$ are points. In fact, we are restricted to using coefficients which sum to one. The interested reader will find the proof of the two next results in Gallier [2011].

Lemma D.1. *If $\alpha_1 + \cdots + \alpha_n = 0$, then the vector in $\overrightarrow{E}$*

$$\alpha_1 \overrightarrow{sa_1} + \cdots + \alpha_n \overrightarrow{sa_n} \tag{D.7}$$

does not depend on the point s.

Lemma D.2. *If $\alpha_1 + \cdots + \alpha_n = 1$, then the point in E*

$$s + \alpha_1 \overrightarrow{sa_1} + \cdots + \alpha_n \overrightarrow{sa_n} \tag{D.8}$$

does not depend on the point s. It is called the barycenter *(or* affine combination*) of the points a_i, assigned the weights α_i.*

D.1.4 Lines and segments

Given two numbers $A < B$, any number X is an affine combination $(1-\lambda)A + \lambda B$ of A and B and is uniquely written in this form. Indeed, since $A \neq B$, the equation $(1-\lambda)A + \lambda B = X$ admits

$$\lambda = \frac{X - A}{B - A} \tag{D.9}$$

as unique solution. In addition, $\lambda < 0$ implies $X < A$, $0 \leq \lambda \leq 1$ implies $A \leq X \leq B$, and $\lambda > 1$ implies $X > B$. This is illustrated in Figure D.1.

By analogy, we call a *line* in E the set of all affine combinations of two distinct points in E (Figure D.1). If a and b are distinct, we note ab the line passing

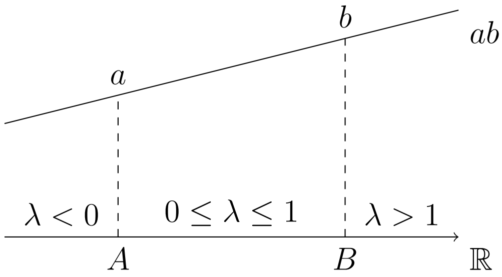


Figure D.1: The real line as a model for all lines.

through a and b, that is

$$ab = \{(1 - \lambda)a + \lambda b \mid \lambda \in \mathbb{R}\} . \tag{D.10}$$

We also define the *segment of line* joining A and B as the set

$$[ab] = \{(1 - \lambda)a + \lambda b \mid 0 \leq \lambda \leq 1\} . \tag{D.11}$$

D.1.5 Homogeneous coordinates

In affine geometry, it is not possible to express the notion of length of a segment of line or orthogonality of vectors. This is why we add an *inner product* to the vector space $\overrightarrow{E}$ to obtain a Euclidian affine space. For two vectors $\mathbf{u} = \left[U_1, U_2, U_3\right]^T$, $\mathbf{v} = \left[V_1, V_2, V_3\right]^T$ in $\mathbb{R}^3$, the inner product is defined as

$$\mathbf{u} \cdot \mathbf{v} = U_1V_1 + U_2V_2 + U_3V_3. \tag{D.12}$$

The result of the inner product of two vectors is a scalar. The inner product of a vector with itself is always non-negative. This allows us to define the length of a vector as

$$||\mathbf{x}|| = \sqrt{\mathbf{x} \cdot \mathbf{x}}. \tag{D.13}$$

Still, it would be advantageous to mathematically treat points and vectors as if they lived in the same world. This is achieved using a “coding trick”. Points and vectors in the 3D space are distinguished by a new fourth coordinate, which is one for points, and zero for vectors. This way of doing actually has firm mathematical grounds, which are described in Chapter 4 of Gallier [2011]. We use the same bold and lowercase notation to denote points and vectors in homogeneous coordinates.

For example, let X be a point in the Euclidian space of dimension three:

$$x = \left[X, Y, Z\right]^T . \tag{D.14}$$

Then, its homogeneous coordinates are

$$\mathbf{x} = \left[X, Y, Z, 1\right]^T . \tag{D.15}$$

These coordinates are defined up to scale, that is

$$\left[X, Y, Z, 1\right]^T = \left[\lambda X, \lambda Y, \lambda Z, \lambda\right]^T . \tag{D.16}$$

The inverse operation that gives euclidian coordinates from homogeneous coordinates is

$$\left[X, Y, Z, \lambda\right]^T \rightarrow \left[X/\lambda, Y/\lambda, Z/\lambda\right]^T . \tag{D.17}$$

Homogeneous coordinates also allow to express a translation as a matrix product [Hartley and Zisserman, 2004]. The translation of vector $\mathbf{t}$ can be expressed as

$$\mathbf{x}' = \left[\begin{array}{c|c} \mathbf{I}_3 & \mathbf{t} \\ \hline \mathbf{0}^T & 1 \end{array}\right] \mathbf{x} \tag{D.18}$$

where $\mathbf{I}_3$ is the identity matrix of dimension three.

D.2 Representation and manipulation of 3D object coordinates

To any object in the Euclidian 3D space, we attach a coordinate system called the body-fixed coordinate system. Then, the pose of the object, which consists in the position and the attitude of the object, is defined by the origin of the body-fixed coordinate system and the rotation matrix linking the world coordinate system to the body-fixed coordinate system.

Consider two coordinate systems $(A) = (o_A, \mathbf{i}_A, \mathbf{j}_A, \mathbf{k}_A)$ (say, the world coordinate system) and $(B) = (o_B, \mathbf{i}_B, \mathbf{j}_B, \mathbf{k}_B)$ (say, the body-fixed coordinate system) that differ by a rigid transformation, i.e. one is translated and rotated relative to the other. Then the coordinates of the point ${}^Bx$ expressed in (B) can be derived from the coordinates of ${}^{A}x$ expressed in (A) through [Forsyth and Ponce, 2003]

$$ {}^{B}x = {}^{B}_{A}\mathbf{R}\,{}^{A}x + {}^{B}o_{A}, \tag{D.19} $$

where the matrix ${}^{B}_{A}\mathbf{R}$ describes the rotation from (A) to (B) and is by definition given by

$$ {}^{B}_{A}\mathbf{R} = \begin{bmatrix} \mathbf{i}_A \cdot \mathbf{i}_B & \mathbf{j}_A \cdot \mathbf{i}_B & \mathbf{k}_A \cdot \mathbf{i}_B \\ \mathbf{i}_A \cdot \mathbf{j}_B & \mathbf{j}_A \cdot \mathbf{j}_B & \mathbf{k}_A \cdot \mathbf{j}_B \\ \mathbf{i}_A \cdot \mathbf{k}_B & \mathbf{j}_A \cdot \mathbf{k}_B & \mathbf{k}_A \cdot \mathbf{k}_B \end{bmatrix}. \tag{D.20} $$

Because ${}^{B}o_{A} = -{}^{A}o_{B}$, Equation (D.19) can equivalently be expressed as

$$ {}^{B}x = {}^{B}_{A}\mathbf{R}({}^{A}x - {}^{A}o_{B}). \tag{D.21} $$

This yields, if $x = o_A$,

$$ {}^{B}o_{A} = -{}^{B}_{A}\mathbf{R}\,{}^{A}o_{B} \tag{D.22} $$

because ${}^{A}o_{A} = \left[0, 0, 0\right]^{T}$.

In homogeneous coordinates, the whole mapping is expressed as a matrix product.

$$ \begin{bmatrix} {}^{B}x \\ 1 \end{bmatrix} = \left[\begin{array}{c|c} {}^{B}_{A}\mathbf{R} & {}^{B}o_{A} \\ \hline \mathbf{0}^{T} & 1 \end{array}\right] \begin{bmatrix} {}^{A}x \\ 1 \end{bmatrix} \tag{D.23} $$

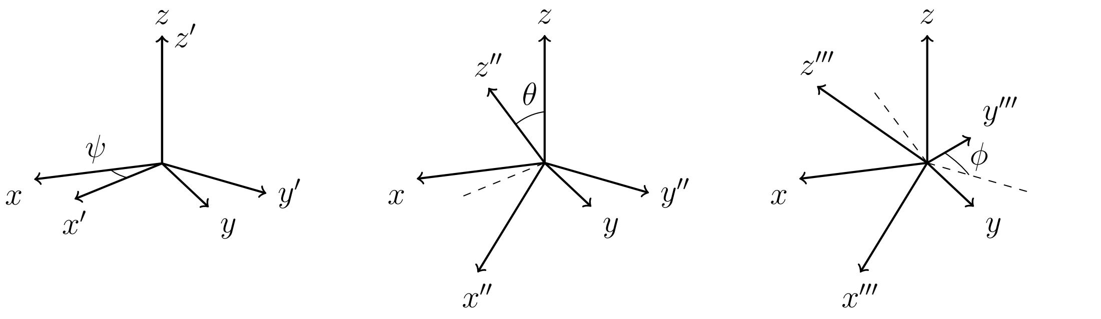
$$ {}^{B}\mathbf{x} = \left[\begin{array}{c|c} {}^{B}_{A}\mathbf{R} & -{}^{B}_{A}\mathbf{R}\;{}^{A}o_{B} \\ \hline \mathbf{0}^{T} & 1 \end{array}\right] {}^{A}\mathbf{x}. \tag{D.24} $$

The matrix ${}^{B}_{A}\mathbf{R}$ can be further decomposed into a product of three matrices corresponding to three rotations, each about a different single coordinate axis [Diebel, 2006]. The triplet by angles defining the rotation is called the Euler angle vector. In **Blender**, the Euler angles $\left[\phi, \theta, \psi\right]$ (specified by the user through $\left[\texttt{rotX}, \texttt{rotY}, \texttt{rotZ}\right]$) that define any object orientation in space, are by default associated with the sequence (X, Y, Z). This means that $\mathbf{R}$ is decomposed in

$$ \mathbf{R} = \mathbf{R}_X(\phi)\mathbf{R}_Y(\theta)\mathbf{R}_Z(\psi) \tag{D.25} $$

$$ = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & \sin\phi \\ 0 & -\sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{D.26} $$

where $\mathbf{R}_X(\phi)$ corresponds to a rotation by angle ϕ about the X-axis, $\mathbf{R}_Y(\theta)$ to a rotation by angle θ about the Y-axis, and $\mathbf{R}_Z(\psi)$ to a rotation by angle ψ about the Z-axis. The detail of each rotation is illustrated in Figure D.2.



(a) Rotation by angle ψ about the z-axis.

(b) Rotation by angle θ about the y'-axis.

(c) Rotation by angle ϕ about the x''-axis.

Figure D.2: The sequence (X, Y, Z) of Euler angles.

Bibliography

Diebel, J., 2006. Representing Attitude: Euler Angles, Unit Quaternions, and Rotation Vectors. 234

Forsyth, D., Ponce, J., 2003. Computer vision: a modern approach. Prentice Hall. 234

Gallier, J., 2011. Geometric Methods and Applications - For Computer Science and Engineering, 2nd Edition. No. 38 in Texts in Applied Mathematics. 230, 231, 232

Hartley, R., Zisserman, A., 2004. Multiple View Geometry in Computer Vision, 2nd Edition. Cambridge University Press. 233

This page intentionally left blank.

Appendix

E

Post-session presence questionnaire

Highlights

- ✓ The questionnaire participants answered in Chapter 7 is given.
- ✓ An example of the presentation a question is given.
- ✓ The list of questions in the Temple Presence Inventory is given.
- ✓ The list of questions in the Swedish viewer-user presence questionnaire is given.
- ✓ The list of questions taken from Bouvier's PhD thesis is given.
- ✓ The lists of questions related to the overall quality of the experience, and to demographics are given.

Contents

In this appendix, the questionnaire participants answered in Chapter 7 is given. First, an example of the presentation of a question is given in Section E.1. Then, the following lists of questions are given: (1) the Temple Presence Inventory in Section E.2, (2) the Swedish viewer-user presence questionnaire in Section E.3, (3) the questions taken from Bouvier's PhD thesis in Section E.4, and (4) the questions related to the overall quality of the experience, and to demographics in Section E.5.

E.1 Presentation of the questions

The questionnaire was presented using seven-item scales with anchors at the endpoints, i.e only the extreme values were verbalized. All bullets are associated to a numerical label, from 1 to 7.

The questions were translated to french.

The answer to a question starting with "To what extent . . . " is given on an intensity scale, which reads

1. Not at all
2. Hardly
3. Slightly
4. Medium
5. Quite a bit
6. Considerably
7. Extremely

which we presented as

not at all	○	○	○	○	○	○	○	extremely
	1	2	3	4	5	6	7	

The answer to a question starting with "How often . . . " is given on a time or frequency scale, which reads

1. Never
2. Rarely
3. Occasionally
4. Sometimes
5. Frequently
6. Very often
7. Every time

which we presented as

never	○	○	○	○	○	○	○	always
	1	2	3	4	5	6	7	

We now give the list of question, in english. Unless otherwise mentioned, the question is answered on an intensity scale.

E.2 Temple Presence Inventory

E.2.1 Spatial presence

PLACE How much did it seem as if the objects and people you saw/heard had come to the place you were?

TOUCH How much did it seem as if you could reach out and touch the objects or people you saw/heard?

OBJECT How often when an object seemed to be headed toward you did you want to move to get out of its way? (frequency scale)

BETHERE To what extent did you experience a sense of being there inside the environment you saw/heard?

LOCALSND To what extent did it seem that sounds came from specific different locations?

TOUCHSMG How often did you want to or try to touch something you saw/heard? (frequency scale)

WINDOW Did the experience seem more like looking at the events/people on a movie screen or more like looking at the events/people through a window? (1: window, 7: screen)

E.2.2 Social presence - actor within medium

PPLSEEU How often did you have the sensation that people you saw/heard could also see/hear you? (frequency scale)

INTERACT To what extent did you feel you could interact with the person or people you saw/heard?

LEFTPLCE How much did it seem as if you and the people you saw/heard both left the places where you were and went to a new place?

TOGETHER How much did it seem as if you and the people you saw/heard were together in the same place?

TALKTOYU How often did it feel as if someone you saw/heard in the environment was talking directly to you? (frequency scale)

EYECONT How often did you want to or did you make eye-contact with someone you saw/heard? (frequency scale)

CONTRINT Seeing and hearing a person through a medium constitutes an interaction with him or her. How much control over the interaction with the person or people you saw/heard did you feel you had?

E.2.3 Social presence - passive interpersonal

FACEEXPR During the media experience how well were you able to observe the facial expressions of the people you saw/heard?

TONEVOIC During the media experience how well were you able to observe the changes in tone of voice of the people you saw/heard?

STYLDRES During the media experience how well were you able to observe the style of dress of the people you saw/heard?

BODYLANG During the media experience how well were you able to observe the body language of the people you saw/heard?

E.2.4 Social presence - active interpersonal

MKSOUND How often did you make a sound out loud (e.g. laugh or speak) in response to someone you saw/heard in the media environment? (frequency scale)

SMILE How often did you smile in response to someone you saw/heard in the media environment? (frequency scale)

SPEAK How often did you want to or did you speak to a person you saw/heard in the media environment? (frequency scale)

E.2.5 Engagement (mental immersion)

MENTALIM To what extent did you feel mentally immersed in the experience?

INVOLVNG How involving was the experience?

SENSEENG How completely were your senses engaged?

SENSREAL To what extent did you experience a sensation of reality?

EXCITING How relaxing or exciting was the experience? (1: very relaxing, 7: very exciting)

ENGSTORY How engaging was the story?

E.2.6 Social richness

No propositions were made in this category. The participant had to choose the number that best described his/her evaluation of the media experience, using only anchors.

REMOTE 1: Remote, 7: Immediate

UNEMOTNL 1: Unemotional, 7: Emotional

UNRESPON 1: Unresponsive, 7: Responsive

DEAD 1: Dead, 7: Lively

IMPERSNL 1: Impersonal, 7: Personal

INSENSTV 1: Insensitive, 7: Sensitive

UNSOCBLE 1: Unsociable, 7: Sociable

E.2.7 Social realism

WOULDOCR The events I saw/heard would occur in the real world. (1: completely disagree, 7: completely agree)

COULDOCR The events I saw/heard could occur in the real world. (1: completely disagree, 7: completely agree)

OCRWORLD The way in which the events I saw/heard occurred is a lot like the way they occur in the real world.

E.2.8 Perceptual realism

FEELLIKE Overall, how much did touching the things and people in the environment you saw/heard feel like it would if you had experienced them directly?

TEMPERAT Overall, how much did the heat or coolness (temperature) of the environment you saw/heard feel like it would if you had experienced it directly?

SMELLIKE Overall, how much did the things and people in the environment you saw/heard smell like they would had you experienced them directly?

LOOKLIKE Overall, how much did the things and people in the environment you saw/heard look they would if you had experience them directly?

SOUNDLKE Overall, how much did the things and people in the environment you saw/heard sound like they would if you had experienced them directly?

E.3 Swedish viewer-user presence questionnaire

We only included the questions related to sound, using the same intensity scale as previously.

SNDIDENT To what extent were you able to identify sounds?

SNDLOCAL To what extent were you able to localize sounds?

SNDREAL To what extent did you think that the sound contributed to the overall realism?

E.4 Bouvier's PhD thesis

E.4.1 Judgement of emotions

E.4.1.1 Pleasure of the experience

PLEASDIM My pleasure in watching the movie quickly diminished. [inverted score]

LASTLONG I wished the experience would last longer.

MOVPLEAS The movie itself was enjoyable.

DISAPEND At the end of the experience, I was disappointed it had ended.

RECOMEXP I will recommend the experiment to my friends.

PLEASWAT I had much pleasure watching the movie.

E.4.1.2 Intensity of the experienced emotions

EMSTIMAG The emotions I felt were almost as strong as if the situation came from my imagination.

EXPSTRES Although I knew the events were virtual, I found myself feeling the same emotions as the characters.

EMSAREAL My emotional response was the same as that which would have been, had the situation been real.

EMSTREAL The emotions I felt were almost as strong as if the situation was real.

EMSAIMAG My emotional response was the same as that which would have been, had I imagined the situation.

E.4.2 Judgment of absorption and immersion

DURATION The movie lasted two and a half minute. The experience seemed to last . . . (1: extremely longer → 7: extremely shorter)

CONSCENV I was conscious of the real surrounding world as I was watching the movie (for example: noise, room temperature, . . .). [inverted score]

PERSTHOU My attention was drawn more on the virtual world than my thoughts (personal concerns, dreaming, . . .)

SUBMVIRT During the experience, the virtual environment overwhelmed you.

REALLABO During the experiment, I could recall I was actually in a laboratory (frequency scale)

E.4.3 Judgment of negative effects

NEGDESOR At the end of the experience, I felt disoriented.

NEGEYES At the end of the experience, my eyes were tired.

NEGHEAD At the end of the experience, I head a headache.

NEGVERT At the end of the experience, I suffered from vertigo.

NEGTIRED At the end of the experience, I felt tired.

NEGNAUSE At the end of the experience, I felt nauseous.

E.5 Additional questions

E.5.1 Overall quality

SEENBEFORE Had you already watched the media which was shown to you before?

PERSRELEV To what extent did you feel concerned by the media?

QUALIMAG How would you rate the image quality of the experience?

QUALSND How would you rate the sound quality of the experience?

QUALSEAT How comfortable was your position during the movie?

QUALOVERALL Overall, how would you rate the quality of the experience?

FREETEXT Use the following text box to give us comments on the experiment in general.

E.5.2 Demographics

AGE How old are you?

GENDER Please indicate your gender.

HOURSTV How many hours do you spend watching television (DVB, DVD, BluRay, . . .) in a typical day? Give an estimation as precise as possible.

SIZETV What size is the screen set you most often watch?

USEVIDEOGAME How often do you play videogames (at home, at work, or at an arcade)?

USERV How many times have you used an interactive virtual reality system?

KNOWBROAD How much do you know about broadcast or film production?

This page intentionally left blank.

Appendix

F

Analysis of variance and multiple imputation

Highlights

- ✓ A method for obtaining point and variance estimates is given.
- ✓ A method for combining results from ANOVAs is given.
- ✓ The statistical methods used in Chapter 7 to combine quantities from multiply imputed datasets are given.

Contents

Often, a dataset to be analyzed contains missing items, for example because one participant forgot to answer one of the questions. Rather than deleting the entire data of the participant or including educated guesses in the dataset, multiple imputation can be used:

> Multiple imputation involves imputing m values for each missing item and creating m completed data sets. Across these completed data sets, the observed values are the same, but the missing values are filled in with different imputations to reflect uncertainty levels. (King et al. [2001])

Multiple imputation assumes that the complete data is multivariate normal. It also assumes that the missing values are missing at random (MAR), meaning that the pattern of missingness only depends on the observed data, not the unobserved data.

F.1 Point and variance estimates

In this section, a method combining the results from several (m) imputed datasets is given. A point estimate can be a mean, a median, ...

We can compute the point and variance estimates of the quantity Q from each imputed dataset [King et al., 2001]. If $\hat{Q}_i$ and $\hat{U}_i$ are the point and variance estimates for the i^{th} dataset, then the mean of the point estimates, i.e.

$$\bar{Q} = \frac{1}{m}\sum_i \hat{Q}_i, \tag{F.1}$$

is the point estimate for Q, and the associated variance estimate is

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B \tag{F.2}$$

where

$$\bar{U} = \frac{1}{m}\sum_i \hat{U}_i \tag{F.3}$$

is the within-imputation variance and

$$B = \frac{1}{m-1}\sum_i (\hat{Q}_i - \bar{Q})^2 \tag{F.4}$$

is the between-imputation variance. The square root of T is the overall standard error associated with $\bar{Q}$.

F.2 Analysis of variance

The analysis of variance (ANOVA) is a general statistical method used to compare group means and test for a variation with a common dependent variable. ANOVA is used many times in this thesis. Therefore, a method is needed that combines the results of ANOVAs carried out on the imputed datasets. Before introducing this method, several quantities of interest are defined and a simple example of ANOVA is given in details.

F.2.1 Background

We give here the definition of several statistical quantities of interest.

The sum of squares (SS) of the values $x_1, \dots, x_n$ is the quantity

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{F.5}$$

where $\bar{x}$ is the mean of $x_1, \dots, x_n$, and each difference $(x_i - \bar{x})$ is called a *deviation*. The SS grows with n, and therefore the comparison between groups with varying sample size requires some scaling. The mean square (MS), or variance, is the quantity

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{F.6}$$

where $n-1$ is the number of *degrees of freedom* of the mean square.

F.2.2 The one-way ANOVA

In order to understand how ANOVA works, we give the example of the simple analysis of variance, where two independent mean squares are compared. In this example, n subjects were distributed amongst k groups. Each member of a given group received the same treatment, but the treatments differed across groups. The collected score for participant i in group j is noted X_{ij}. We wish to determine which model fits the data better between

$$X_{ij} = \mu + \epsilon_{ij} \tag{F.7}$$

and

$$X_{ij} = \mu + \alpha_j + \epsilon_{ij}. \tag{F.8}$$

ANOVA is based on the algebraic identity

$$X_{ij} - \bar{X} = (\bar{X}_j - \bar{X}) + (X_{ij} - \bar{X}_j), \tag{F.9}$$

which shows that the deviation between any score X_{ij} and the grand mean $\bar{X}$ is a result of a deviation between the group mean and the grand mean $\bar{X}_j - \bar{X}$, equal for all the members of group j, and a deviation inside the group $X_{ij} - \bar{X}_j$.

As shown in Rietveld and Hout [2005], (F.9) generalizes to a relationship between the sums of squares:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X})^2 = \sum_{j=1}^{k} n_j(\bar{X}_j-\bar{X})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j)^2 \tag{F.10}$$

when there are n_j participants in each group. The total SS can therefore be partitioned in a between-groups SS and a within-groups SS. The MSs are obtained by dividing the SSs by their number of degrees of freedom. Note that the number of degrees of freedom follow a similar relationship:

$$n-1 = (k-1)+(n-k). \tag{F.11}$$

The statistical significance is tested by comparing the within-groups MS (or variance) and between-groups MS (or variance). The ratio of these two quantities

$$F = \frac{\text{between-groups variance}}{\text{within-groups variance}} = \frac{\sigma_{\text{between}}}{\sigma_{\text{within}}} \tag{F.12}$$

follows an F-distribution with $k-1$ and $n-k$ degrees of freedom. The expected value of F is $1+n\sigma_{\text{between}}/\sigma_{\text{within}}$, when $n_j = n$ for all j. Under the null hypothesis that there is no effect of the treatment, the expected value of F is exactly 1. If the null hypothesis is rejected, the expected value of F is greater than 1.

F.2.3 ANOVAs and multiple imputation

We give in this section the method developed by Raghunathan and Dong [2011] to obtain F-statistics and their associated p-value when conducting ANOVAs on a multiply imputed dataset.

Suppose that the analysis is based on $l = 1,\ldots,m$ datasets. We denote the elements of the numerator in (F.12) by a subscript N. Similarly, the elements of the denominator are indicated by a subscript D. Then, the following quantities are defined,

$$A_N = \frac{1}{m}\sum_l \frac{1}{\mathrm{SS}_N^{(l)}} \tag{F.13}$$

$$B_N = \frac{1}{m}\sum_l \frac{1}{\left(\mathrm{SS}_N^{(l)}\right)^2 \mathrm{df}_N^{(l)}} \tag{F.14}$$

$$C_N = \frac{1}{m-1}\sum_l \left(\frac{1}{\left(\mathrm{SS}_N^{(l)}\right)^2} - A_N\right)^2, \tag{F.15}$$

and similarly A_D, B_D, and C_D for the denominator. Then, the F-statistic is

$$F_{\mathrm{MI}} = \frac{A_D}{A_N} \tag{F.16}$$

with the number of degrees of freedom (r_N, r_D) given by

$$r_N = 2A_N^2 \left(2B_N + \frac{m+1}{m} C_N\right) \tag{F.17}$$

$$r_D = 2A_D^2 \left(2B_D + \frac{m+1}{m} C_D\right). \tag{F.18}$$

Bibliography

King, G., Honaker, J., Joseph, A., Scheve, K., 2001. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. Am. Polit. Sci. Rev. 95, 49–69. 247, 248

Raghunathan, T., Dong, Q., 2011. Analysis of variance from multiply imputed data sets. Tech. rep., Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI. 250

Rietveld, T., Hout, R. V., Jan. 2005. Statistics in Language Research: Analysis of Variance. Walter de Gruyter GmbH & Company KG. 249

This page intentionally left blank.

Appendix
G

The chi–squared test for differences between proportions

Highlights

- ✓ The statistical procedure used in Section 8.3.3 is detailed.
- ✓ The χ^2 test for homogeneity is described.

We describe here the χ^2 test for homogeneity [Berenson et al., 2012] for the dataset in Section 8.3.3. The test compares c (two or more) different independent groups, or *populations* (corresponding to `AVangle` in our work) on a binary outcome (yes or no).

The problem is usually presented in a $2 \times c$ *contingency table*. The statistic compares the given contingency table (Table G.1(a)) to that under the null hypothesis, where the proportions are equal (Table G.1(b)).

	Group 1	Group 2
yes	X_1	X_2
no	$n_1 - X_1$	$n_2 - X_2$

(a) A 2×2 contingency table.

	Group 1	Group 2
yes	πn_1	πn_2
no	$(1 - \pi)n_1$	$(1 - \pi)n_2$

(b) A 2×2 contingency table under the null hypothesis of the χ^2 test.

Table G.1: 2×2 contingency tables.

The statistic to be computed is denoted by χ^2 and is defined as

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}, \tag{G.1}$$

where, for a given cell in the table, f_o is the observed frequency, and f_e is the expected frequency. The observed frequency f_o in the cell j is the ratio between the count in the cell, either X_j or $n_j - X_j$, and the sample size n_j of the group. The expected frequency in the cell j would be the observed frequency, were the null hypothesis be true, that is, were the proportions of the different groups equal. It is computed by multiplying the global proportion $\bar{p}$

$$\bar{p} = \frac{X_1 + \cdots + X_c}{n_1 + \cdots + n_c} \tag{G.2}$$

by the sample size for cells pertaining to success (yes),

$$f_{e_j} = \bar{p} n_j, \tag{G.3}$$

and by multiplying $1 - \bar{p}$ by the sample size for cells pertaining to failure (no),

$$f_{e_j} = (1 - \bar{p}) n_j. \tag{G.4}$$

The statistic χ^2, defined in Equation (G.1), approximately follows a χ^2 distribution with $c - 1$ degrees of freedom. In the statistical software **R**, the statistic is computed with the function `chisq.test`.

The χ^2 approximation requires that all expected frequencies must be large. In practice, a common criteria states that at least 80% of the cells should have an expected frequency greater than 5 and that no cell should have an expected frequency less than 1.

When the statistic χ^2 is not significant, there is no statistical difference between the proportions of answers yes or no in each group. This is the null hypothesis:

H_0: the proportions in each group are the same.

The alternative hypothesis, H_1, is that not all proportions are equal:

H_1: the proportions in each group are different.

Under H_0, the proportions in each group vary only by chance and can be collapsed into the global proportion $\bar{p}$ given in (G.2).

Under H_1, one can only reach the conclusion that some proportions differ. A multiple comparisons procedure is needed to determine which groups have differing proportions. The Marascuilo procedure is one such procedure [Marascuilo, 1966]. The Marascuilo procedure consists in computing the observed absolute differences $|f_{e_i} - f_{e_j}|$ $(i \neq j)$ for all $c(c-1)/2$ possible pairs of groups. Then, each observed difference is compared to its critical range CR

$$\mathrm{CR} = \sqrt{\chi^2_U}\sqrt{\frac{f_{e_i}(1 - f_{e_i})}{n_i} + \frac{f_{e_j}(1 - f_{e_j})}{n_j}}, \tag{G.5}$$

where χ^2_U is the upper-tail critical value of the χ^2 distribution at a given significance level p with $c-1$ degrees of freedom. A given pair of groups has significantly differing proportions if the observed difference $|f_{e_i} - f_{e_j}|$ is greater than its corresponding critical range.

Bibliography

Berenson, M. L., Levine, D. ., Krehbiel, T. C., 2012. Basic Business Statistics, 12th Edition. Prentice Hall. 253

Marascuilo, L. A., 1966. Large-sample multiple comparisons. Psychol. Bull. 65 (5), 280–290. 254

This page intentionally left blank.

List of Figures

List of Tables