UNIVERSITÉ PARIS 13

THÈSE

pour obtenir le grade de

DOCTEUR DE l'UNVERSITÉ PARIS 13,
SORBONNE PARIS CITÉ

Discipline : **Signaux et Images**

présentée et soutenue publiquement

par

Abderrahmane SMIMITE

le 25 Juin, 2014

Titre:

Immersive 3D sound Optimization, Transport and Quality Assessment

Optimisation du son 3D immersif, Qualité et Transmission

Jury :

Pr.	Azeddine BEGHDADI	L2TI, Directeur de thèse
Pr.	Ken CHEN	L2TI, Co-Directeur de thèse
Dr.	Abdelhakim SAADANE	Polytech Nantes, Rapporteur
Pr.	Hossam AFIFI	Telecom SudParis, Rapporteur
Pr.	Younes BENNANI	LIPN-CNRS, Examinateur
Mr.	Pascal CHEDEVILLE	Digital Media Solutions, Encadrant

Declaration of Authorship

I, Abderrahmane SMIMITE, declare that this thesis titled, 'Immersive 3D sound Optimization, Transport and Quality Assessment' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed:

Date:

Résumé

Cette étude porte sur trois problématiques reliés à l'utilisation du son multicanal 3D dans le contexte des applications Audio Professionnelles.

Le système *SIRIUS* est présenté. Il s'agit d'une technique de transport du son multicanal qui répond aux contraintes de *Fiabilité*, *Synchronisation* et *Latence* des applications professionnelles et garantie un compromis entre ces différents aspects. Le système peut fonctionner sur les infrastructures LAN classiques et coexister avec d'autres types de trafic réseau. Il est aussi basé sur une couche protocolaire n'utilisant que des protocoles standards, ce qui lui procure un certain niveau d'interopérabilité avec des technologies équivalentes.

La seconde contribution est la méthodologie *AQUA*. Il s'agit d'une nouvelle approche pour l'évaluation de la qualité du son multicanal qui propose des outils efficaces pour l'analyse subjective et objective de la qualité. La partie subjective consiste en un nouveau protocol pour les tests d'écoute qui combine l'analyse de l'information perceptive et spatiale. La précision de la localisation est évaluée grâce au suivi des gestes des auditeurs. Notre méthode, basée sur l'utilisation de la Kinect, permet d'obtenir cette information d'une façon rapide et précise. Le protocole utilise notamment l'analyse EEG pour étudier les biais psychologiques et filtrer efficacement les sujets. La partie objective repose sur un moteur binaural qui convertit le flux multicanal en un flux stereo binaural plus simple à analyser et qui préserve l'information spatiale. Le signal audio résultant est analysé par un modèle perceptif et un modèle spatial qui permettent d'estimer une représentation interne équivalente. Les variables la constituant alimentent ensuite un Réseau de Neurones Artificiel qui permet d'obtenir une note objective de qualité. Parallèlement, le modèle psychologique simule le comportement humain en ajustant la note en fonction des notes précédentes. Les performances obtenues montrent que le système peut être utilisé pour prédire la qualité perceptive et spatiale du son multicanal avec un grand niveau de précision et de réalisme.

Le dernier axe d'étude porte sur l'optimisation de la qualité d'écoute dans les systèmes audio surround. Etant donné leur problème de Sweet Spot, et la complexité des systèmes suggérant de l'élargir, on propose une technique basée sur le suivi de la position réelle des auditeurs. Le suivi est réalisé d'une façon non-intrusive par l'analyse d'images thermiques. Les canaux audio initiaux sont considérés comme des sources virtuelles et sont re-mixés par VBAP pour simuler leur déplacement vers l'auditeur. Les performances obtenues montrent un suivi efficace et une amélioration de l'expérience d'écoute.

Abstract

In this work, three complementary topics regarding the use of multichannel spatial audio in professional applications have been studied.

SIRIUS, is an *audio transport* mechanism designed to convey multiple professional-grade audio channels over a regular LAN while maintaining their synchronization. The system reliability is guaranteed by using a FEC mechanism and a selective redundancy, without introducing any important network overload. The system also offers a low latency that meet the professional applications requirements and can operate on the existing infrastructures and coexist with other IT traffic. The system relies on standard protocols and offers a high level of interoperability with equivalent technologies. The overall performances satisfy Pro Audio requirements.

The second contribution is *AQUA*, a comprehensive framework for multichannel audio *quality assessment* that provides efficient tools for both subjective and objective quality evaluation. The *subjective* part consists of a new design of reliable listening tests for multichannel sound that analyze both perceptual and spatial information. Audio localization accuracy is reliably evaluated using our *gesture-based* protocol build around the Kinect. Additionally, this protocol relies on EEG signals analysis for psychological biases monitoring and efficient subjects screening. The *objective* method uses a *binaural model* to down-mix the multichannel audio signal into a 2-channels binaural mix that maintains the spatial cues and provides a simple and scalable analysis. The binaural stream is processed by a perceptual and spatial models that calculate relevant cues. Their combination is equivalent to the internal representation and allows the cognitive model to estimate an objective quality grade. In parallel, the psychological model simulate the human behavior by adjusting the output grades according to the previous ones (i.e., the *experience effect*). The overall performance shows that AQUA model can accurately predict the perceptual and spatial quality of a multichannel audio in a very realistic manner.

The third focus of the study is to optimize the listening experience in surround sound systems (*OPTIMUS*). Considering the *sweet spot* issue in these systems and the complexity of its widening, we introduce a tracking technique that virtually moves the sweet spot location to the actual position of listener(s). Our approach is non-intrusive and uses *thermal imaging* for listeners identification and tracking. The original channels are considered as virtual sources and remixed using the *VBAP* technique. Accordingly, the audio system virtually *follows* the listener actual position. For home-cinema application, the kinect can be used for the tracking part and the audio adjustment can be done using HRTFs and cross-talk cancellation filters. The system shows an improvement of the localization accuracy and the quality of the listening experience.

Acknowledgements

This work has been carried out under a CIFRE contract between Digital Media Solutions, the funding company and L2TI the academic research lab. I would like to thank everyone in DMS for this opportunity and for providing the environnement and tools that allowed this work to come to life.

I would like to extend my greatest appreciation and gratitude to my advisers at L2TI, Pr. Azeddine Beghdadi and Pr. Ken Chen, for their guidance, understanding and patience.

I am really thankful to my colleagues Walid, Pascal, Ludovic, Justin and Ouadîe for their help with the tests, for the stimulating discussions, and most importantly, for their friendship.

I also would like to thank the following gentlemen: I. Choi, H. Wierstorf, A. Bovik, M. Dietz and A. Csapo, for kindly exchanging their valuable work with me and for the chats, although small, were very inspiring and helpful.

Finally and most importantly, I would like to thank my parents, my family, my wife-to-be and my high-school english teacher for their infinite help and support.

Also Thanks to *Dr. House* and *Dr. Lightman* for the inspiration and the outside-the-box ideas ☺.

Contents

List of Figures

List of Tables

Abbreviations

AES	**A**udio **E**ngineering **S**ociety
EBU	**E**uropean **B**roadcasting **U**nion
ITU	**I**nternational **T**elecommunication **U**nion
LAN	**L**ocal **A**rea **N**etwork
IP	**I**nternet **P**rotocol
UDP	**U**ser **D**atagram **P**rotocol
RTP	**R**eal-**T**ime **P**rotocol
PTP	**P**recision **T**ime **P**rotocol
QoS	**Q**uality **o**f **S**ervice
CODEC	**CO**der-**DEC**oder
FEC	**F**orward **E**rror **C**orrection
AVB	**A**udio **V**ideo **B**ridging
OSC	**O**pen **S**ound **C**ontrol
ZeroConf	**Zero Conf**iguration
PEAQ	**P**erceptual **E**valuation of **A**udio **Q**uality
HRTF	**H**ead **R**elated **T**ransfer **F**unction
BRIR	**B**inaural **R**oom **I**mpulse **R**esponse
ITD	**I**nteraural **T**ime **D**ifference
ILD	**I**nteraural **L**evel **D**ifference
IACC	**I**nter**A**ural **C**ross-correlation **C**oefficient
MSO	**M**edial **S**uperior **O**live
LSO	**L**ateral **S**uperior **O**live
EEG	**E**lectro-**E**ncephalo-**G**raphy
ANN	**A**rtificial **N**eural **N**etwork
VBAP	**V**ector **B**ase **A**mplitude **P**anning

To my parents, my family and my wife-to-be for their infinite and unconditional love and support.

"If you stay close to nature, to its simplicity, to the small things hardly noticeable, those things can unexpectedly become great and immeasurable."

Rainer Maria Rilke
Letters to a Young Poet

Chapter 1

Introduction

1.1 Context

The past decade has seen a rapid development and deployment of audio and video technologies in many fields. The new generation of multimedia contents is richer, more realistic, and offers a whole new experience to end users where they are no longer mere spectators, but get completely immersed in the scene. They can actually become a part of the show, virtually speaking, and even capable of interacting with it through interactive virtual environments.

Thanks to the recent developments in computer sciences, electronics and signal processing, interest has been renewed in audio technologies. Like video, we speak today of *3D audio* or *Multichannel 3D audio*. This actually refers to a branch of technologies based on the combination of several virtual audio sources localized in space, which creates a listener illusion of 3D immersion in the scene. This is the natural evolution of the well-known *Surround-sound* systems (5.1, 7.1, etc.), which enabled so far, only an envelopment of the listener on a 2D plane (horizontal dimension).

These technologies are reaching more people today, mainly through the cinema industry. In fact, the synergy resulting from combining techniques that stimulate more than one sense, offers a whole new multimedia experience, in the so-called *cinema of the future*. By associating 3D video, 3D audio, mechanical (haptic) feedback or even fluids projections, the immersion sensation is total and extremely realistic. Across the globe, movie theaters and entertainment complex are equipped today with such technologies (*6D cinemas* are available in The United Kingdom, Belgium, Malaysia, ..).

For the audio part, this effect is generally obtained by using carefully arranged loudspeakers arrays, that interact to create an auditory environment or spatial sound scene. It is also possible to emulate 3D sound using only headphones, by using the *binaural* characteristics of the human auditory system, as explained later in this paper.

In the scope of our study, we essentially focus on issues associated with professional sound applications. It mainly relates to audio technologies used in cinemas and studios, but can also be found in other high performances audio applications. Generally speaking, the audio chain in this context is presented in figure 1.1:

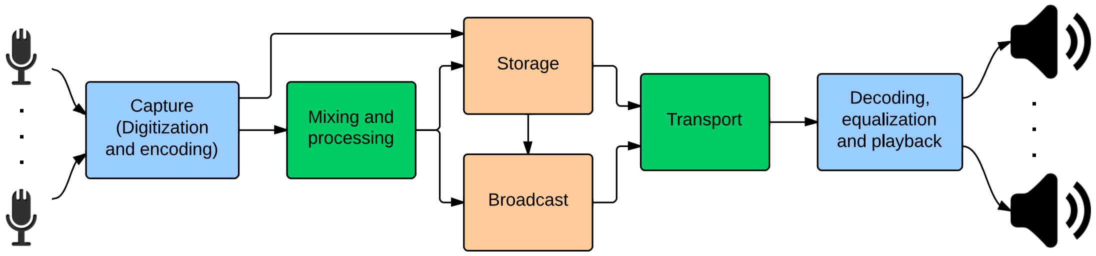

Capture (Digitization and encoding)
Mixing and processing
Storage
Broadcast
Transport
Decoding, equalization and playback

FIGURE 1.1: Generic Audio Chain

First, the analog signals emitted from the audio sources are captured, digitized and encoded. Next, they are stored and transmitted to the audio authoring and mixing system (typically a Mixing Console and a ProTools workstation). The result is either stored for later distribution or live broadcasted using specific transport mechanisms. Once the media content reaches the end user, it is decoded and processed to fit the playback system. Accordingly, we can identify two phases: *production* and *playback*.

A closer look on the audio *production* chain, focusing on the data flow of the authoring part is illustrated in figure 1.2.

A preliminary study with the assistance of industrial partners (movie theaters managers, sound engineer and technicians), regarding the use of this chain for multichannel audio, allowed us to identify the following aspects:

- The transport mechanisms are obsolete and are no longer adapted for the increasing requirements of multichannel pro audio.
- The existing audio transport technologies lack scalability, inter-compatibility and efficient management tools.
- There is no efficient way to locally monitor the quality level, not to mention the absence of spatial quality assessment tools.

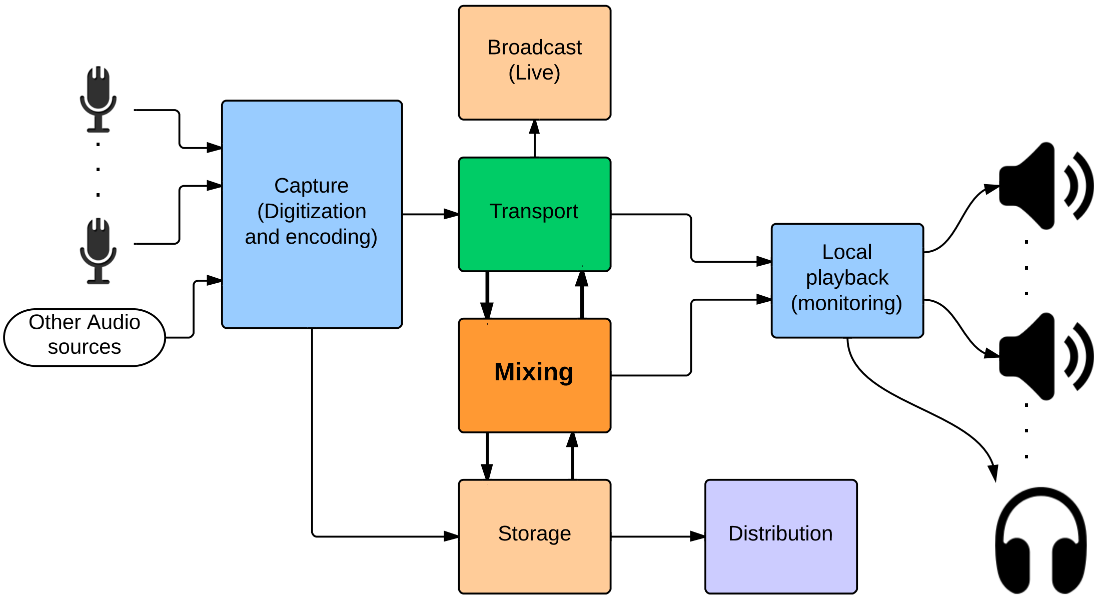


Figure 1.2: Pro Audio: Production Audio Chain

A closer look at the end-user part of the chain is presented in figure 1.3. It illustrates what is commonly known as A and B chains, which respectively refer to sound track processing and reproduction chains. At this level, the main audio source is the media server. At its outputs, and alongside the decoded video signal that goes to the projector, a multichannel audio stream [1] is transmitted to the sound processor. The latter performs additional decoding if needed and applies the room equalization (RoomEQ). The final audio stream is sent to amplifiers and then to the loudspeakers.

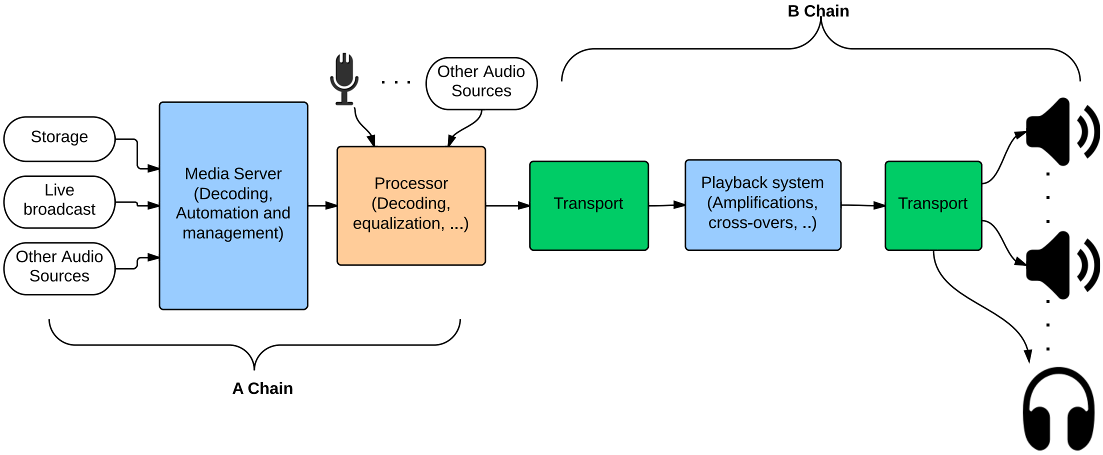

B Chain
Other Audio Sources
Storage
Live broadcast
Other Audio Sources
Media Server (Decoding, Automation and management)
Processor (Decoding, equalization, ...)
Transport
Playback system (Amplifications, cross-overs, ..)
Transport
A Chain

Figure 1.3: Pro Audio: Playback Audio Chain

[1] Live stream such as sport event or opera or decoded media file from the server hard drive

For this part, the previous study identified the following aspects:

- The usual transport techniques lack flexibility and are quite expensive.
- The lack of an efficient tool to assess the perceptual and spatial quality of sound, unless consulting experts, which is time consuming and expensive.
- Before moving to the next generation of 3D audio, most of the current setups still rely on surround sound systems, which suffer from the sweet spot limitation.

The sweet spot is the focal point between the speakers, where an individual is fully capable of hearing the audio mix the way it was intended to be heard by the mixer. Unless he is within, a listener would most likely hear the spatial mix differently.

For all these aspects, this study seeks to remedy these problems by critically analyzing the literature and suggesting new solutions to answer each of the following needs:

- How to transport multichannel audio efficiently, meet the requirement of high performance audio systems and keep the cost-effectiveness ?
- How to evaluate multichannel audio quality on a perceptual and spatial scale in a realistic and rapid manner ?
- How can we widen the sweet spot in surround sound system or circumvent this issue using a low complexity and cost-effective solution ?

These are the questions that we are answering through the next chapters, where we present our approach for Audio Transport, Quality Assessment and Suround-Sound Optimization.

This work has been conducted within a CIFRE contract[2], between *A. SMIMITE*, PhD candidate and Signal Processing Engineer, *Digital Media Solutions* (DMS), the financing company and *Laboratoire de Traitement et Transport de l'Information* (L2TI), the academic supervising lab. Accordingly, this PhD has a strong engineering orientation, as the purpose is to establish algorithms and solutions intended to be embedded in real products.

[2] www.anrt.asso.fr

1.2 Study Organization

The central thesis of this study is about multichannel audio. As presented above, we address three issues regarding this theme: Transport, Quality and Optimization. Although they might seem separate, these three axis are strongly connected as illustrated in figure 1.4. Therefore, the present work discusses three complementary components of the same chain.

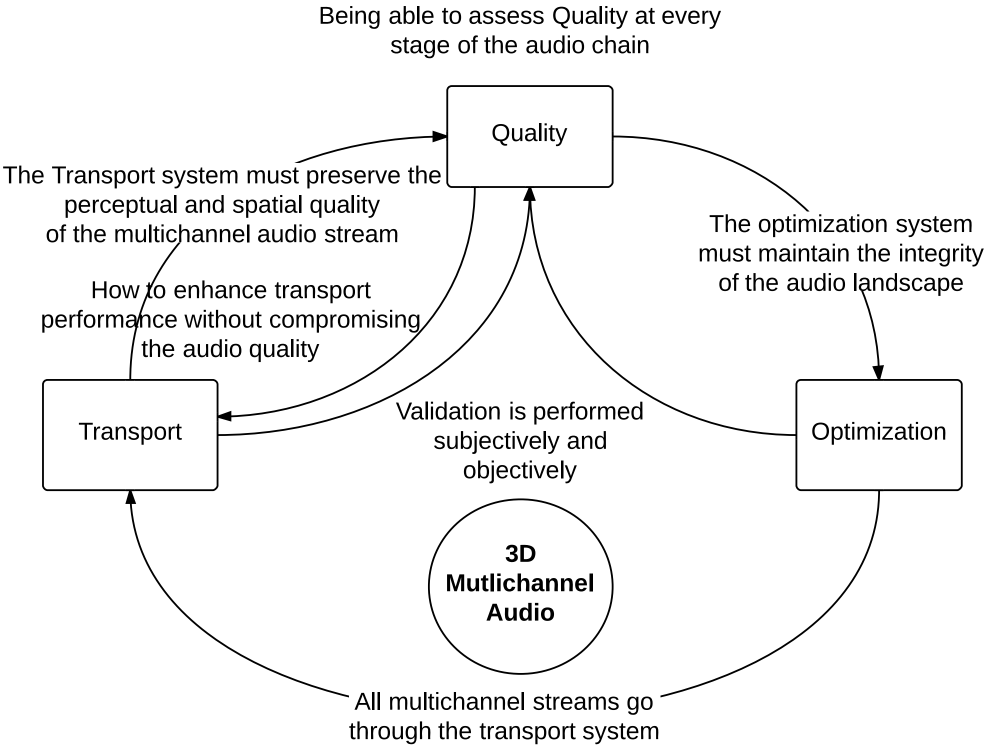

Being able to assess Quality at every stage of the audio chain
Quality
The Transport system must preserve the perceptual and spatial quality of the multichannel audio stream
The optimization system must maintain the integrity of the audio landscape
How to enhance transport performance without compromising the audio quality
Transport
Validation is performed subjectively and objectively
Optimization
3D Mutlichannel Audio
All multichannel streams go through the transport system

Figure 1.4: Thesis Organization

This study has been conducted using a multidisciplinary approach, including not only audio signal processing and computer networks, but also thermal image processing, EEG, psychology, cognitive sciences, Machine learning and Embedded systems. The main contributions are:

- A Multichannel Audio Transmission Protocol (RTP profile and extensions) based on an optimal architecture offering a trade-off of low latency, high synchronization and high reliability,
- A Multichannel Audio Quality assessment framework with tools for subjective and objective evaluation, that combine perceptual, spatial and psychological information and offer a very realistic assessment.

- An Optimization technique for surround sound system, based on tracking the listener(s) position and virtually move the sweet-spot to his or their actual location.

Also, various Software and Hardware tools have been developed during this study to achieve these goals, as detailed through this report.

1.3 DMS/L2TI

This PhD was carried out under a CIFRE contract between A. SMIMITE, DMS and L2TI.

Digital Media Solutions (DMS) is a specialized company in the design, manufacturing and marketing of 2D and 3D audio solutions, for the cinema, Hi-Fi, V.I.P Home Cinema, Television and Music industries. Thanks to DMS associates 25 years of experience and passion for audio solutions and services, DMS meets your technical needs for analog and digital audio. DMS is one of the global specialist in immersive sound, which essentially consists of virtually positioning sound in space. The basic 3D audio system consists of a 16 channels sound system, but can interact with formats from 1.1 to 23.1 and even higher orders. The technology used by DMS allows a clean conversion between stereo, 5.1 and 7.1 into 3D immersive sound. More than 20 worldwide movie theaters are equipped today with this technology, offering a unique experience at the edge of technology and the heart of action for the first time, with an unprecedented realism. With the help of a wide network of certified integrators, DMS provides its assistance from the acoustical design to the product customization. Thanks to the recent acquirement of Storm audio, DMS started a new Residential division to bring its state of the art solutions to home cinema.

More details on the website: www.dms-company.com

Le Laboratoire de Traitement et Transport de l'Information (L2TI) was founded at the university of Paris 13 in 1998, and was recognized as "Equipe d'Accueil" (EA3043) by the French Ministry of Education and Research in 1999. The L2TI main research focus is on the development of applied and theoretical research on Data Processing and Transmission. More specifically, the topics addressed by the L2TI are related to multimedia content processing and analysis, and computer networking. The L2TI is structured into 2 teams, working on these two fields of research. At L2TI there are 17 permanent senior researchers, 15 PhD students, 1 post-docs, 1 engineer and 1 administration officer. The current director is Pr. Azeddine BEGHDADI.

More details on the website: www-l2ti.univ-paris13.fr

1.4 Document Layout

The first chapter gives an overview of the thesis context, goals and organisation. The next chapters are as follow:

Chapter 2. 3D multichannel audio: This section presents the characteristics of the human auditory system and how it inspired 3D sound systems,

Chapter 3. Audio Transport (project codename, SIRIUS): This chapter examines the literature as well as the requirements for multichannel audio transport and presents our suggested approach and architecture to do so. In addition, a deployment optimization algorithm is introduced.

Chapter 4. Audio Quality Assessment (project codename, AQUA): This section presents our work to establish a new comprehensive quality assessment technique designed for multichannel audio, that can perform both perceptual and spatial analysis and integrate a psychological model as well to enhance the approach realism. In addition, a new subjective protocol is introduced to offer an efficient way to conduct more reliable listening tests in the context of spatial audio.

Chapter 5. Optimization (project codename, OPTIMUS): This chapter presents our solution to resolve the sweet spot issue in cinema and home cinema environment by virtually moving the sweet spot to the actual listener(s) position instead of the theoretical one.

Chapter 6. Conclusion and future directions: The last chapter discusses the three solutions and presents the future directions to enhance each one.

1.5 Contributions

Publications

- A. Smimite, Ken Chen, and A. Beghdadi. *Next-generation audio networking engineering for professional applications.* In Telecommunications Forum (TELFOR), 2012 20th, pages 1252–1255, 2012. doi : 10.1109/TELFOR.2012.6419443.

- Abderrahmane Smimite, Azeddine Beghdadi, and Ken Chen. *Vers une nouvelle approche de l'evaluation de la qualité audio multicanal.* In Compression et Représentation des Signaux Audiovisuels, Novembre 2013.

- Abderrahmane Smimite, Azeddine Beghdadi, and Ken Chen. *Investigating "the experience effect" in audio quality assessment.* In Telecommunications Forum (TELFOR), 2013 21st, pages 769–772, 2013. doi: 10.1109/TELFOR.2013.6716343.

- Abderrahmane Smimite, Azeddine Beghdadi, Ouadie Jafjaf, and Ken Chen. *A new approach for spatial audio quality assessment.* In 2014 International Conference on Telecommunications and Multimedia (TEMU2014), Heraklion, Greece, July 2014 (Accepted).

- Abderrahmane Smimite, Azeddine Beghdadi and Ken Chen, *Audio Quality Assessment: Recent development and suggestion for a new approach.* Submitted to Journal of Signal Processing, Elsevier, 2014.

Patents

- Patent proposal 01: Multichannel audio transport system (Système de transmission du son multicanal).

- Patent proposal 02: Multichannel Audio Quality assessment techniques (Technique d'évaluation de la qualité du son multicanal).

- Patent proposal 03: Surround sound optimization for cinema and home cinema (Technique d'optimisation du son multicanal dans les cinémas et home-cinema).

Software/Hardware

- AQUA-ST: A .Net software for subjective listening tests with a Kinect interface for sound localization assessment. An iPad version is also available.

- AQUA-O/SHIELD: A set of Matlab/R scripts for objective quality assessment of multichannel audio.

- SIRIUS-Daemon: A C-written, linux based daemon for multichannel audio transmission, successfully tested on Raspberry Pi.

- SIRIUS-Deploy: A deployment utility to assist integrators for optimal setups of audio networks, with 3D visualization.

- Optimus-Tracker: thermal images processing utility to extract listeners positions.

- Ongoing/future: Sweet Spot Mover, Web-based audio quality rating platform, DSP implementation of AQUA-O, EEG-multimedia correaltion analysis, etc.

Chapter 2

Sound Localization & 3D Multichannel Audio

Multi-channel loudspeaker systems are used to create immersive auditory displays. This chapter presents a brief overview of the most used 3D multichannel audio techniques. For a better understanding of the fundamentals of these technologies, we first present a brief description of the human auditory system and how sound localization is performed by the brain.

2.1 Sound perception

Our approach is fundamentally based on the analysis of the natural phenomena involved in the listening process and try to mimic them as close as possible. Also, and although everyone agrees on the fact that hearing is not listening, the most existing approaches focus only on the hearing part. In fact, we believe that while hearing is more of physiological process, listening is a cognitive one. As explained further, they are both involved. This section describes the characteristics of the human auditory system, how sound localization is performed and how the current 3D audio systems use these feature to create spatial sound.

2.1.1 Auditory system

The auditory system is the sensory system for hearing. It includes both the sensory organs, the ears, and the auditory parts of the sensory system. The ear function is to transmit acoustic wave energy, composed of compressions and rarefactions in longitudinal waves, into mechanical vibratory energy amplified enough to go through transduction, where the wave is then converted into electrical energy that can be passed on as neural messages to the brain. Figure 2.1 illustrates the anatomy of the human ear. The auditory system consists of two parts: the *Peripheral* system and the *Central* system.

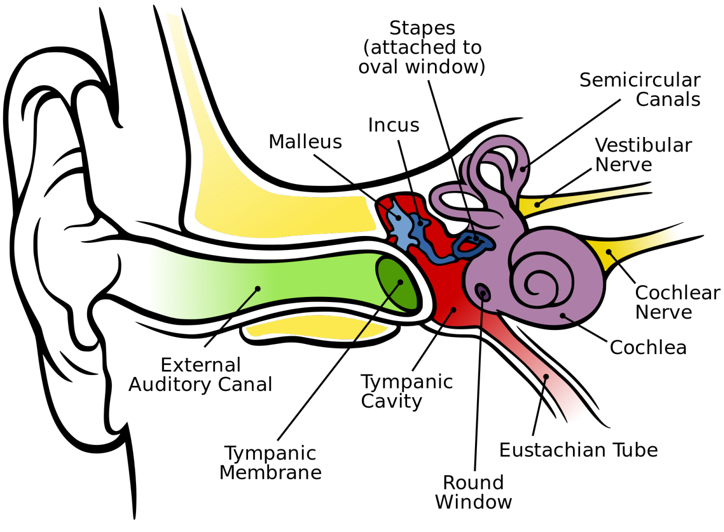


Figure 2.1: Anatomy of the human Ear. Credits: Wikimedia Commons.

The Peripheral system essentially consists of three parts: Outer ear or ear canal (green) middle ear (red) and inner ear (purple). The function of the outer ear is mainly to direct and amplify sound waves into the middle ear. The pinna (the visible part of the ear) gathers the sound energy, which is amplified by the ear canal. The middle ear is the space between the tympanic membrane and the oval window which is the entrance to the fluid-filled inner ear, and its function is to transmit the acoustic energy from air to cochlea. The inner ear consists of the cochlea and several non-auditory structures and its function is to transform the mechanical waves into electric (neural) signals [1–3]. Within the cochlea, the hair cells line the basilar membrane as illustrated in figure 2.2. The physical characteristics of the basilar membrane cause different frequencies to reach maximum amplitudes at different positions. Much as on a piano, high frequencies are at one end and low frequencies at the other. This is commonly known as *Tonotopic* Organization.

According to different theories of hearing, both the hair cells and the membrane respond to certain resonant frequencies, depending on the individual stiffness and mass of the cells and membrane at their respective locations. Resonance theories hold that the whole membrane responds differently to these multiple frequencies, consequently informing the brain of the pitch of the signal received. Place theories state that each hair cell corresponds to a certain resonant frequency, and the specific hair cell alerts the brain when it is stimulated by its frequency. Most likely a combination of the two allows the brain to determine pitch from frequency.

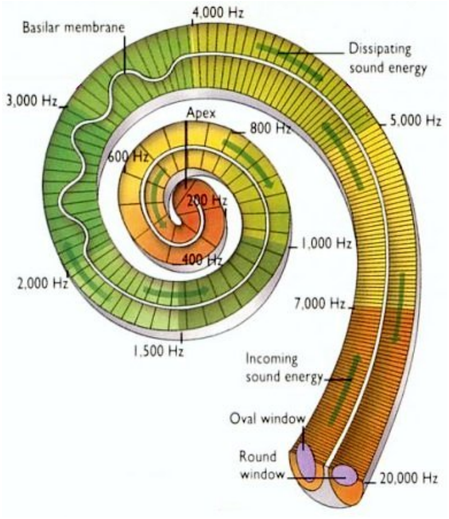


Figure 2.2: Tonotopic organization of the Human Cochlea. Credits: universe-review.ca.

The sound information, now re-encoded in form of electric signals, travels down the auditory nerve (acoustic nerve, vestibulocochlear nerve, VIIIth cranial nerve), through intermediate stations as illustrated in figure 2.3. The information eventually reaches the thalamus, and from there it is relayed to the cortex. In the human brain, the primary auditory cortex is located in the temporal lobe.

An interesting aspect to highlight is the interaction of the neural signals of the left and right ear, which is an important feature for sound localization, as explained later.

It should also be mentioned that auditory information trigger different sorts of behavioral outcomes as shown in figure 2.4. Memory and voluntary motricity responses are some of the most important ones and key features for our study, as explained in chapter 4.

2.1.2 Sound localization

Audio localization refers to the listener's ability to identify the location of a detected sound in direction and distance. It also refers to the methods in sound engineering to simulate the placement of an auditory cue in a virtual 3D space.

The auditory system uses several cues for sound source localization, including time and level differences between both ears, spectral information, timing analysis, correlation analysis, and pattern matching.

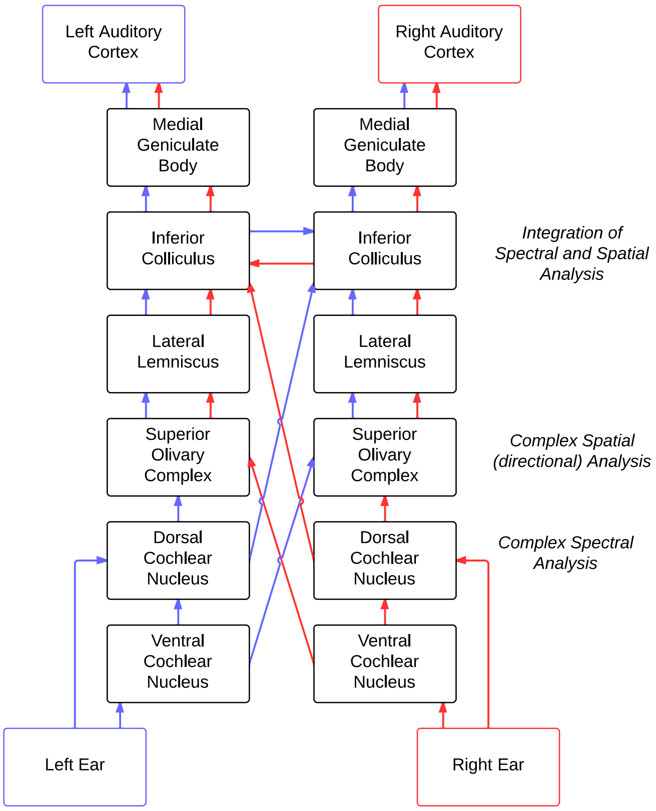


Figure 2.3: Auditory Neural Pathway and Ipsilateral and Contralateral Interactions in the Brainstem. Credits: Based on SPH307 lecture, Macquarie University.

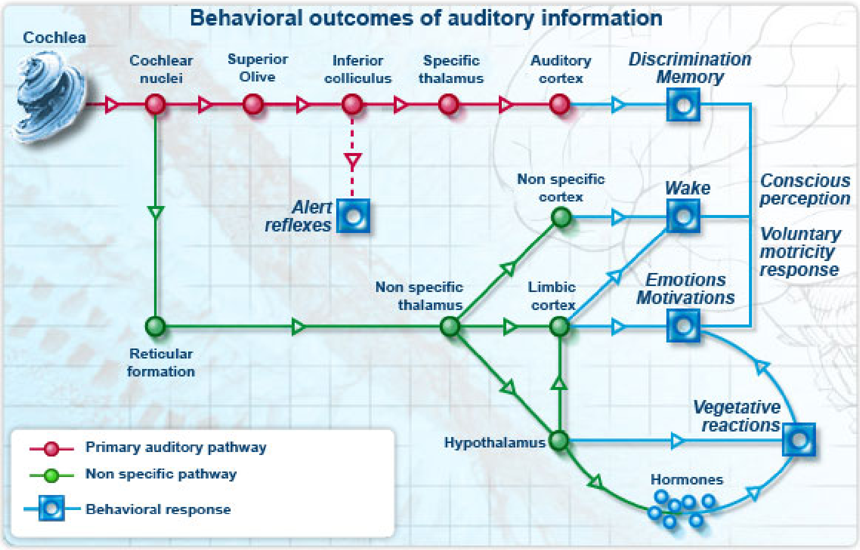
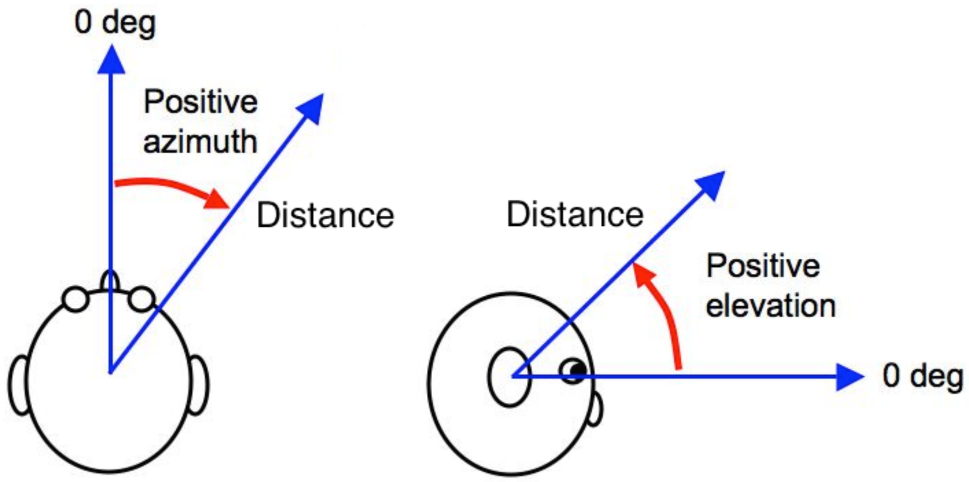


Figure 2.4: Behavioral outcomes of auditory information. Credits: cochlea.eu



(A) Azimuth Angle

(B) Elevation Angle

Figure 2.5: Spatial Coordinates for sound localization. Credits: based on [4].

The brain utilizes subtle differences in intensity, spectral, and timing cues to allow us to localize sound sources. Localization can be described in terms of three-dimensional position: the *azimuth* or horizontal angle, the *elevation* or vertical angle, and the *distance* (for static sounds) or velocity (for moving sounds) (Fig. 2.5).

Sound localization mechanism is based on two different means of analyzing the acoustic waveform. The first constitutes a spectral analysis in which the sound energy across different frequency bands arriving is compared between the two ears, and provides sound-localization abilities in the vertical dimension, including distinctions between sources to the front from those behind (Fig. 2.6). Although better performance based on frequency spectra may be possible using both ears, it essentially represents a monaural cue for sound localization, generated largely by the direction-specific attenuation of particular frequencies by the pinna and concha of the outer ear.

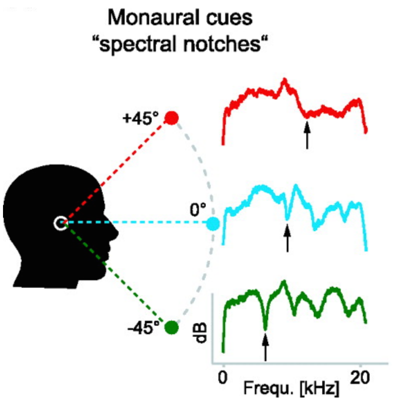


Figure 2.6: Monaural Spatial cues: Spectral Notches. Credits: [5]

The second means by which sound localization is achieved is based on detecting and comparing differences between the signals at the two eardrums. This is known as binaural computation, which takes place mainly within narrowband sound-frequency channels, and enables sound localization in the horizontal dimension. Two interaural differences are available to such binaural analysis. First, for sounds arising not directly from front or behind, the signal arrives earlier at one ear than at the other, creating an Interaural Time Delay or Difference (ITD). Second, for wavelengths roughly equal to, or shorter than, the diameter of the head, a shadowing effect is produced at the ear further from the source, creating an Interaural Intensity, or Level, Difference (IID or ILD) [5–8]. Figure 2.7 illustrates these differences, as described in Eq. 2.1. To a first-order approximation, binaural localization cues are ambiguous: many source locations give rise to nearly the same interaural differences. For sources more than a meter away, binaural localization cues are approximately equal for any source on a cone centered on the interaural axis (i.e., the well-known *cone of confusion*). The interaction between the two mechanisms resolves this issue.

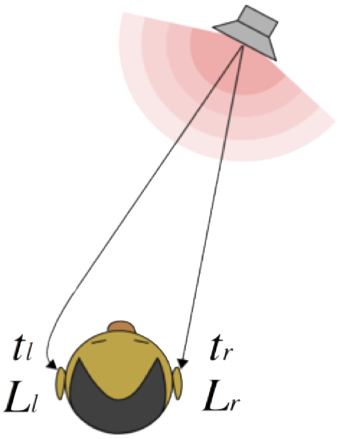


FIGURE 2.7: Binaural cues principle

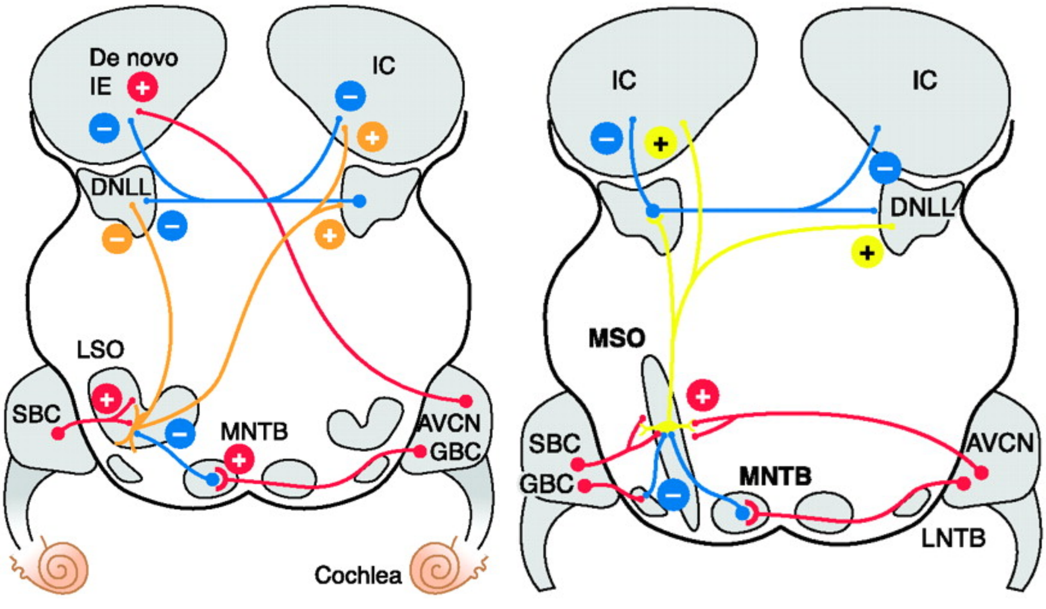
$$ITD = t_l - t_r \tag{2.1}$$

$$ILD = L_l - L_r \tag{2.2}$$

Humans show exceptional abilities in sound localization, discriminating changes of just 1–2 degrees in the angular location of a sound source. Studies under listening conditions using headphones (which enables the isolation of specific cues) confirm the remarkable accuracy of human spatial hearing, with thresholds as low as 7-10 μs for ITD and 1–2 dB for ILDs for presentations of binaural clicks [5, 6].

Lord Rayleigh's studies (1907) and their confirmation by by Stevens and Newman (1934) demonstrate the *Duplex theory* of sound localization. It states that ILDs are mainly employed in high-frequency localization tasks while timing differences (ITDs) are used for localization in low-frequency.

Despite its general acceptance, the dichotomy suggested by the duplex theory is not strict. First, significant ILDs can occur for low-frequency sounds located in the near field [9, 10]. Second, extensive psychophysical evidence indicates that sensitivity to ITDs is conveyed by the envelopes of high-frequency complex sounds [5, 11]. In this regard, recent studies have shown that when provision is made for temporal information in the envelopes of high-frequency modulated tones to match as closely as possible temporal information normally present in the output of low-frequency channels, ITD discrimination thresholds can be as good as, and in some cases surpass, those for low-frequency tones [12].



(A) ILD Neural Processing (B) ITD Neural Processing

Figure 2.8: Interaural Differences Neural Processing. AVCN: Anterio-Ventral Cochlear Nucleus, DNLL: Dorsal Nucleus of the Lateral Lemniscus, GBC: Globular Bushy Cells, IC: Inferior Colliculus, LSO: Lateral Superior Olive, MNTB: Medial Nucleus of the Trapezoid Body, MSO: Medial Superior Olive, SBC: Spherical Bushy cells. Credits: [5].

The neural pathway of ILD detection is illustrated in figure 2.8a. The initial site of ILD processing is generally considered to be the LSO (Lateral Superior Olive). ILD sensitivity is first created by convergence of excitatory inputs from SBCs (Spherical Bushy cells) located in the *ipsilateral* AVCN (Anterio-Ventral Cochlear Nucleus) and inhibitory inputs from ipsilateral MNTB (Medial Nucleus of the Trapezoid Body), which is itself innervated by GBCs (Globular Bushy Cells) of the *contralateral* AVCN. Projection of the LSO is excitatory to the contralateral DNLL (Dorsal Nucleus of the Lateral Lemniscus) and inhibitory to the ipsilateral DNLL and IC (Inferior Colliculus). In the IC, ILD sensitivity is created de novo by the convergence of monaural contralateral excitatory input from the AVCN and binaural inhibitory input from the DNLL.

Neural coding of ITDs demands the highest precision of any temporal process known to exist within the mammalian, reptilian, or avian brain. For humans, it appears that neurons in both of the major nuclei of the lower auditory brain stem, the MSO (Medial Superior Olive) and the LSO, are capable of extracting ITD information from their binaural inputs, although the MSO has traditionally been considered the major site of ITD processing. Figure 2.8b illustrates its neural pathway, where MSO principal cells receive binaural excitatory inputs from SBCs in the ipsilateral and contralateral AVCN, as well as binaural inhibitory inputs from the LNTB and MNTB. MSO neurons send then the excitatory projections to the DNLL and IC [4, 5].

It is important to note that the audio localization mechanism is still an active research area. Two theories have come to dominate our understanding of how the brain localizes sounds: the peak coding theory, which says that only the most strongly responding brain cells are needed, and the hemispheric coding theory, which says that only the average response of the cells in the two hemispheres of the brain are needed. A recent study [13] states that it is neither of them, and that the evidence of the previous studies only works because their experiments used unnatural/idealized sounds. If more natural sounds are used, both theories perform very badly. The study shows that to enhance performances with realistic sounds, one needs to use the whole pattern of neural responses, not just the most strongly responding or average response. It also showed two other key things: first, it has long been known that the responses of different auditory neurons are very diverse, but this diversity was not used in the hemispheric coding theory. This study shows how this aspect is relevant, particularly for natural sounds. Second, previous theories are inconsistent with the well-known fact that people are still able to localize sounds if they lose one half of their brain, but only sounds on the other side, which this study demonstrates.

The key point in interaural difference processing is the interaction between the ipsilateral and contralateral brain cells, and between the LSO and MSO. These findings are essential to our approach for spatial quality assessment as presented in section 4.3.3.2.

The next section enumerates the main 3D audio technologies and how these cues are used to virtually create spatial sounds.

2.2 3D audio technologies

The most basic audio system is Mono. Mono or monophonic is a technique where all the audio signals are mixed together and routed through a single audio channel. Mono systems can have multiple loudspeakers, and even multiple widely separated ones, but the signal contains no level and arrival time/phase information that would replicate or simulate directional cues. Common types of mono systems include single channel center clusters, mono split cluster systems, and distributed loudspeaker systems with and without architectural delays. Mono systems can still be full-bandwidth and full-fidelity and are able to reinforce both voice and music effectively. The big advantage to mono is that everyone hears the very same signal, and, in properly designed systems, all listeners would hear the system at essentially the same sound level. This makes well-designed mono systems very well suited for speech reinforcement as they can provide excellent speech intelligibility [14].

As the need of including spatial information arises, new technologies have been introduced as described below.

2.2.1 Multichannel systems

The most common and known spatial audio technology is *Stereo Sound*. Stereo sound systems can be divided into two categories: The first is *true* or *natural* stereo in which a live sound is captured, with any natural reverberation or ambience present, by microphones array. The signal is then reproduced over multiple loudspeakers to recreate, as closely as possible, the live sound. The second is *artificial* or *pan-pot* stereo, in which a mono sound is reproduced over multiple loudspeakers. An artificial direction, relative to the listener, can be then suggested by varying the relative amplitude of the signal sent to each speaker. The control that is used to vary this amplitude is known as a *pan-pot* (panoramic potentiometer). By combining multiple *pan-potted* mono signals together, a complete, yet entirely artificial, sound field can be created [15].

As a result, a mono signal that is panned somewhere between the channels does not have the requisite phase information to be a true stereophonic signal, since it is based on level difference only.

An additional requirement of the stereo playback system is that the entire listening area must have equal coverage of both the left and right channels, at essentially equal levels. This explains the *sweet spot* phenomenon between the loudspeakers, where the level differences and arrival time differences are small enough for the stereo image and localization to be both maintained. The sweet spot is generally limited to a relatively small area between the loudspeakers. When a listener is outside that area, the image may collapse and only part of the channels is heard. Although a sweet spot in a living room might be acceptable, it is more problematic in a larger venue (church, theater, cinema, etc) where the sweet spot might only include a third of the audience, leaving the others wondering about the program quality.

Moreover, a stereo playback system must have the correct absolute phase response input to output for both channels. This means that a signal with a positive pressure waveform at the input to the system must have the same positive pressure waveform at the output of the system. When the absolute polarity is flipped the wrong way, a stable center channel image cannot be achieved and will wander around away from the center, localizing out at both the loudspeakers [14]. Also, stereophonic systems cannot create the elevation impression or at least with some considerable constraints [16].

New technologies have been introduced to resolve these limitations. *Vector Based Amplitude Panning* (VBAP) is an efficient extension of stereophonic amplitude panning techniques, applied to multi-loudspeaker setups. In a horizontal plane around the listener, a virtual sound source at a certain position is created by applying the tangent panning law between the closest pair of loudspeaker. This principle was also extended to project sound sources onto a three dimensional sphere and assumes that the listener is located in the center of the equidistant speaker setup [17].

In VBAP the number of loudspeakers can be arbitrary, and they can be positioned in an arbitrary 2-D or 3-D setups. This technique produces virtual sources that are as sharp as possible with current loudspeaker configuration and amplitude panning methods, since it uses at one time the minimum number of loudspeakers needed (one, two or three). VBAP is a method to calculate gain factors for pair-wise or triplet-wise amplitude panning. It uses a triangulation of the convex hull around the given sound source [18]. In pair-wise panning it is a vector reformulation of the tangent law. Differing from the tangent law, it can be generalized easily for triplet-wise panning. More details on this method are available in chapter 5.

Ambisonics is a more advanced technique developed for encoding soundfields, and decoding them onto speaker arrays [19]. Initially it was used only to 1st order, with 4 signals that encode a full sphere of sound around a central listener. More recently, Ambisonics has been employed at higher orders, whereby it is possible not only to increase the angular resolution of distant sources, but also to extend the listening region and recreate accurately the soundfield from nearfield sources. Ambisonics encoding of any order is referred to as B-format, borrowing the original terminology for 1st order. Using high-order encodings, the listener receives distance cues about near sources exactly as they would for the real soundfield, because the soundfield around the listener can be reconstructed arbitrarily well [20, 21]. The sounds directions are encoded instead of the speakers signals and the positions are defined in term of spherical harmonics.

Ambisonics is basically a microphoning technique, where specifically designed microphones capture the 3D sound-field [22]. However, it can also be used to simulate a synthesis of spatial audio [23, 24]. In this case it is equivalent to an amplitude panning method in which a sound signal is applied to all loudspeakers placed evenly around the listener with gain factors calculated accordingly.

The most advanced and promising technique currently is *Wave Field Synthesis* (WFS). It was initially invented in the late 80's by Berkhout and has been further developed at the TU Delft [25, 26]. The basic idea is related to the Huygens' Principle which states, that an arbitrary wave front may be considered as a secondary source distribution. Regarding the propagating wave from the given wave front we cannot differentiate if it was either emitted by the original sound source (the primary source) or by a secondary source distribution along this wave front. As a consequence, the secondary source distribution may be substituted for the primary source, in order to reproduce the primary sound field. Based on this, WFS reproduces sound waves using distributed loudspeaker arrays [27].

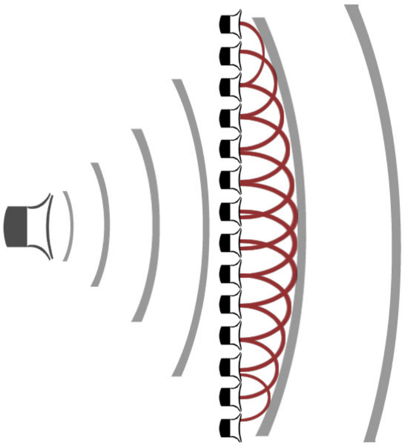


Figure 2.9: Wave Field Synthesis principle. Credits: Sound And Vision Magazine.

The main advantages of WFS are that the direction of rendered point sound sources is independent of the listeners' position. In other words, no sweet spot. Also, sound source characteristics and distance can be easily controlled. The primary limitations are the very high number of loudspeakers that it requires to properly reconstruct the sound waves, and accordingly the considerable computational resources that it needs and the deployment constraints that it implies.

Simpler techniques can be used to create spatial sound on headphones, using the properties of human perception, as explained in the next section.

2.2.2 Binaural Synthesis

Binaural synthesis is inspired by the natural process of sound localisation as illustrated in figure 2.7.

For a considered sound source, $s(t)$, at a specific position in space, the sound wave will reach respectively the left and right ear, at time t_l and t_r, and level L_l and L_r. The Interaural Time difference (ITD) and the Interaural Level Difference (ILD), as explained above, are the main cues used by the brain to determine the position of the audio source. This also implies that the signal $s(t)$ will be perceived by the left and right ear as $s_l(t)$ and $s_r(t)$ (Eq. 2.3). $h_l(t)$ and $h_r(t)$ are the Binaural Room Impulse Response (BRIR) that match the spatial position of the audio source and also comprises the room acoustical properties and the listener's body characteristics. BRIR is the equivalent time-domain to the well-known Head Related Transfer Functions (HRTF).

$$s_l(t) = s(t) * h_l(t) \tag{2.3}$$

$$s_r(t) = s(t) * h_r(t) \tag{2.4}$$

Consequently, binaural synthesis consists of convolving the audio source with the BRIR that matches a specific position. In this way, we can virtually place a sound at any position, as long as the impulse response is available for it. If measured correctly, an HRTF contains the interaural differences as well as the spectral cues (notches).

Binaural synthesis is designed to be used primarily on headphones, otherwise, additional processing is required. It also supposes that the listener is not moving his/her head because the soundscape will be moving along with, which may cause inside-head localization [17, 28].

2.2.3 Object-based approach

One of the recent standardization activities in the MPEG audio group is Spatial Audio Object Coding (SAOC). This is a technique for efficient coding and flexible, user-controllable rendering of multiple audio objects based on transmission of a mono or stereo downmix of the object signals [29]. Unlike channel-based techniques where audio sources are mixed as channels for a specific target setup, object-based production consists of encoding audio sources together with side information (meta data) in order to form audio objects [30]. They are then rendered by the receiver who adjust to the playback setup.

Currently, one of the most promising format for spatial audio object coding is Multi-Dimensional Audio (MDA). It is an open end-to-end audio platform that allows sound engineers to create object-based audio content, channel-based or a hybrid combination of the two. More details on MDA are available on DTS website[1].

2.3 Summary

This chapter describes the characteristics of human sound localization and how it inspired a lot of audio spatialization techniques. Besides binaural, these systems require an important number of channels to reproduce the spatial immersion. Also, these channels are highly synchronized and should be kept so to maintain the auditory display. In addition to professional audio applications constraints, these are the basic foundations of our transport technique as explained in the next chapter. The processing mechanism of auditory information is the theoretical basis for our approach for quality assessment.

[1] www.dts.com

Chapter 3

Audio Transport

3.1 Context

Recent developments in multimedia technologies have heightened the need for efficient transmission mechanisms. As stated in the previous section, transport is one of the most important and critical component of audio chains. In fact, it can even influence almost every part of the chain and therefore the whole system design.

Audio transport is not the same as regular data. It requires mechanisms that guarantee the timing and reliability constraints that come with, especially for professional applications.

In this chapter, we present our work to establish an optimal architecture for audio transport, specifically designed for the spatial sound systems used in professional applications. Through the study of the existing techniques, we highlight the missing aspects regarding the growing requirements of multichannel sound in pro audio and suggest our tailored solution to meet these needs.

3.2 Literature review

Traditionally, audio signals are directly transported in their analog form. Here is a brief description of the basis of the classical transmission techniques and their limitations.

3.2.1 Analog Audio Transport

Two types of cable are used for that purpose: unbalanced and balanced lines [31, 32].

Unbalanced lines are characterized by the fact that the cable and connectors use only two conductors, a center conductor surrounded by a shield. The shield stays at a constant ground potential while the signal voltage in the center conductor varies in a positive and negative manner relative to it. Very little or no interference will be able to reach the center conductor and interact with desired signal, since the shield will intercept them. Because the shield is one of the two conductors, it must always be connected at both ends of the cable. This may set up a condition called *ground loop*, that might produce a *hum effect* when the grounds of different electrical equipment are connected to each other.

If outside electrical interference does manage to penetrate the shield, it will mix with the desired signal present in the center conductor and be amplified right along with it as noise or buzz. In environments containing a lot of interference or when an unbalanced signal is sent long distances, such as down a snake, it will become more and more susceptible to unwanted interference.

This problem can be alleviated with the use of balanced lines. They use two conductors for the signal at center, surrounded by a shield. This shield is connected to ground like unbalanced lines but it is not required as one of the signal conductors. Its sole purpose is to provide defense against interference.

A benefit of this configuration is that the shield only needs to be connected to ground at one end of the cable in order to work, which eliminates the ground loop problem of unbalanced lines, except when transmitting phantom power. The two center conductors of a balanced line act as the main conduit for the signal and operate in a *push-pull* manner. Both conductors are equal in voltage but opposite in polarity (differential transmission).

If an electrical interference manages to penetrate the shield, it will equally interact with both conductors but with the same polarity. Thanks to the differential amplifier at the receiving circuit, the two interferences simply cancel each other out. This ability of balanced lines to reject noise and interference makes them popular when it is necessary to send signals over long distances [31, 32].

Whenever multiple audio signals need to be conveyed, snake cable are used. This is the case of professional applications dealing with multichannel audio streams. Snake cable is an audio multicore cable that contains from 4 to 64 individual audio cables inside a common outer jacket. Additional amplifications might be required to compensate the cable resistive loss and cover long distances.

The main issue with analog audio is that each channel requires a separate physical circuit. For instance, each microphone in a studio or on a stage must have its own circuit back to the mixer, which makes the the signals routing inflexible. Also, according to this aspect, analog audio transport is energy-intensive and expensive.

Because of these limitations and thanks to digitization and the progress of computer networks, digital audio transport has gained a new momentum as explained in the next section.

3.2.2 Digital Audio Transport

Digital audio is traditionally wired in a similar way to analog. The main difference is that several channels can share a single physical circuit, thus reducing the number of cores needed in a cable. Routing of signals is still inflexible and any change to the equipment in a location might involve new cabling [33].

Here is some of the most used technologies for digital audio transmission in professional context:

AES3 also known as AES/EBU or S/PDIF for the consumer grade version, is a standard format of serial digital audio transmission for a two-channel Linear Pulse Code Modulated (LPCM) sound. It allows data to be run at any rate, and encodes the clock and the data together using Biphase Mark Code (BMC). For multichannel stream, parallel AES3 interfaces can be used to carry every two channels separately. An ongoing AES project (AES-X196) aims to extend the current specifications to carry up to 8 channels.

ADAT historically refers to Alesis Digital Audio Tape but it becomes so popular that it describes now a protocol based on an optical interface to transfer multichannel audio. ADAT uses a Non Return to zero, Inverted (NRZI) coding and can carry eight channels of uncompressed sound at a 48 KHz sampling rate. For higher rates, the signal can split up into multiple channels.

MADI also known as AES10, is a Multichannel Audio Digital Interface that supports serial digital transmission of multiple channels over coaxial cable or fibre-optic lines of 28, 56, or 64 channels. It is an AES3 extension to multichannel streams.

FireWire also known as IEEE 1394 or i-LINK. is a serial bus interface standard for high-speed communications and isochronous real-time data transfer. Thanks to its high bandwidth (up to 3.2 Gbps with the IEEE 1394b-s3200 version), a high number of multiplexed audio channels can be transfered but the cable length cannot exceed 4.5 m.

Digital point-to-point transmission methods as the ones listed have the critical disadvantage of requiring separate wiring. As for analog audio transport systems, this make the signals routing inflexible and still not cost-effective.

For all these reasons, network-based solutions make perfect sense for multichannel audio transmission in this context of professional applications. As a matter of fact, and for more than a decade now, audio networks have been growing in popularity, used in studios, professional broadcast and other commercial facilities to deliver uncompressed, multi-channel, low-latency digital audio over standard networks [33, 34].

Ethernet-based audio networks are becoming the preferred solution as it can make use of the existing IT infrastructure. Thanks to its low cost and scalability, many companies have been using Ethernet as a basis of their audio networks. In addition, it allows an easier monitoring and management thanks to the existing protocols. The studies in [33, 34] list the existing technologies that can be classified in two categories:

- Layer 3 solutions: based mainly on IP (Internet Protocol). It is also known as Audio-Over-IP.
- Layer 2 solutions: based directly on Ethernet (IEEE 802.3). It is also known as Audio-Over-Ethernet.

The term *layer* refers to the OSI (Open Systems Interconnection) model, as IP and Ethernet are respectively on the 3^{rd} and 2^{nd} layer of the model.

In the context of professional applications, the following features are the starting point for a network-based audio transmission system for multichannel audio:

- Reliability
- Latency
- Synchronisation
- Channels number and Audio quality
- Maximum supported Distance
- Topology and Routing options
- Control interface and network management

These prerequisites are detailed in section 3.2.6.

The next section describes the characteristics of each category, and how well they meet the previous requirements. A complete benchmark and analysis is not possible though since most of the existing techniques are proprietary. We focus here on some of the most used and documented ones.

3.2.3 Layer 2 Solutions

Layer 2 solutions are based on Ethernet (IEEE 802.3 standards).

Ethernet is a set of computer networks technologies that are widely used for Local Area Networks (LAN). Its physical layer has evolved over time and include coaxial, twisted pair and fiber optic physical media interfaces, with speeds ranging from 10 Mbit to 100 Gbit. The most commonly used forms are 10BASE-T, 100BASE-TX, and 1000BASE-T, and utilize twisted pair cables and RJ45 (8P8C) modular connectors. They respectively run at 10 Mbit/s, 100 Mbit/s, and 1 Gbit/s. Fiber optic variants of Ethernet offer high performance, electrical isolation and distance but at a higher cost. In general, network protocol stack software will work similarly on all varieties [35].

In IEEE 802.3, a datagram is called a packet or frame and includes the payload (data) as well as additional information to manage the transmission (Fig. 3.1). A frame starts following a 7 bytes Preamble and one byte Start Frame Delimiter (SFD). The frame begins then with a frame header featuring source and destination MAC (Media Access Control) addresses as well as the protocol used for the payload (Ether Type). The middle and main section consists of payload data that can include additional protocols. The frame ends with a 32-bit Cyclic Redundancy Check, which is used to detect data corruption during the transmission. An inter-packet gap (IPG) of a minimum of 12 bytes is finally inserted to separate packets [36] .

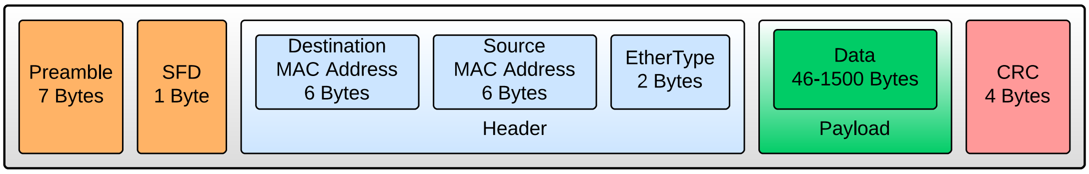


FIGURE 3.1: Ethernet Type II Frame format

The basic configuration of Ethernet does not make it suitable for real time audio transport. CobraNet is one of the first methods that used Ethernet for this purpose [37, 38]. It introduces additional mechanisms as illustrated in figure 3.2. This technique uses three basic packet types. All packets are identified with a unique protocol identifier (0x8819) stated in the EtherType field.

The first one is the *Beat Packet*. It is a multicast addressed packet that contains network operating parameters, clock and transmission permissions. The beat packet is transmitted from a single CobraNet device on the network (the conductor) and indicates the start of the isochronous cycle. Since the beat packet carries the clock for the network, it is sensitive to delivery delay variation. Failure to meet the delay variation specification may prevent devices from locking their local sample clock to the network clock. The beat packet is typically small (100 bytes) but can be large on a network with numerous active bundles.

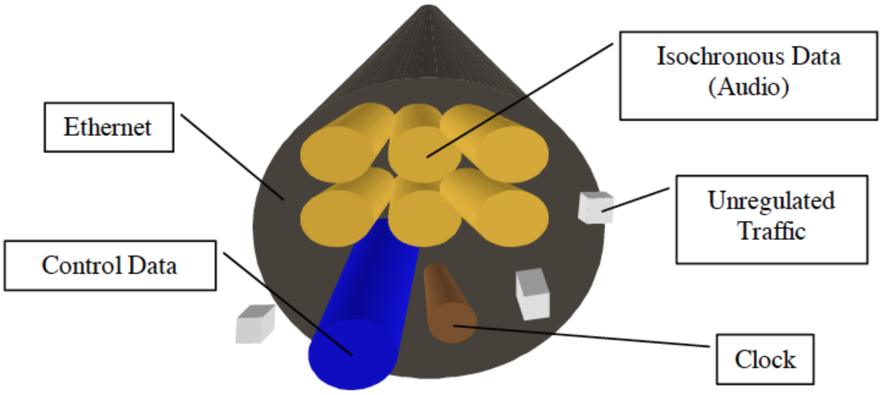


FIGURE 3.2: CobraNet Principle. Credits: [37].

The second one is the *Isochronous Data Packet* that contains the audio data, and which can be a unicast or multicast packet, depending on the number of destinations. Buffering is performed in the CobraNet devices thus out of order delivery of data packets is acceptable. Data packets are typically large (1000 bytes) to reduce the protocol overhead. The last one is the *Reservation Packet* which is a multicast addressed packet as well. CobraNet devices typically transmit a small reservation packet once per second for node management. It is theoretically able to convey up to 64 channels but can practically carry around 56 channels.

A typical uncompressed audio channel in pro audio is coded on 24 bit and sampled at 48 KHz. Accordingly, the bandwidth required for a channel is:

$$BW_{ch} \quad = WordSize \times Fs = 24 \times 48K = 1,152 \quad Mbps \tag{3.1}$$

Since the available ASICs for CobraNet only support 100 Mbps Ethernet and the Protocol Efficiency (PE) is relatively low, this reduces the actual maximum Number of Channels (CN) to:

$$PE = \frac{PayloadSize}{PacketSize} = \frac{1000}{1538} = 65,02\% \tag{3.2}$$

$$CN = int\left[\frac{BW_{total} \times PE}{BW_{ch}}\right] = int\left[\frac{100 \times 65,02}{1,152}\right] = 56 \quad channels \tag{3.3}$$

Also, the used multicast packets are actually broadcast packets as Ethernet does not distinguish between the two frames, which eventually flood the network.

Besides, audio artifacts may occur in CobraNet networks if other data exist on the Ethernet network and interrupt the timing. It is possible though to separate the CobraNet time-critical audio data from asynchronous external data by using manageable switches that support VLANs in line with IEEE 802.1q [34, 39]. But then again, this introduces additional requirements.

In the same category, EtherSound is a protocol developed by the Digigram to convey pro audio, with the purpose of offering a low jitter and latency [40]. An Ethersound frame has a fixed size of 236 bytes as illustrated in figure 3.3.

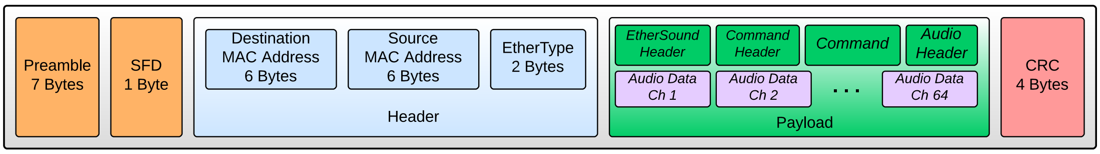


Figure 3.3: EtherSound Frame format of a fixed size of 236 bytes. A 24-bit sample from all channels is embedded within a single frame.

Up to 64 channels of audio may be transmitted in both directions using Ethersound [41]. As described above, this system combines a 3-Bytes sample from all the 64 channels within a single frame. The frames are transmitted at the sampling rate (fixed rate of 48 KHz), which allow the Digital-to-Analog and Analog-to-Digital Converters on each side to lock on the start of a frame, and consequently reduce jitter.

Ethersound uses daisy chain topology. It means that the devices are connected in a single line and audio data are transmitted from one to another. The wiring system is practical for live sound and tours but lacks reliability as one device can leave only a half of the network working. As for cobraNet, Ethersound does not include any routing mechanism (not ip-based), which may limit the network deployment (subnets).

Similarly to the previous one, non-audio traffic on the network deteriorate Ethersound performance. It also shows compatibility issues between devices using the same technology (the ES-100 and ES-Giga interfaces cannot coexist on the same network).

Audio Video Bridging (AVB), the newest addition to this category, is the IEEE standard under work for audio transport over Ethernet [42, 43].

From an application view, Ethernet AVB is not really a complete, ready-to-use solution, but a set of standards and protocols [44]. In addition to Ethernet standards, IEEE 802.1 focus on the bridging and management of audio and video streams. There are four IEEE 802.1 AVB protocols that build the so-called AVB *Plumbing*:

- IEEE 802.1BA [45] that describes the system specifications.
- IEEE 802.1AS [46], which relies on the Precision Time Protocol [47] to ensure devices synchronization.
- IEEE 802.1Qat [48], Stream Reservation Protocol (SRP), that manages the network resources allocation to guarantee the QoS (Quality Of Service) requirements.
- IEEE 802.1Qav [49], Traffic Shaping that provides mechanisms to guarantees packets delivery.

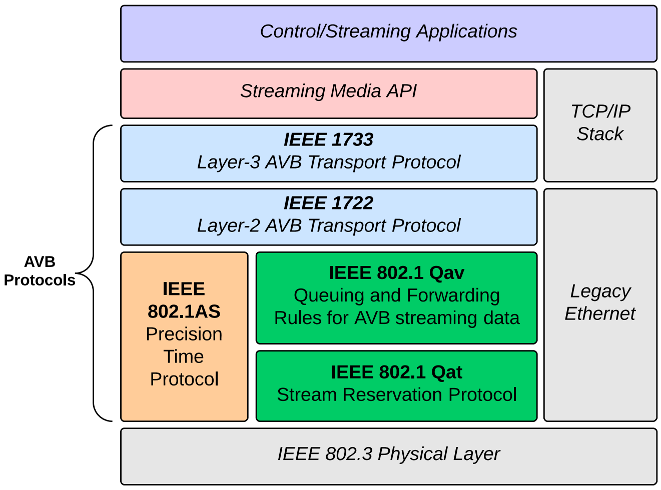

Control/Streaming Applications
Streaming Media API
TCP/IP Stack
IEEE 1733
Layer-3 AVB Transport Protocol
IEEE 1722
Layer-2 AVB Transport Protocol
AVB Protocols
IEEE 802.1AS Precision Time Protocol
IEEE 802.1 Qav
Queuing and Forwarding Rules for AVB streaming data
IEEE 802.1 Qat
Stream Reservation Protocol
Legacy Ethernet
IEEE 802.3 Physical Layer

Figure 3.4: Overview of Ethernet AVB

Additionally, it relies on IEEE 1722 and IEEE 1733 which are respectively the layer-2 and layer-3 transport protocols for AVB, used for control and management [42, 50].

The main drawback of Ethernet AVB is that it requires the implementation of special AVB-compliant switches to handle the modified Ethernet frames and manage synchronization, which eventually increases the solution cost. In addition, AVB switches reserve up to 75% of the network bandwidth to media traffic [51]. Although XMOS has released AVB chips and interfaces, it still on the beta stage.

Moreover, the preliminary study mentioned in the first chapter, shows that despite the *best effort* design of the previous solutions, several users have reported packet drops while using them (except AVB that shows more stable performances[1]).

3.2.4 Layer 3 solutions

Layer 3 solutions operate on a slightly higher level since they are based on Internet Protocol (IP). They use the same physical infrastructures as the previous categories but with additional mechanisms, without changing the lower layers configuration.

Internet Protocol (IP) [52, 53] is the well-known network protocol (L3) used in packet-switched computer networks. IPv4 (Fig. 3.5) is the most used version of IP in WAN and LAN but IPv6 is progressively being deployed in both. The use of IP-based technologies has become a strategic element in the design, development and use of telecommunication networks.

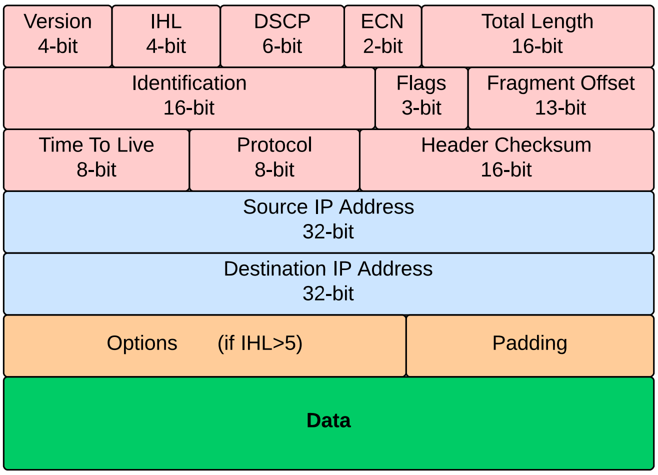


Figure 3.5: IPv4 Frame Format

In order to handle real-time audio transport, IP requires additional mechanisms that are introduced by the followings systems. UDP (User Datagram Protocol) is the main used protocol for this purpose as TCP (Transfer Control Protocol) mechanisms will interfere with timing management [51].

Dante is a commercial network-based audio transmission system developed by Audinate (2006). It transmits the audio data over IP using UDP packages in a 100 Mbps or 1 Gbps Ethernet. As AVB, it uses PTP [47] for synchronization. Dante relies on the QoS features of the network switches, just as they are used for VoIP traffic, to prioritize

[1] The study in [50] shows no Packet Loss for AVB but has been performed in a non realistic simulation mode.

audio data over external ones [54]. Audio Data are encapsulated using a propretary Audio Transport Protcol (ATP) based on the Real Time Protocol (RTP).

The special feature of Dante compared to the previous methods is that it can be integrated into a conventional router network, since it transports over IP. Moreover, it uses IP-based zero configuration concepts such as ZeroConf [34, 55, 56], to offer plug-and-play capability to its devices. Dante allows the transport of 48 channels in a 100BaseT link and ten times that amount on a Gigabit interface.

Dante does not handle the eventual packet drops. For this purpose, the *Core Module* is equipped with two network connectors and which, when breakdowns occur, should make it possible for a network to transfer without any glitches to the other one. Dante uses non standard control protocol, which, as explained in the next section, is a decisive criterion for network deployment.

On the same category, Ravenna [57] is an open solution based on IP and introduced by Alc Networx. Accordingly, it offers the same advantages as presented above. It employs a collection of standardized network protocols to guarantee real time audio transport as shown below:

- Real Time Protocol (RTP) [58] for media streaming.
- PTPv2 (IEEE1588-2008) for synchronization.
- DiffServ [59] for QoS management.

A receiver can subscribe to any existing Ravenna stream through RTSP/SDP protocol which are supported by most common media players. Ravenna assumes that packet delivery in the underlying network is quite reliable and incurs little delay variance. It does not automatically guard against packet loss, hence lost packets will usually be noticeable as dropouts in a stream. Network redundancy is offered as a way to guard against this [57].

3.2.5 Control protocols

Devices control and management is one of the fundamental requirements of professional audio networks.

The most famous one is Open Sound Control (OSC) [60]. Considering the OSI model, OSC is classified as a Presentation Layer (Layer 6) entity. It uses a hierarchical tree structure to address control points that correspond to remote function calls of the device. It is a content message format that can be compared to XML or JSON.

OSC streams are sequences of frames defined with respect to a point in time called a timetag (NTP timestamp). The frames are called bundles. Inside a bundle are some number of messages, each of which represent the state of a sub-stream at the enclosing reference timetag. The sub-streams are labelled with a human-readable character string called an address [61]. Figure 3.6 illustrates the OSC format.

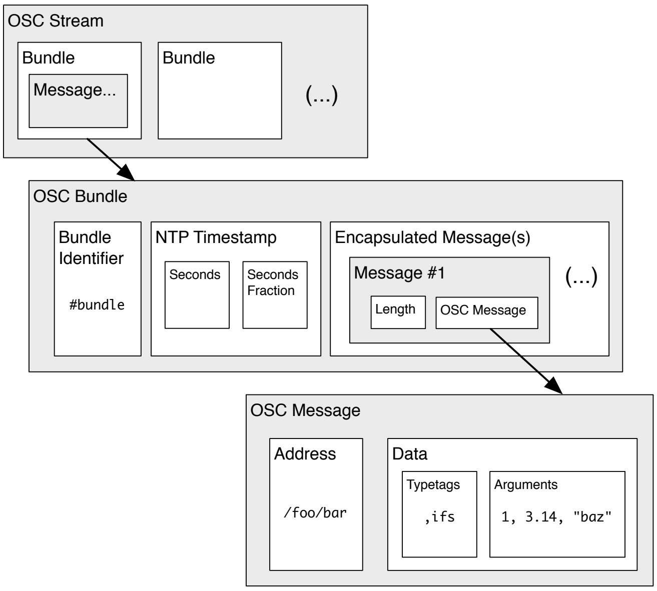


Figure 3.6: Structure of the OSC content format. Credits: [61]

Alternatively, XFN is an IP-based peer to peer network protocol, in which any device on the network can send or receive connection management, control, and monitoring messages [62]. The basis of XFN is that each parameter in a device is addressable via a hierarchical structure that reflects the functional layout of the device.

XFN defines a fixed 7-level hierarchy for modeling a device's parameters [63]. In XFN, parameters refer to the features on a device which have values that can be retrieved and/or modified. It also provides control capabilities such as:

- Grouping of networked devices.
- Joining of different parameters (or controls) regardless of which device the parameters belong to.
- Modifiers that allow management of complex relationships between parameters.

IEC62379 is implemented using the Simple Network Management Protocol (SNMP) [64], and represents signal paths through a device by means of connections between standard, predefined functional blocks. Functional blocks combine and route audio signals, and can implement audio processing functions found at points on a signal path. Although, higher-level groupings of signal processing functions such as an equalization section or channel strip are not supported. Parameter data is not independently accessible; parameters associated with signal processing functions are contained within the corresponding functional blocks. Service enumeration proceeds by tracing the connections between functional blocks. These connections represent all signal paths within a device. Enumeration of all device inputs and outputs supports connection management. Generally speaking, it lacks flexibility [65].

Open Control Architecture (OCA) is a system control and monitoring architecture [66] under work by several companies (OCA Alliance). It is designed to cooperate with current and future media signals transport standards. It is part of the ongoing AES standardization project, X210. The OCA definition has three parts:

- An Architectural Framework, OCF, that defines the set of structures and mechanisms upon which the rest of OCA is based.
- An object-oriented Class structure, OCC. It is an expandable, evolvable hierarchical structure which defines OCA's repertoire of control functions.
- A protocol definition, currently only OCP.1 exists which describes an implementation of OCA for standard TCP/IP networks.

OCP.1 uses Apple's ZeroConf mechanism, the Bonjour service [67] for devices discovery and management. Fig. 3.7 illustrates an example of OCA communication.

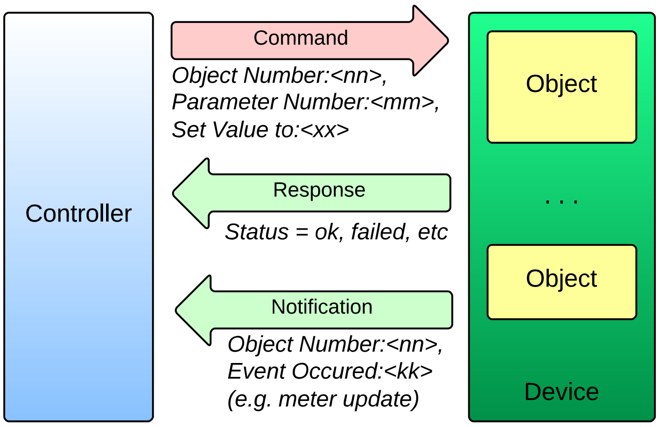


FIGURE 3.7: Example of OCA communication. Objects can receive and execute commands or emit notifications.

3.2.6 Summary

The classical analog point-to-point audio transmission is no more adapted for the requirements of high quality multichannel audio in the context of professional applications. Digital transmission is more flexible, scalable, less-expensive and more energy efficient.

The existing IT networks are capable of transporting multichannel Audio streams over IP or directly over Ethernet, with the help of additional mechanisms. As discussed, each approach has its pros and cons. Table 3.1 summarize the main ones. We used AVB and Ravenna as example for Layer 2 and Layer 3 comparison as they are the only two open methods. More details can be found in [33, 44, 51].

Table 3.1: Comparison of Layer 2 and Layer 3 audio transport techniques

Criterion	**Layer 2**	**Layer 3**
Channel Number	High	High
Reliability	Bandwidth Reservation	Full network Redundancy
Synchronization	IEEE 802.1AS gPTP	IEEE 1588-2008 (PTPv2)
Latency	Guaranteed	Variable
Infrastructure	Dedicated network and special switches	Any. Manageable switches and IT traffic management are recommended though
Management	1722.1-2013 (Tunnel for control protocols)	Variable

In the next section, we present our approach to resolve some of the issues of the previous methods. More details on the currently existing technologies are available in appendix A.

3.3 Suggested approach

Less is more[2], as the german architect would say, is the basis of our approach. In order to fulfill the previous requirements, we established a minimalistic architecture to address them while maintaining a low latency and high flexibility, as presented through this section.

Audio networks requirements are detailed in [34, 68]. To summarize, a high performance audio network should guarantee the following criteria:

- Reliability: packet delivery must be guaranteed. A reliable form of transport is required, otherwise redundancy should be used.
- Synchronization: the quality of the auditory display depends on the inter-channel synchronization, which accordingly, should be maintained. According to AES11 standard [69], for a word clock frequency of 48 kHz, two network nodes should be within 2 μs of each other.
- Latency: the transport system should offer the lowest latency as possible (time transparency). Latency should also be deterministic so integrators can integrate it in their Data Flow. Generally, 2 to 5 ms are considered as acceptable values[3].
- Bandwidth: 3D audio systems requires a high number of channels: from 64 channel for Dolby ATMOS to more than 400 channels for WFS-based systems [70]. Pro applications generally use uncompressed audio on 24-bit at sampling rates ranging from 48 to 192 KHz. Accordingly, a high bandwidth is required to transport all these channels in parallel.
- Management options: the audio network should offer simple and efficient control protocols or at least allow such traffic.

In addition to these requirements, we were concerned about the following aspects:

- Low-complexity for an easier implementation on chip
- Scalability
- Cost-effectiveness
- Interoperability with other systems (at least at a minimal functioning mode).

The architecture that we are introducing offers these possibilities.

[2] Ludwig Mies van der Rohe

[3] Latency within A/V applications cannot exceed 40 ms.

3.3.1 Architecture

Our architecture as introduced in [71], is illustrated in figure 3.8. It is a minimalistic functional approach for multichannel audio transport based on IP.

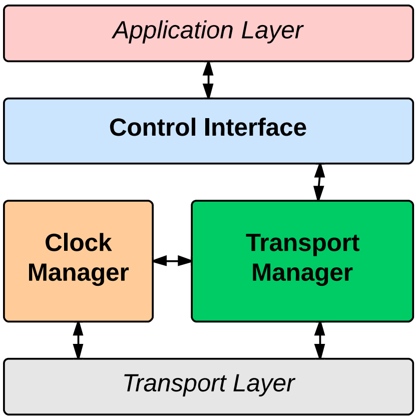


Figure 3.8: Minimalistic system architecture

SIRIUS system consists of three main components:

- Clock Manager handles nodes synchronization using a shared wall clock,
- Transport Manager creates SIRIUS RTP packets and includes additional information to enhance the system reliability,
- Control Interface enables device management and monitoring,

The system has been designed as a layer 3 solution so it can be deployed on any standard IT network, and to offer a certain level of interoperability as explained further. We essentially relied on IETF/IEEE standards for the same purpose.

SIRIUS is based on Real Time Protocol and is compliant with EBU (european Broadcasting Union) tech-3326 requirements [72–74]. Since we were involved in the AES-X192 project, we also took into consideration most of the remarks, which makes it compliant with the new AES-67 standard (announced on September, 2013 and issued by the end of March, 2014) [75].

3.3.2 Topology

A lot of audio networks use Daisy Chains topology as it uses less cable, but as discussed before, it is not cost-effective (two network interfaces are required) and its reliability is questionable. Since our concern is mainly about the system reliability, we choose to work with a two-level hierarchical tree as illustrated in 3.9. This topology is more manageable and scalable, and also allows us to suggest an approach to optimally reduce cable lengths for static setups (e.g. cinemas). Moreover, the hierarchical tree is more practical when it comes to fault-detection.

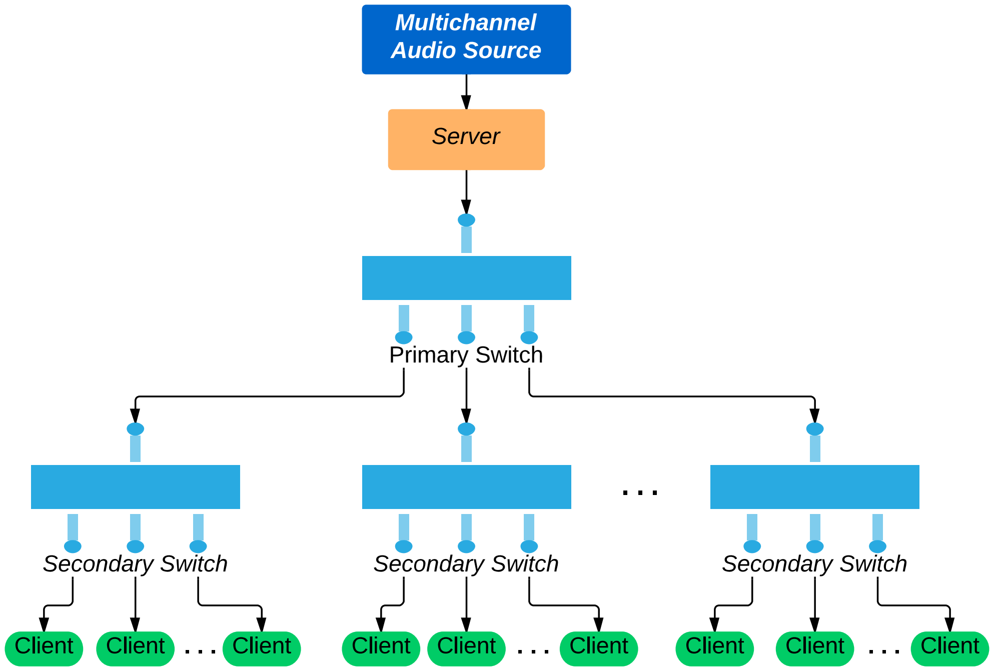


FIGURE 3.9: SIRIUS system topology. A hierarchical tree offers more scalability and easier management of professional audio networks.

Multichannel audio systems use extensively loudspeakers arrays. The loudspeakers are generally arranged into neighboring groups, which makes this deployment scheme even more relevant. Since this structure is derived from star topology, a fault at the primary switch may jeopardize the whole network performances.

3.3.3 SIRIUS Protocol Stack

As mentioned above, our approach is based on IP. Although, IPv4 is completely suitable and managed by our system, we used the IPv6 [76] version (Fig. 3.10) for the following reasons:

- IPv6 has a simpler and fixed size 40-Byte header which is better for implementation.

- QoS management can be performed at IP level using the *Flow Label* field [77] without requiring additional mechanisms. DiffServ can also be utilized using the *Traffic Class* field.
- The audio stream fragmentation into packets is performed by a SIRIUS device, which is compliant with IPv6 fragmentation strategy and frees the Router and Switch resources.
- The use of IPv6 allows us to avoid additional checksum calculation as it is already performed by UDP and Ethernet.
- Security options are enhanced in IPv6 (native support for IPSec).
- IPv6 nodes can perform stateless address AutoConfiguration [78] which offers more flexibility for audio solutions.

IPv6 addressing and routing capabilities are not discussed in the scope of our study. Although, they can be useful for an end-to-end audio network management.

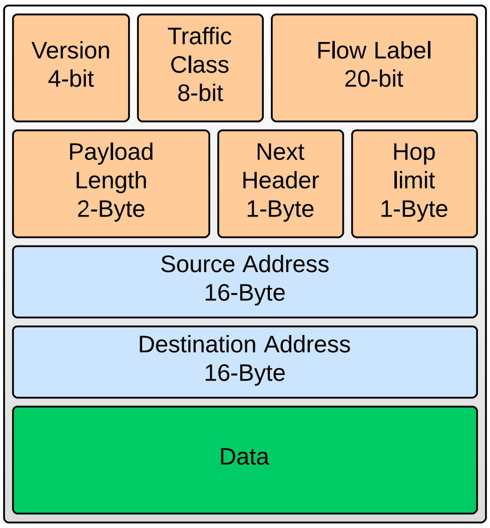


FIGURE 3.10: IPv6 Frame Format

In our approach, Real Time Protocol (RTP) is used on UDP for audio streaming instead of TCP because TCP mechanisms interfere with timing management.

RTP [58, 79] aims to provide useful services for the transport of real-time media, such as audio and video, over IP networks. It includes timing recovery, loss detection, payload and source identification, reception quality feedback, media synchronization, and membership management. RTP was originally designed for use in multicast conferences, using the lightweight sessions model. Since that time, it has proven useful for a range of other applications: video conferencing, webcasting, TV distribution, etc. The protocol has been demonstrated to scale from point-to-point use to multicast sessions with thousands of users, and from low-bandwidth cellular telephony applications to the delivery of uncompressed High-Definition Television (HDTV) signals at gigabit rates.

RTP does not handle those aspects per se, but offers a simple framework for other protocols to do so. Figure 3.11 illustrates its header. Fields description is as follow:

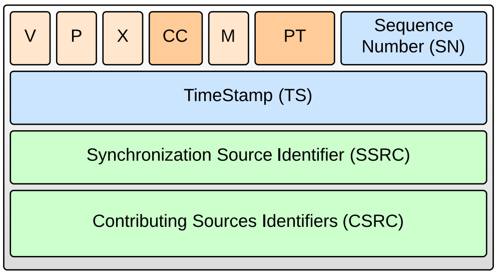

V
P
X
CC
M
PT
Sequence Number (SN)
TimeStamp (TS)
Synchronization Source Identifier (SSRC)
Contributing Sources Identifiers (CSRC)

FIGURE 3.11: Real-Time Protocol Header

- **V**, 2-bit, indicates the protocol Version (currently 2).
- **P**, 1-bit, indicates if the packet contains additional Padding octets at the end.
- **X**, 1-bit, informs if the fixed header is followed by a header extension.
- **CC**, 4-bit, the CSRC (Contributing Source) count contains the number of CSRC identifiers that follow the fixed header.
- **M**, 1-bit, the interpretation of the marker is defined by an additional profile. It is intended to allow significant events such as frame boundaries to be marked in the packet stream.
- **PT**, 7-bit, this field identifies the format of the RTP payload as well as the encoding/compression schemes.
- **SN**, 16-bit, the sequence number increments for each RTP data packet sent.
- **TS**, 32-bit, the timestamp reflects the sampling instant of the RTP data packet. The sampling instant must be derived from a clock that increments monotonically and linearly in time, to allow synchronization and jitter calculations.
- **SSRC**, 32-bit, Synchronization Source, identifies the synchronization source (stream). This identifier is chosen randomly, with the intent that no two synchronization sources within the same RTP session will have the same SSRC identifier.
- **CSRC**, a list of 1 to16 items, 32 bits each, enumerates the Contributing Sources for the payload contained in this packet. The number of identifiers is given by the CC field.

Additionally, An extension mechanism is provided to allow custom implementations. The header can contain a profile-specific extension header using the following fields: Extension Header ID, (2-Byte), Extension Header length, (2-Byte), and the actual Extension Header. Though the RFC recommends including such data as part of the payload instead of extending the header [80].

RTP is designed to work in conjunction with the auxiliary Real Time Control Protocol (RTCP) to get feedback on quality of data transmission and to provide minimal control and identification functionality.

The most important fields in our case are the Sequence Number and the Time Stamp. The Sequence Number is a randomly initialized integer that allows packet loss detection and to maintain sequences order. In fact, it is the main entry for the transport manager.

The Time Stamp, as the name indicates, marks the timing information for each packet, so they can be presented to the receivers accordingly. RTP itself is not responsible for the synchronization but it is a PTP task, as presented in the next section.

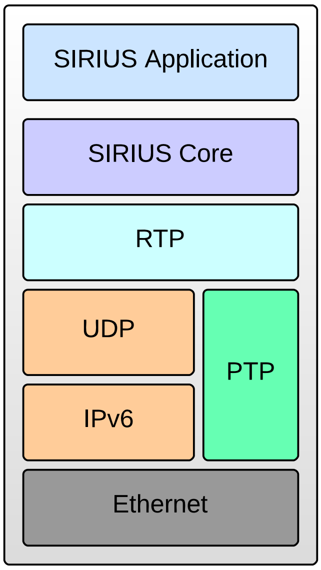
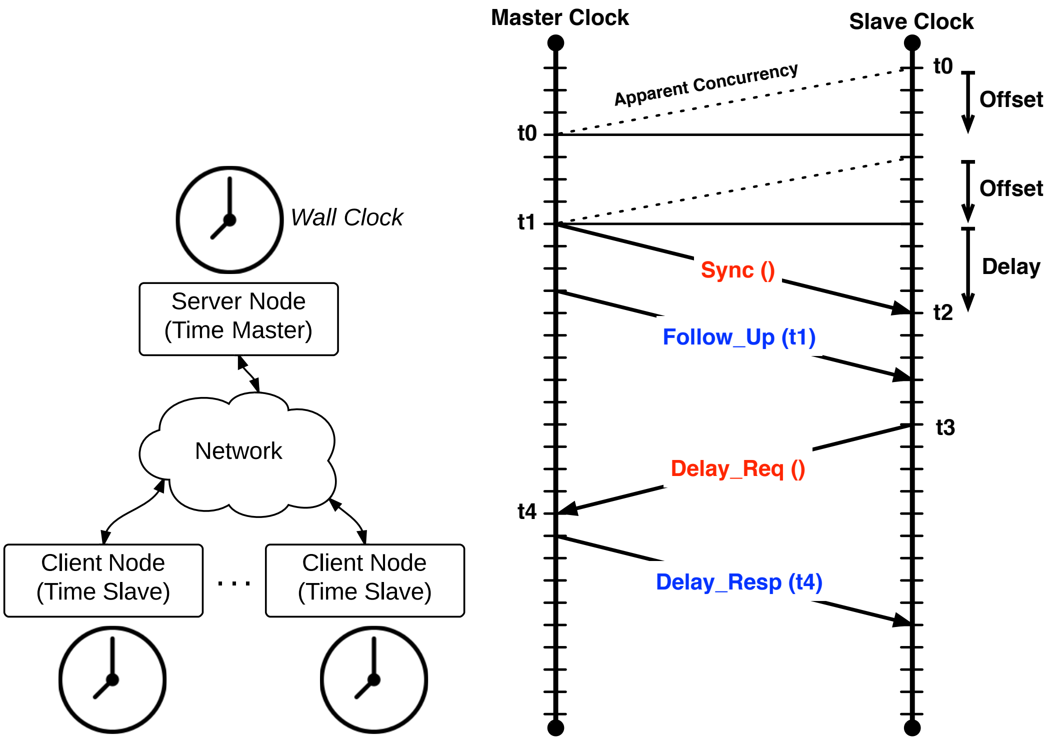


Figure 3.12: SIRIUS protocol stack

3.3.4 Synchronization

The network nodes synchronization is performed by the Precision Time Protocol (PTP), as defined in the IEEE 1588 standard [47]. PTP is based on a Master-Slave hierarchy and defines a procedure allowing many spatially distributed real-time clocks



(A) PTP Wall Clock principle (B) PTP offset estimation mechanism

Figure 3.13: PTP principle

to be synchronized to an accurate Master Wall Clock, through a packet-based network, such as Ethernet (Fig. 3.13a).

Synchronization and management of a PTP system is achieved through the exchange of specific messages as shown in figure 3.13b:

Initially and widely speaking, a slave node has a time offset with the master node that we need to adjust:

$$\text{Offset} = clock_{slave} - clock_{master} \tag{3.4}$$

The master sends a *Sync()* message to initiate the synchronization process and estimate the network delay. It is received by the slave at t_2:

$$t_2 = t_1 + \text{Delay} + \text{Offset} \tag{3.5}$$

Optionally, the server sends a *Follow Up* message containing the emission time t_1 to allow the slave to estimate Eq. 3.5.

The slave sends then a *Delay Request* message at t_3. The Master gets it at t_4.

$$t_4 = t_3 - \text{Offset} + \text{Delay} \tag{3.6}$$

For $A = t_2 - t_1$ and $B = t_4 - t_3$, we can estimate the Offset and Delay as follow:

$$\begin{aligned} A = t_2 - t_1 &= \text{Delay} + \text{Offset} && (3.7) \\ B = t_4 - t_3 &= \text{Delay} - \text{Offset} && (3.8) \\ \text{Delay} &= \frac{A+B}{2} && (3.9) \\ \text{Offset} &= \frac{A-B}{2} && (3.10) \end{aligned}$$

The Master finally sends a Delay Response message with the time t_4 to allow the slave to perform the calculation above and adjust its clock accordingly.

The accuracy of PTP is defined by two major factors: 1) The symmetry of the transmission path, assuming that one-way transmission delay is half of the round trip time and 2) the accuracy of TimeStamps [81].

Audio synchronization requirements are defined by the AES11 standard [69]. Respectively, the clock distribution system of the network must be able to keep each node within 5% of an audio WordClock period relative to the clock master. Accordingly we get the following values:

TABLE 3.2: Synchronization requirements

Sample Rate	44,1 KHz	48 KHz	96 KHz	192 KHz
Tolerated error to Master	1,13 μs	1,04 μs	520,83 ns	260,41 ns
Tolerated error to another channel	2,27 μs	2,08 μs	1,04 μs	520,83 ns

High precision time stamps can be achieved with the support of specialized hardware interfaces in the physical layer of the network. Well designed implementations can provide accuracy below 100 ns [82]. Unfortunately, many legacy systems lack such hardware interfaces. Accordingly, a hardware approach goes against our precept to use the existing infrastructures.

The requirements in 3.2 are made to ensure a low-jitter digital transmission. The auditory system though, can detect in the best cases a 10 μs ITD, in the case of binaural audio [5]. To ensure synchronization at a low cost, we use Branicky et al. [83] software-only implementation, namely the PTPd. It is an open source PTP implementation that can guarantee a 2-10 μs accuracy, as explained later in section 3.4.

As it has shown good performance [84], PTP is now used in most media transport technologies and part of the AVB and AES67 specifications. Using it as a synchronization system guarantees a higher interoperability with other audio transport techniques.

Finally, the timing correlation between PTP and RTP is established via the Session Description Protocol (SDP) [85, 86].

3.3.5 Transport Manager

The transport manager is the main component of our system and responsible of building the packet and manage timing and recovery information.

Since we based our system on UDP instead of TCP for timing constraints, we lost the reliability mechanisms of the latter. In fact, the basic form of CSMA/CD[4] based Ethernet suffers from the capture effect and is unsuitable for supporting real-time multimedia traffic. It behaves poorly under heavy load conditions, leading to an excessive delay, throughput degradation and packet loss because of the excessive collisions [87]. For networks with no CSMA/CD issues, using switches for instance, queuing under a high network load will be the main cause for packets drop [88].

Using IPv6 QoS mechanism offers packets prioritization but does not guarantee their delivery. Since UDP does not guarantee delivery either, and particularly under high network congestion, we introduce our specific Forward Error Correction (FEC) scheme to enhance the reliability of our system, based on the study established in [89].

Our approach is based on the statical assumption[5] that our transport channel follows a Gilbert-Elliott Model [90] for packet loss as illustrated below:

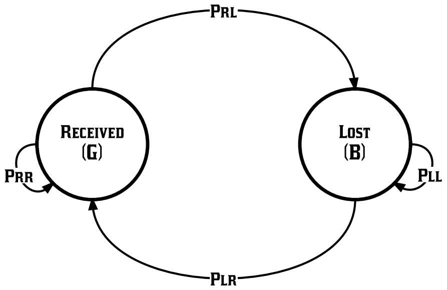


FIGURE 3.14: Sirius Channel Model

With a good network design, we noticed that the probability of *successively* losing two or more packets is relatively very low (ε=0,0038), which was confirmed by our tests.

Based on that assumption, our method consists of combining recovery data from the previous packet within the current packet to send.

[4] CSMA/CD is disabled when Ethernet switches are used but some networks still use Hubs.

[5] Measurements performed for several tests, using a hardware implementation under a normal network load.

TABLE 3.3: Channel Model properties

Probability	**Description**	**Value**
P_{RR}	Probability to receive a packet after a received one	0,5
P_{RL}	Probability to lose a packet after a received one	0,5
P_{LL}	Probability to lose a packet after a lost one	ε
P_{LR}	Probability to lose a packet after a received one	$1-\varepsilon$

The transport manager builds our packets using the mechanism of figure 3.15.

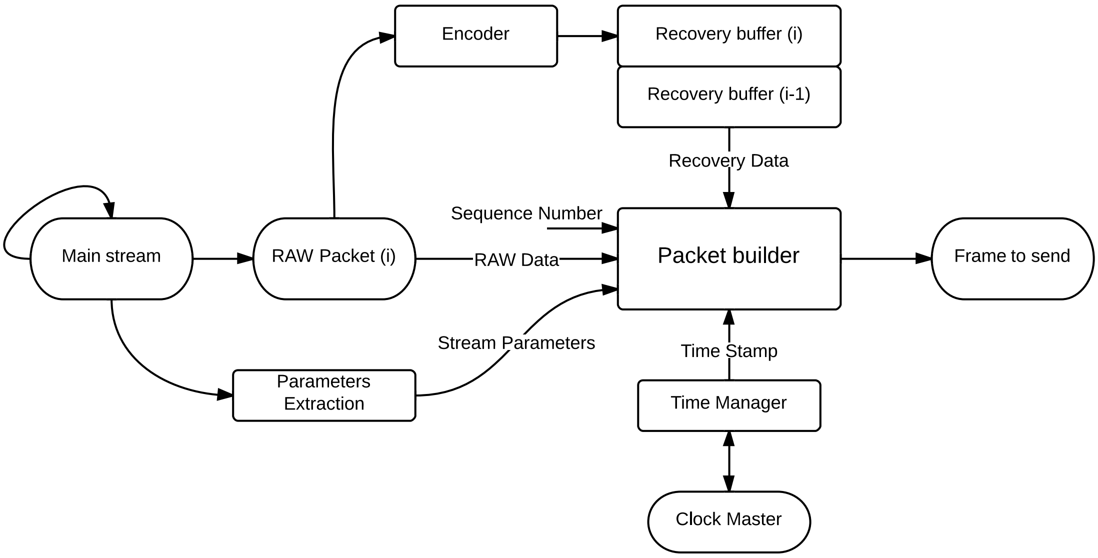


FIGURE 3.15: Packet building mechanism

As a result, the transport manager at the receiver is able to recover lost packets without any additional retransmission requests that may compromise the low latency and timing constraints, as described in figure 3.16.

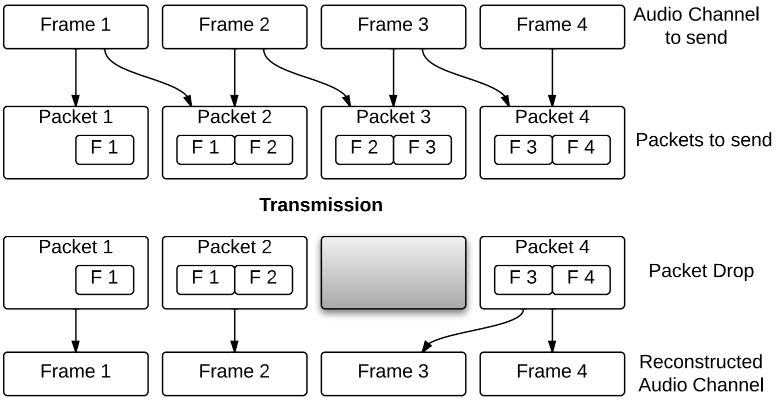


FIGURE 3.16: Packet Recovery mechanism

Using the information within the RTP header, a receiver is able to detect packet loss and reconstruct the initial stream without any additional latency. If it fails (more than one packet is lost), a simple repetition of the previous packet is used to conceal the packet drop and maintain the stream timing. For, applications that are not sensitive to latency issues, more advanced techniques can be used to enhance the concealment quality [91, 92].

Most of the existing audio transports technologies rely on redundancy to enhance the network reliability. Generally speaking, it is a costly process either in terms of bandwidth consumption or hardware complexity. Ironically, a full redundancy on the same network can cause congestion and consequently packet loss.

In order to support our recovery method and enhance our system reliability, a *Selective Redundancy* mechanism has been added. It consists of activating a redundancy pipe only when a packet drop is detected and maintains it only for few seconds. Unlike, the previous methods, a selective redundancy will not excessively use the network bandwidth and offer a better coexistence with other traffic.

In case of packet drops on several channels, the user can define a hierarchical resources allocation to avoid congestion and mainly guarantee the quality of critical audio channels. Selective redundancy channels are activated and managed by Resource Reservation Protocol (RSVP)[6] [93].

Multichannel audio systems can contain time-critical channels where a tight synchronization must be guaranteed. For instance in cinema, the front channels (Left, Right and Center) can be considered as time-critical, since they mainly contain dialogues, which are generally the fundamental part of the movie.

For this purpose, we introduce an additional packet format that relies on IP *multicast* capabilities [94]. If the PTP software implementation performance is insufficient, this scheme can be used to enhance the synchronization, as the packet is delivered to the multicast group members with a lower inter-delay compared to unicast. A critical stream will contain a group of channels as illustrated in figure 3.17.

Using Multicast optimizes bandwidth consumption as one copy of the message is delivered to all members. The group members number should be kept low though to maintain a good protocol efficiency (we limited it to 3). Latency is also reduced in this mode.

[6] Can eventually be replaced by the General Internet Signaling Transport (GIST) as part of the *Next Steps In Signaling* (NSIS). More information are available on the RFC 5971.

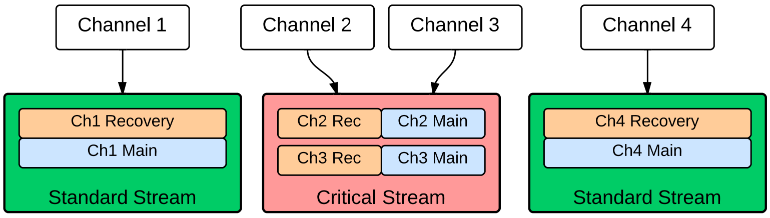


FIGURE 3.17: Time-Critical Packet structure

3.3.6 Packet building & Latency reduction

As part of the professional audio requirements, latency must be deterministic and as low as possible. The generic delay path for a network-based audio transport is illustrated in figure 3.18.

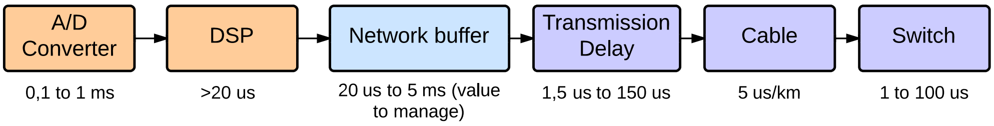


FIGURE 3.18: Latency path in a network-based audio transmission.

The main delay in our chain results actually from the buffering during the packet building phase (Fig. 3.18). A small buffer size leads to low latency, but when it comes to buffer settings, there is a trade-off to achieve between lowering latency and reducing the strain on the processing unit (e.g. CPU).

Most audio application use specific sizes for audio buffer: 64, 128, 256, 384 or 512 samples. Choosing a different value at the transmitter can introduce latency on the other side of the transmission, as the receiver will need additional buffering to adjust to the new size.

Another important factor is the size of the payload available to use. In fact, since we decided to use the standard frame size in both IP and Ethernet[7], the buffer size should be chosen in respect to the word size the available size on the packet payload.

Accordingly, we use a buffer of *128 samples* for standard streams, and *64 samples* for time-critical ones. Consequently, the latency of the standard mode is ≈ 3 ms and for the critical one ≈ 1,5 ms, including the latency per hops.

These performances can only be obtained with some effort on the hardware design, as well as the software optimization, so that the processing delay can be ignored.

[7] No Jumbo Frames are used since they are not supported by most of the existing infrastructures and no IP segmentation as well as we use IPv6.

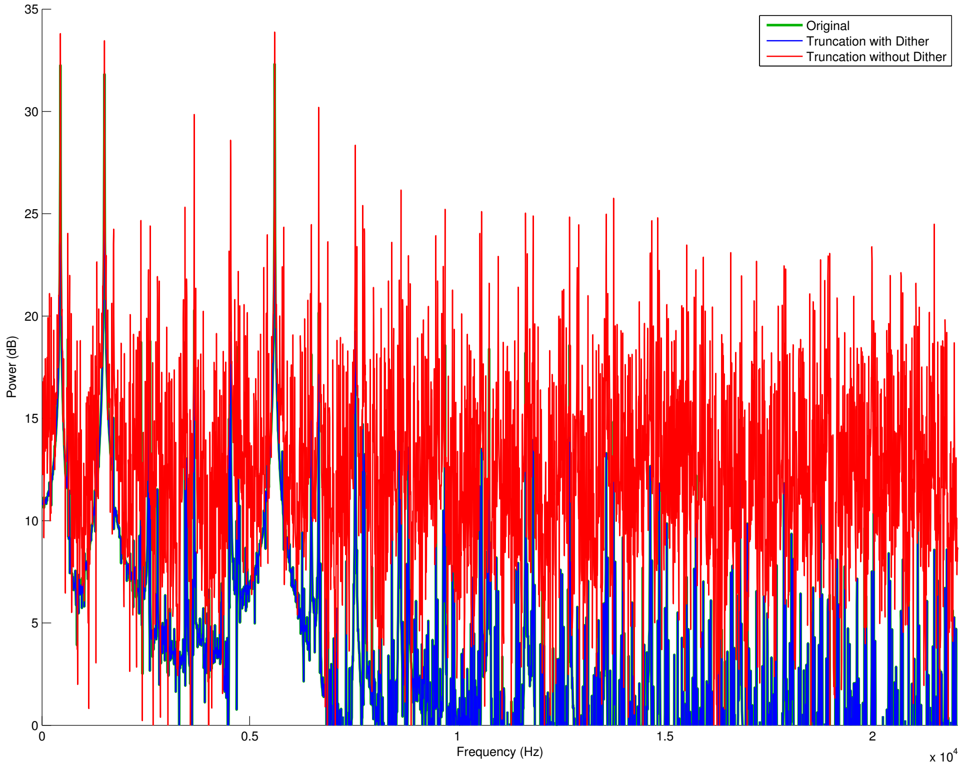


Figure 3.19: Spectrum of a three-components audio signal illustrating the Dithering effect. Adding noise before truncation compensate the re-quantification error.

For applications tolerating more latency and requiring higher buffer size, an alternative solution is suggested. In fact, we simply reduce the resolution of the recovery data, allowing more samples to fit within the packet. In this configuration, a packet contains 256 samples with $\approx$ 10 ms latency.

In this context, we studied several audio compression techniques to create the recovery part without compromising the timing constraints. The main elected ones are *Dithering* [95] and the *Opus Codec* [96].

In the Compact Disk industry, a quantization operation is frequently used to reduce the word length of audio data from the original 24-bit master recording to 16-bit before placing it on the CD. A common practice in mastering is to use dithering and noise shaping processes to prevent correlation of quantization noise with audio waveform and to reduce the perceived amount of added noise respectively [97].

Dithering is a straight-forward technique that consists of adding noise to the signal before performing the truncation. The efficiency of a dithering method depends on the noise shaping technique used to compensate the re-quantification error. Figure 3.19

illustrates the beneficial effect of dithering. For a three components signal (440, 1500 and 5600 Hz), we can see how the dithered signal approximates better the original one and lowers the quantification noise. The noise can be locally stored on a buffer, so that this operation will only cost the addition operation time.

The Opus Codec is an open source, royalty-free audio codec, defined by the RFC 6716. It is a real-time interactive audio codec designed to meet the requirements of an audio transmission through internet network [98].

Opus is composed of a two layers: the first one is based on Linear Prediction Coding (LPC) [99] and the second one is based on the Modified Discrete Cosine Transform (MDCT) [100]. The idea behind using two layers is that in speech, linear prediction techniques (such as Code-Excited Linear Prediction, or CELP) code low frequencies more efficiently than transform domain techniques (e.g., MDCT used in MP3 or Ogg Vorbis), while the situation is reversed for music and higher speech frequencies. Thus, a codec with both layers can operate over a wider range of audio signals and achieve better quality by combining the two techniques.

Opus uses SILK on the LPC layer, a low-latency codec for speech coding used by Skype [101] and the Constrained-Energy Lapped Transform (CELT) codec [102] for the MDCT layer, the one used for Ogg Vorbis compression. Both techniques can perform low-latency audio compression with delays down to 5 ms.

Despite the fact that both methods are lossy, they showed very good results during our tests. The performances of the two techniques have been evaluated using our audio quality assessment process, as presented in the next chapter.

The overall performances fits the 5 ms latency requirements for most of the professional applications and the 10 ms for live audio [75]. It should also be noted that latency is less critical in the cinema context as the video decoding add a considerable delay (digital media servers can manage up to 100 ms latency).

3.3.7 Packet Format

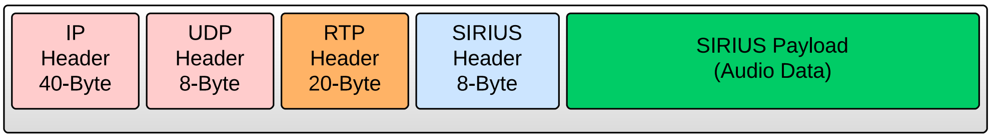


FIGURE 3.20: General Sirius Packet Format

Bearing in mind the previous considerations, a Sirius packet will contain both audio data and timing information as illustrated in figure 3.20.

We defined two types of packets: standard and time-critical which payload is respectively presented in figures 3.21 and 3.22. Since no Jumbo Frames have been used, the packet has been designed to fit within Ethernet and IP Maximum Transmission Unit (MTU).

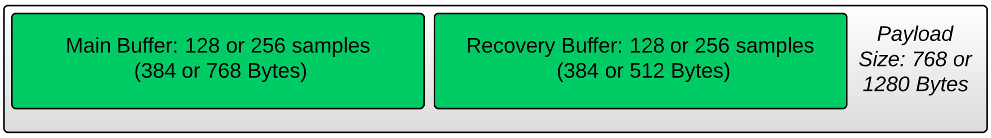

Main Buffer: 128 or 256 samples (384 or 768 Bytes)
Recovery Buffer: 128 or 256 samples (384 or 512 Bytes)
Payload Size: 768 or 1280 Bytes

Figure 3.21: Sirius Packet Format: standard mode

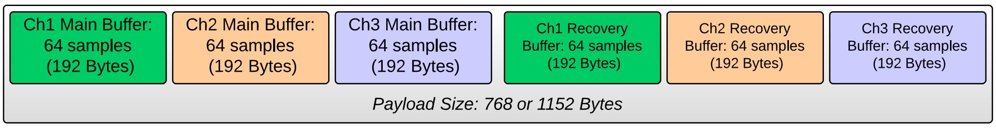

Ch1 Main Buffer: 64 samples (192 Bytes)
Ch2 Main Buffer: 64 samples (192 Bytes)
Ch3 Main Buffer: 64 samples (192 Bytes)
Ch1 Recovery Buffer: 64 samples (192 Bytes)
Ch2 Recovery Buffer: 64 samples (192 Bytes)
Ch3 Recovery Buffer: 64 samples (192 Bytes)
Payload Size: 768 or 1152 Bytes

Figure 3.22: Sirius Packet Format: time-critical mode

The header is 8-Byte long and contains the following information:

- Time Stamp of the recovery buffer, 4-Byte.
- Sampling Frequency, 6-bit (Annex).
- Audio resolution (Word Size), 2-bit.
- Packet Mode, 4-bit.
- Codec ID, 4-bit.
- Reserved 2-Byte for future uses.

More details are available on appendix B.

3.3.8 Nodes Management & Control

As mentioned before, the management capabilities are one of the main requirements for professional audio applications. A devices management system must provide two fundamental functions:

- Network Discovery: the user should be able to scan the network and identify audio devices and their capabilities.
- Control Interface: a reliable control interface must be available to send commands to audio devices and request their statuses.

Additionally, configuration-free devices with *plug-and-play* capability are really appreciated for professional applications, especially for tours and live events.

For network management and services discovery, *ZeroConf* comes as a natural choice thanks to its numerous advantages [103]. ZeroConf[8] relies on a set of technologies [55] that can be described as follows:

- The address auto-configuration process replaces the traditional Dynamic Host Configuration Protocol (DHCP) server, introducing a link-local method of addressing coupled with the mechanism of auto-configuration described in IPv6 [78] as well as in IPv4 [104].
- The name-to-address resolution replaces the need for a Domain Name System (DNS) server. This process resolves its queries using IP multicast (hence the name Multicast DNS, or mDNS) [105]. It works in conjunction with link-local addressing but is self-dependent.
- The Service Discovery [106] enables the user to find the available services over the network, replacing consequently the directory service. This process is known as DNS-SD which is compatible with mDNS.
- The last component is the multicast address allocation which replaces the need for a multicast server. The ZeroConf Multicast Address Allocation Protocol (ZMAAP) handles this process and solves most of multicast addresses conflict.

Since Zeroconf protocols are not secure as compared to standard configured network protocols, the development of lightweight security mechanisms remains an area of improvement. If needed, IPsec framework can provide such feature.

For the control aspects, we worked on adding an OSC (Open Sound Control) capability to our devices as it is the most used one in the industry. The Open Control Architecture (OCA) is also a promising solution since it is designed to ensure a high level of interoperability but still at its early stages.

OSC has the advantage of offering a simple, flexible and efficient framework for audio devices management. It uses an URL-style symbolic naming scheme with pattern

[8] *Bonjour* is Apple's commercial implementation of ZeroConf.

matching capability to specify multiple recipients of a single message, which we associated with IP multicast capability.

OSC messages consist of an Address pattern, a Type tag string, Arguments and an optional time tag. It also allows several messages to be bundled together into a single packet. We associated the time tag with our previously presented TimeStamp to ensure highly synchronized commands. An OSC-capable device can broadcast *events messages* to inform about its status. Here some examples of the OSC messages used in our implementation:

```
/Devices/Audio/SiriusNet/MyDevice1/MainMute ,T ,t (timetag)
/Devices/Audio/SiriusNet/MyDevice2/MainLevel ,f 0.7
/Devices/Audio/SiriusNet/MyDevice3/NodeSymName ,s "Room01_CenterChannel"
/Devices/Audio/SiriusNet/MyDevice4/Channels/Ch1/Level ,f 0.5 ,t (timetag)
```

The lack of a standard control protocol, as well as a lack of interoperability between the existing ones constrains the capabilities of networked application. Although it has not been fully implemented and tested in our case, we highly recommend a *proxy* approach to establish a *message translator* in order to provide communication between different protocols. The study suggested in [107] is a good example.

3.4 Results & discussion

The architecture was tested first on a simulation environment using OMNET++ IDE[9] and INET framework[10] to validate the concept.

The simulation network, as illustrated in figure 3.23, consists of a server sending RTP packets to two clients. Additionally, we added a host that generate burst data to simulate the network load. The server is also the PTP clock master.

After the network initialization, the server reads a stereo WAV file and starts the transmission. We essentially analyzed three parameters: reliability (packet loss), synchronization and latency, as presented in the next section.

The system was tested then in more realistic environment using Raspberry Pi (RPi) boards [11]. The Raspberry Pi is a credit-card-sized single-board computer running a Linux-based OS. It includes a stereo audio card as well as an Ethernet 100Mbps interface.

[9] www.omnetpp.org

[10] inet.omnetpp.org

[11] www.raspberrypi.org

The OS features also a dual-stack TCP/IP implementation (compatible with IPv4 and IPv6). The software is written in C++ and uses the standard BSD sockets and POSIX threads.

The synchronization is performed by the PTP Daemon (*PTPd*) [83]. Additionally, *Avahi*[12] daemon is used for ZeroConf management (mDNS/DNS-SD) and is fully compatible with Bonjour. Finally, the OSC interface for devices management is based on *liblo*[13] library. The network setup for the hardware-based test platform is the same as the one used for simulation (figure 3.23).

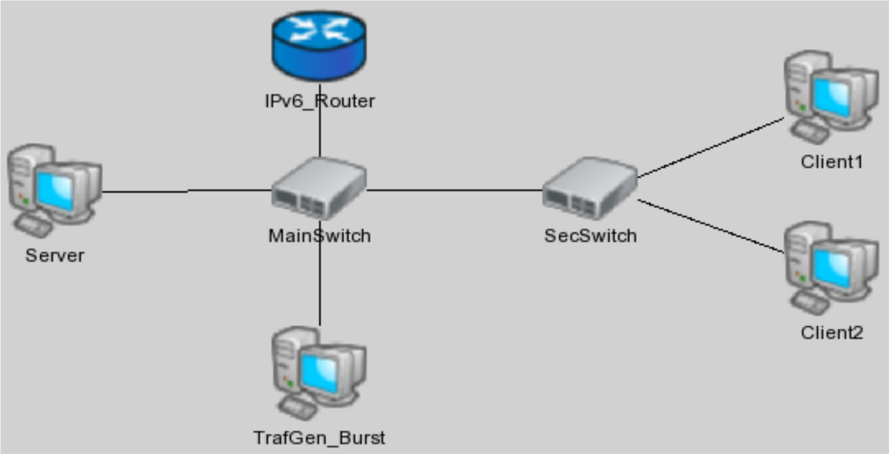


FIGURE 3.23: OmNet++ Simulation Topology.

3.4.1 Performances Analysis

The analysis is based mainly on the hardware-based test platform. The OmNet-based simulation environment has been essentially used to test the system under heavier network loads (i.e., stress-test).

Since we are in the context of a professional application, the main concern is the system reliability. Accordingly, we used the packet loss rate (PLR) as a metric to measure this feature. Figure 3.24 illustrates the average PLR measured under a normal network load. Using the recovery technique introduced in the previous section, the loss rate is lowered by ≈ 20 times. The remaining PLR is due to the fact that more than two consecutive packets have been lost, which caused the recovery failure.

Using the simulation environment, we were able to test the recovery mechanism under higher network loads as presented in figure 3.25. It shows how our method maintains

[12] avahi.org

[13] liblo.sourceforge.net

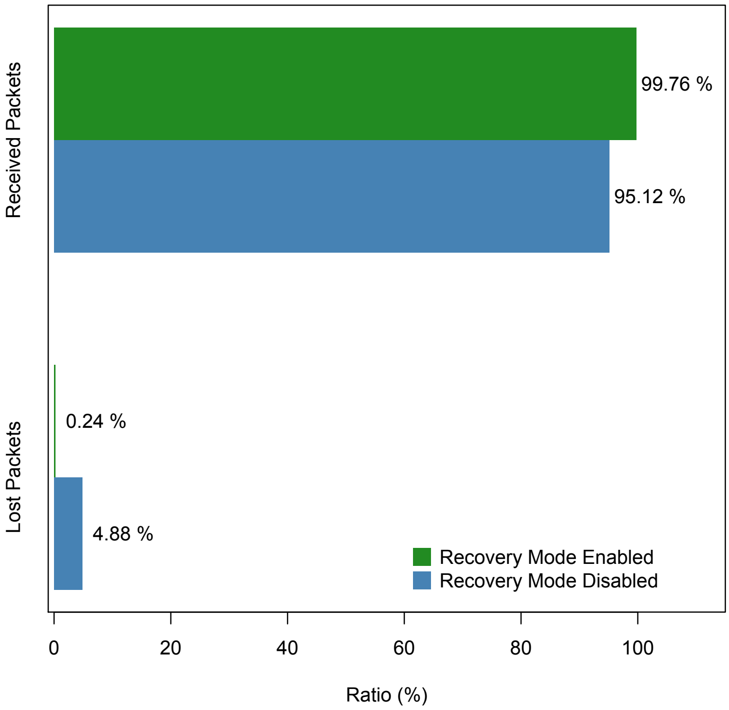


FIGURE 3.24: Network Performances: Measured Packets Recovery Ratio. Averaged over 10 mn on a normal network load.

the transmission reliability with a relatively low PLR, even in the context of a highly congested network.

The system showed also overall good performances during the subjective assessment (i.e., listening tests), as presented in the next chapter.

The second factor that we analyzed is the synchronization accuracy between the server and its clients. Using Posix time-stamping, we monitored the time offset between the clock master and a slave node, as illustrated in figure 3.26. Since the PTPd requires a convergence period, measurement starts only after ≈ 200 seconds. The synchronization accuracy roughly follows a normal distribution (mean: 2.1, sd: 1.76) μs, which matches the initial requirements, as shown in figure 3.27.

Finally, we measured the transport latency over the network, using the same approach as for synchronization. We also included the initial latency due to the audio buffer size (respectively 1,33 and 2,66 ms for 64 and 128 samples). As presented in figure 3.28, the overall latency is slightly higher than the initial value and varying because of the network hops and the changes in the network load. An additional latency is also introduced by the protocols stack but the overall performances fits the initial requirements.

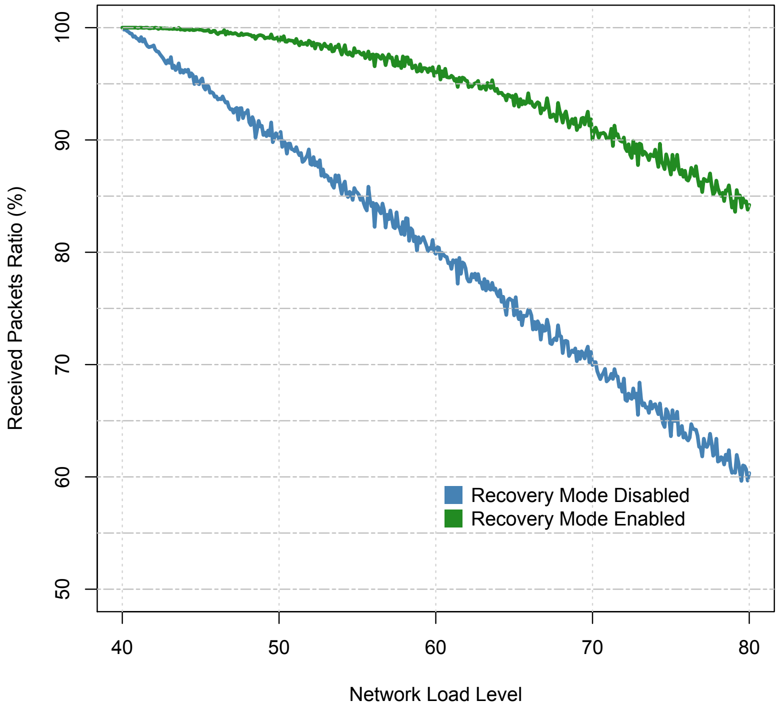


FIGURE 3.25: Network Performances: Simulation of Packets Recovery under different network loads.

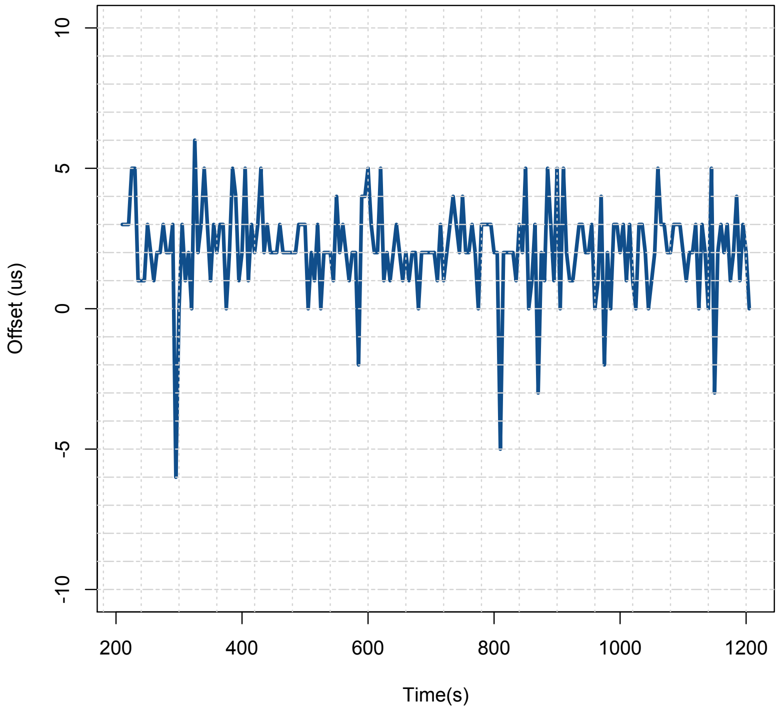


FIGURE 3.26: Network Performances: Offset measurement. The offset could only be expressed in μs due to *gettimeofday* function limitation.

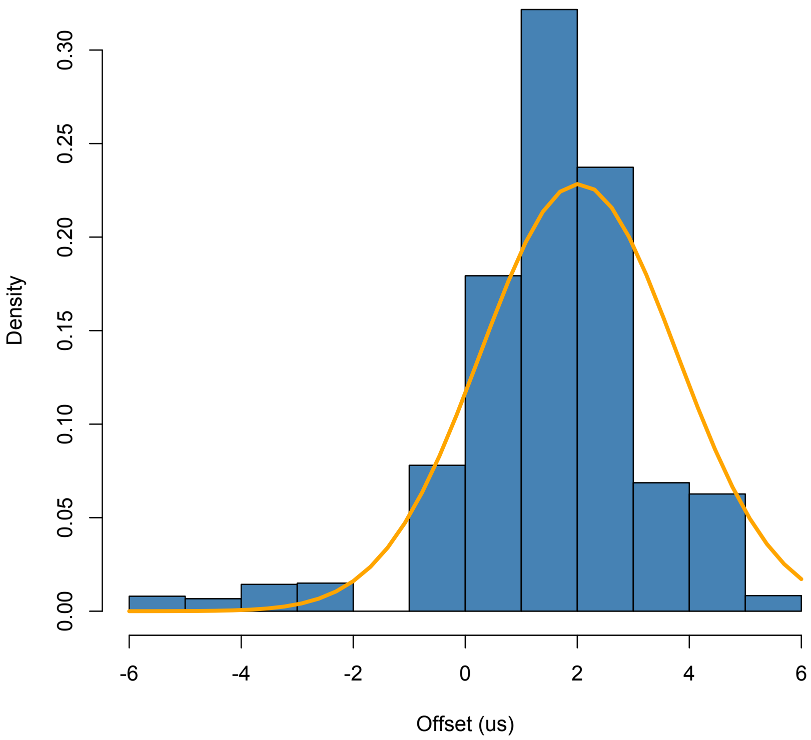


FIGURE 3.27: Network Performances: Synchronization accuracy (mean: 2.1, sd: 1.76).

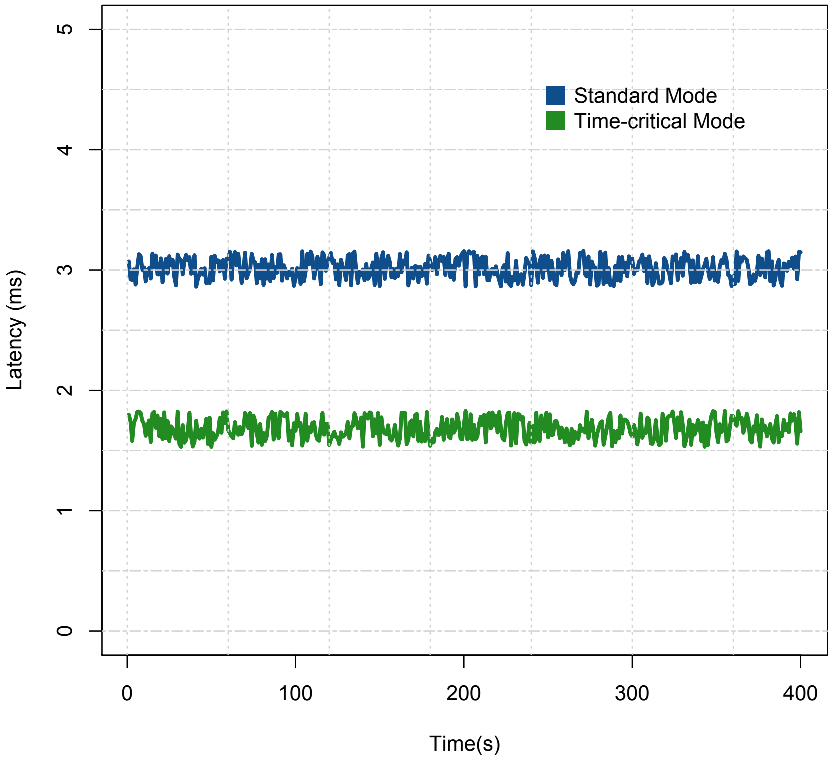


FIGURE 3.28: Network Performances: Latency monitored over time.

3.4.2 Discussion

Generally speaking, the overall system performances meet the initial requirements of professional applications, as presented in the previous section.

Although the synchronization accuracy using a software-only implementation does not match the AES11 specifications, the obtained accuracy is still below the minimum noticeable ITD.

The fact that there is currently no actual method to measure how these synchronization issues will affect the auditory display, is one of the main motivation of the spatial quality assessment technique that we present in the next chapter.

As mentioned before, the use of IPv6 allows an easier QoS management. In fact, it also helps with the latency, as a simpler header will take less time to be processed. Additionally, the *Hop Limit* field (Fig. 3.10) can be used to maintain the latency below a certain threshold.

Audio packets prioritization using QoS enhance the system reliability but cannot guarantee the packet delivery delivery, especially when the network is shared with other IT traffic. The combination of the recovery mechanism and the selective redundacy introduced in this study offers a simple solution to lower the packet loss rate without adding any latency, cost or intensive bandwidth consumption. In fact, network protocols that use aggressive retransmissions to compensate for packet loss tend to keep systems in a state of congestion even after the initial load has been reduced, which eventually and ironically leads to more packet loss.

A full redundancy (two communicating interfaces on two separated networks) can be used if higher performances are required but will induce higher cost and system complexity.

The present framework is designed to be used on standard network infrastructures and to offer a high level of interoperability. This is the reason behind using mainly standard and widely used protocols. Consequently, the current work will be submitted to be registered as an RTP profile.

As part of the study, we also investigate the possibility of using a wireless technology to transport audio. The next section highlight the main issues.

Dealing with the timing constraints of multichannel audio as well as the packet loss rate raises the question of how they actually affect the perceptual and spatial perception by the listener. This is how this part relates to the next chapter where we establish our multichannel quality assessment technique.

3.5 Wireless Transmission

As mentioned in the previous sections, audio transport in professional applications requires three main features: Reliability, Latency and Synchronization. Additionally, a high bandwidth is required in order to transport the high number of uncompressed audio channels involved.

Currently, two technologies are mainly used for wireless audio transport: BlueTooth and WiFi (IEEE 802.11). Based on the previous requirements, here is the main limitations of these technologies:

- RF interference can be generated by almost any device that emits an electromagnetic signal and may eventually compromise the system reliability.
- The timing constraints cannot be guaranteed due to the instability of the physical layer.
- The bandwidth is generally low and limited which leads to audio compression and quality deterioration (e.g. Apple AirPlay).

Moreover, as part of the playback system, the receivers will eventually need a relatively high electrical power to drive the loudspeakers (at least 100 W). Currently, wireless electricity transfer cannot provide such power, and even though, there are some coverage distance and security issues to go through.

The new generation of Bluetooth audio devices is capable of using the Advanced Audio Distribution Profile (A2DP) [108] as a transmission protocol, which if used with an advanced codec such as Apt-X, can offer a CD-close audio quality with a latency of 32 ms. It is still limited though to stereo audio only.

Also, the latest version of WiFi, namely the 802.11ac [109] holds an interesting potential thanks to its high bandwidth (theoretically 1,3 Gpbs, actual performance is closer to 250-300 Mbps), its low latency compared to previous versions[14], and its capacity to cover a wider range[15] [110]. The study in [111] is a good example of the potential use of WiFi for high quality audio transport.

[14] Thanks to the use of the 5 Ghz band instead of the 2.4 Ghz.

[15] 802.11ac uses BeamForming which detects where the devices are and intensifies the signal in their directions.

3.6 Deployment Optimization

As mentioned before, the suggested hierarchical tree topology offers a high level of scalability. It also the basis of our deployment optimization algorithm in the context of static setup such as cinemas, as presented below.

The previous topology is equivalent to the graph in figure 3.31 where:

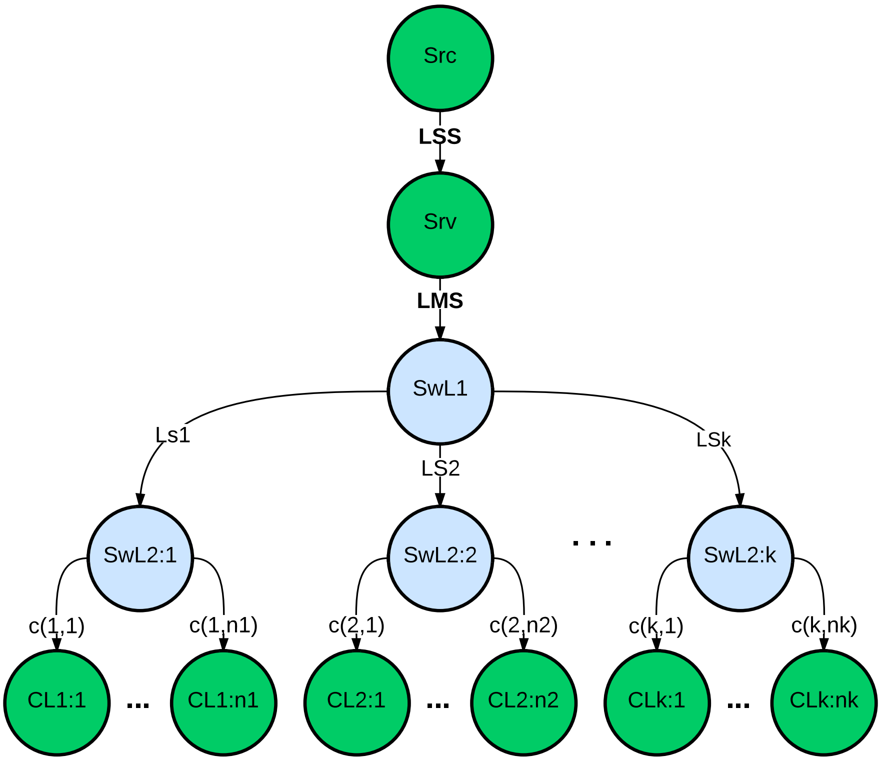


FIGURE 3.29: Deployment Optimization Basis

- *Src*, the multichannel audio source (e.g. Media Server).
- *Srv*, the Network Server, it is the actual audio stream source.
- *LMS*, the cable Length between the Server and SwL1, the first level switch (Main switch).
- $L_s(i)$, the cable Length between the main switch and the i-th secondary switch (K secondary switches).
- $c(i, j)$, the cable Length between the i-th secondary switch and the j-th client connected to it (n_K clients per secondary switch).

Accordingly, the Total cable Length (TL) is:

$$TL = L_{MS} + \sum_{i=1}^{K} L_s(i) + \sum_{i=1}^{K} \sum_{j=1}^{n_K} c(i,j) \tag{3.11}$$

The Media Server and Network Server are generally parts of the same device or mounted on the same rack. Consequently, The length of the cable linking them, L_{SS}, can be ignored in comparison to the other cables lengths. The model inputs are the spatial locations of each audio node as follow:

- Clients coordinates, $C_n(x,y,z)$, depend on the loudspeakers layout.
- Server position, $P(x,y,z)$, depends on the rack position in the projection room.
- Available switches number, K>1.

Accordingly, this algorithms aims to estimate the optimal positions for the switches to minimize the total cable length. As a results, the model outputs are the switches coordinates, $S_i(x,y,z)$, where $S_0(x,y,z)$ are the coordinates of the Main switch.

The approach is based on the fact that most of the cable is used to connect secondary switches to the clients. By putting a switch as close as possible to a group of neighbor clients, cables $c(i,j)$ will get shorter. By placing the main switch as close as possible to the secondary ones, cables $L_s(i)$ get shorter too. Eventually, it comes to increasing the cable L_{MS}, linking the server to the main switch but the overall gain is more interesting.

The mathematical gives the following:

$$\begin{aligned}
\underset{S_i(x,y,z)}{Argmin}(TL) &\Rightarrow \underset{S_i(x,y,z)}{Argmin}(TL^2) && (3.12)\\
&\Rightarrow Argmin \sum_{i=1}^{K} L_s(i)^2 + Argmin \sum_{i=1}^{K} \sum_{j=1}^{n_K} c(i,j)^2 && (3.13)\\
where \quad c(i,j) &= \ell(S_i(x,y,z), C_n(x,y,z)) && (3.14)\\
and \quad L_s(i) &= \ell(S_i(x,y,z), S_0(x,y,z)) && (3.15)
\end{aligned}$$

For this kind of installations, the length function ℓ is generally a *taxicab* distance (d_1) as described in equation 3.16 [112]. Thanks to the *triangle inequality* (eq. 3.17), we can simply consider the *euclidean* distance for minimization.

$$d_1(p,q) = \|p-q\|_1 = |p_x - q_x| + |p_y - q_y| + |p_z - q_z| \tag{3.16}$$

$$d_1(p,q) \geq d(p,q) \quad where \quad d(p,q) = \sqrt{(q_x-p_x)^2 + (q_y-p_y)^2 + (q_z-p_z)^2} \tag{3.17}$$

Using Euclidean distances, we get the following:

$$Argmin \sum_{i=1}^{K} L_s(i)^2 \Rightarrow Argmin \sum_{i=1}^{K} d(S_i, S_c)^2 \tag{3.18}$$

$$Argmin \sum_{i=1}^{K} \sum_{j=1}^{n_K} c(i,j)^2 \Rightarrow Argmin \sum_{i=1}^{K} \sum_{j=1}^{n_K} d(C(i,j), \mu_i)^2 \tag{3.19}$$

where S_c is the centroid of points $S_{1 \ldots K}$ and μ_i is the centroid of the i-th group of points $(C(i, j = 1 \ldots n_K)$.

Accordingly, equation 3.19 refers to the well-known *K-means* clustering algorithm [113]. K-means clustering aims to partition the n observations into k sets ($k \leq n$) S = S1, S2, . . . , Sk so as to minimize the within-cluster sum of squared distances. More details are available on [114].

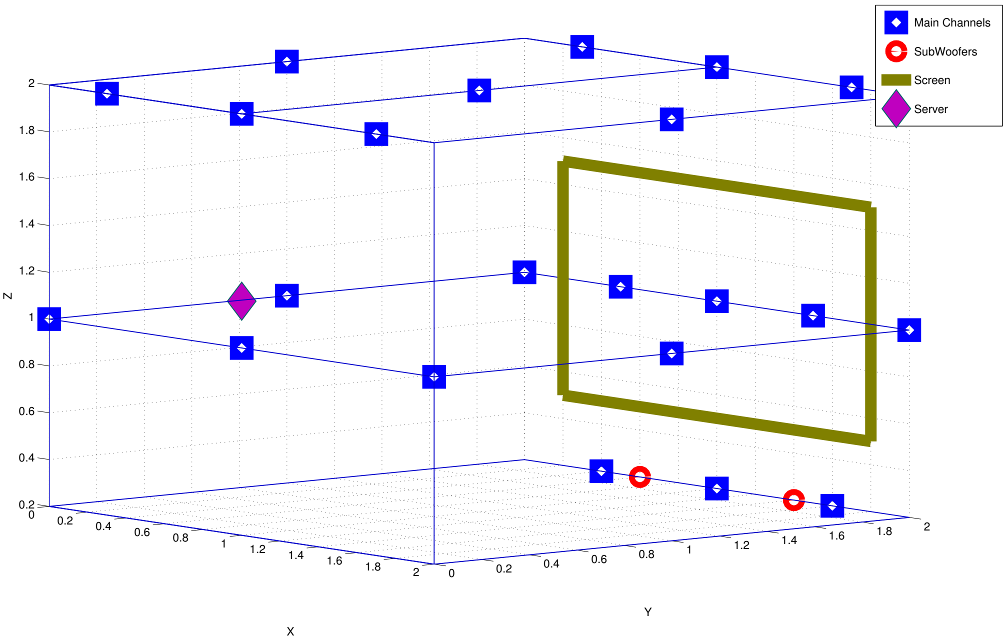


FIGURE 3.30: NHK 22.2 loudspeakers setup. The blue squares circle the three layers of the system.

In the case of static setups, the cables are generally passed through cable trays and not freely placed. Consequently, the results are most likely not applicable since they are based on euclidean distance. To resolve this, the calculated coordinates should be projected on the nearest plane and then on the nearest tray. Eventually, if more data on the room configuration and constraints are available, the current model should be adapted to fit that. If the Main switch is movable, the optimal position will be as close as possible to the centroid of secondary switches (eq. 3.18).

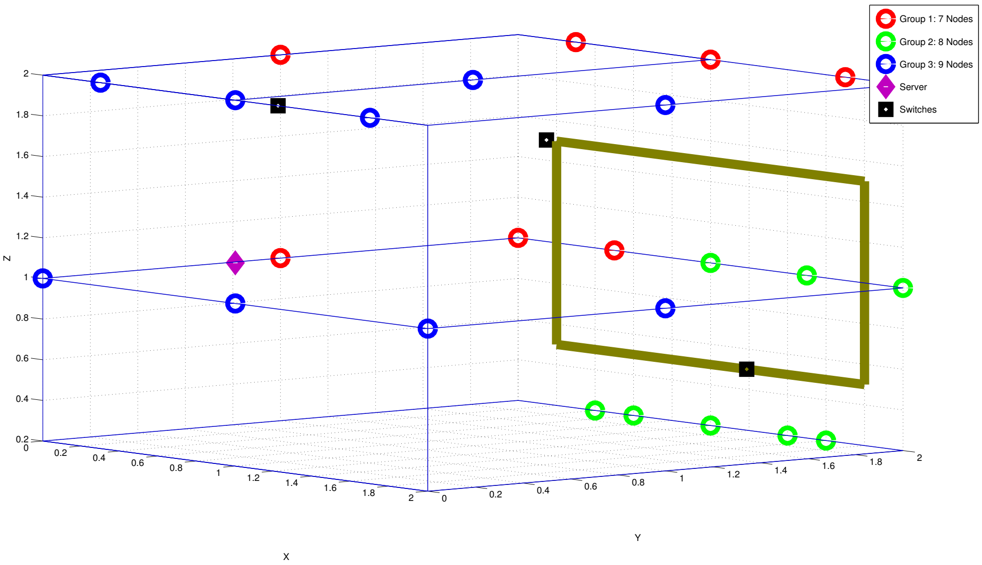


FIGURE 3.31: Optimization algorithm results for the NHK 22.2 setup

Figure 3.32 shows the algorithm steps. The example in figures 3.30 and 3.31 illustrates the NHK 22.2 loudspeakers setup and how the algorithms estimates the switches coordinates. For this layout, more than 37,5 % cable gain was established using this method.

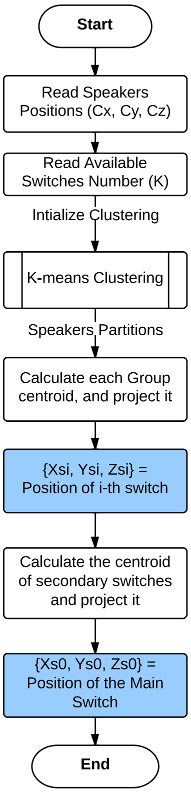

Start
Read Speakers Positions (Cx, Cy, Cz)
Read Available Switches Number (K)
Intialize Clustering
K-means Clustering
Speakers Partitions
Calculate each Group centroid, and project it
{Xsi, Ysi, Zsi} = Position of i-th switch
Calculate the centroid of secondary switches and project it
{Xs0, Ys0, Zs0} = Position of the Main Switch
End

FIGURE 3.32: Deployment Optimization Algorithm

Chapter 4

Quality Assessment

4.1 Introduction

The recent years have witnessed a rising interest towards multimedia applications across the globe with an increased concern about the content quality, from the consumer's side as well as equipment manufacturers and content providers. This led to the foundation of an entire branch of applications dealing with quality assessment and enhancement techniques, as it became vital in the design and deployment of multimedia services.

Furthermore, with the current trend of multimedia technologies combining 3D visual and auditory stimulus, the spectators can be completely immersed in the scene and even capable of interacting with it. Consequently, they become more sensitive and demanding when it comes to the quality of the content and the realism of the scene.

Although 3D audio technologies have existed for a while now, only a few works have attempted to establish efficient quality measurement techniques to assess both the perceptual and spatial aspect of multichannel 3D sound.

Generally speaking, quality assessment methods can be classified into two main categories: *Objective* methods where the combination of specific metrics based on perceptual models, helps getting a quality grade without a human assistance, and a more intuitive one which is the *Subjective* methods, where the quality level is based on the statistical analysis of the opinion of a representative group of human subjects, who, according to the application, might be experts or non-experts (i.e. Listening tests).

On the one hand and largely speaking, formal listening tests are more reliable if they are conducted properly. Their main and constraining limitation is that they are time and resources consuming. On the other hand, objective methods resolve this issue and are even the only applicable solution in the case of a time sensitive evaluation (e.g., real-time audio quality monitoring). The critical aspect of the latter approach is its reliability, which is defined by its correlation to a subjective reference. In other words, how well it *mimic* the human perception. This shows the inescapable duality linking the two approaches, and which unfortunately, is not always remembered.

The existing quality assessment techniques do not cover all the characteristics of multichannel sound. In fact, most of them deal basically with perceptual information while overlooking spatial and psychological aspects. This is actually what motivates this part of our work and defines our approach. By going back to the *natural* phenomena involved in quality evaluation, we study how people interact with sound and its quality in order to mimic it in the most accurate and realistic way.

To cover these aspects, the layout of this chapter will be as follow: the second section presents an overview of the existing standards, mainly the ITU recommendations, and the recent studies to improve them. The next one introduces our method that combines a subjective and objective approach to improve and enhance the existing methods. The main aspects of our experimental framework are presented in the fourth section with the results of our method, while the last one present a conclusion of our study and our future directions in multimedia quality assessment and enhancement techniques.

4.2 Literature Review

There has been several good studies and papers that review the ITU standards for audio quality assessment and the later studies to improve them [115–119]. In this section, we present a brief survey of the most used ones and highlight the missing aspects regarding the analysis of spatial and psychological information within multichannel sound.

4.2.1 Subjective Assessment

As mentioned before, formal listening tests are the most efficient way to get reliable results if conducted correctly. Moreover, even the objective methods require subjective data as a reference to evaluate their performance and correlation with the human perception. This shows how important and inescapable listening tests are, despite their inconvenient constraints.

In this context, the ITU issued few recommendations to design and conduct proper listening tests. The ITU-R BS. 1284-1 standard [120] presents the general methods for subjective assessment and refers to two main subsets according to the level of the studied impairments. They are described by the following recommendations:

- Rec BS.1534-1, as defined in [121], describes a method for the subjective assessment of intermediate quality level for coding systems, also know as MUltiple Stimuli with Hidden Reference and Anchor or MUSHRA test. As the name suggests, this approach is more suitable for intermediate impairment evaluation by using a double-blind multi-stimuli with hidden reference and hidden anchor(s). The subject is presented with a sequence of trials. In each trial the subject is presented with the reference version as well as all versions of the test signal processed by the systems under test, and is allowed to switch instantly among them. The listener is asked to score the stimuli according to the continuous quality scale (CQS) using five levels (Bad, Poor, Fair, Good and Excellent) for an overall scale from 0 to 100. A grade of 0 is the lowest of the "bad" category, while a grade of 100 is the highest of the "excellent" category.

- Rec BS.1116, as described in [122], is more appropriate for the assessment of small impairments and relies on double-blind triple stimulus with a hidden reference. It is also known as the ABC method since it uses three stimuli "A", "B" and "C". The known reference is always available as "A" while the hidden reference and the studied object are randomly assigned to "B" and "C". The subject is asked to assess the impairments on "B" compared to "A", and "C" compared to "A", according to the ITU continuous five-grade impairment scale as shown in table 4.1.

Table 4.1: ITU Grading Scale

Grade	**Impairment**	**Quality**
5.0	Imperceptible	Excellent
4.0	Perceptible but not annoying	Good
3.0	Slightly annoying	Fair
2.0	Annoying	Poor
1.0	Very annoying	Bad

These recommendations deal basically with perceptual impairment since they are designed to describes only one feature: the annoyance level. Considering that the evaluation of audio quality involves both *sensory* and *affective* judgements, the term *annoying* will reflect mainly and only the affective part. This factor is also highly subjective [123], very general (confusing) and cannot inform about specific aspect like the spatial distortion. Therefore, it is not efficient for a thorough audio evaluation.

In fact, previous studies like Letowski's [124] showed that the audio quality does not depend only on the timbral characteristics. Accordingly, the MURAL model (fig. 4.1) clusters the attributes that influence the audio quality in three main categories: *Timbral*, *Spatial* and *Technical*, and which combination is responsible for the overall perceived quality [125].

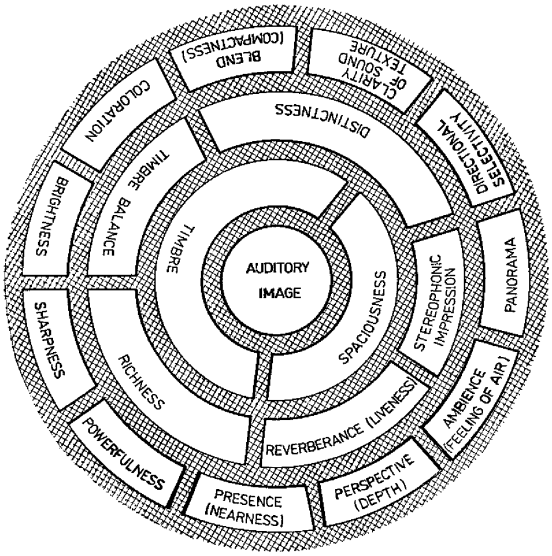


Figure 4.1: Letowski's MURAL Model. Credits: [124]

On that basis, Rumsey and Berg studied extensively how to improve the listening tests by including the missing features. The approaches presented in [125, 126] are based on the rating of additional spatial attributes. Additionally, and in order to reduce the difference in interpretation of such descriptors, the process of evaluation is divided into three main phases:

- During the first one, the subjects express individually the perceived spatial sensations in a verbal form instead of the experimenter.
- The second phase consists of the analysis and clustering of the verbal responses which will lead to the generation of attribute scales.
- Finally, the generated *customized* scales are used to rate stimuli.

More studies have been conducted in order to supply tools to analyze such features. Paquier et al. in [127] suggest a modified version of the MUSHRA test that consists of two sessions. The first one focus on the evaluation of the Overall Quality (OQ) and the second one deals with the assessment of three families of attributes: *Timbre* (T), *Space* (S) and *Defects* (D), where the subjects evaluate each one separately. Their work led to a *regression* model linking the Overall Quality to the other attributes as described by equation 4.1:

$$OQ = 0,65 \times D + 0,44 \times S + 0,3 \times T - 0,32 \tag{4.1}$$

Although this approximation cannot be generalized since it is only valid in the context of their study, the resulting model shows that, in addition to the perceptual attributes, other factors do influence the perceived overall quality.

As a general observation, verbal descriptors are not the most efficient way when it comes to spatial audio quality analysis as highlighted in [125, 128, 129]. In fact, and unless it is supervised, subjects tend to express their spatial perception in a very subjective and inaccurate way. Additionally, as mentioned in numerous studies and summarized by Zielinski et. al. in [130, 131], the very foundation of the existing listening tests may lead to some noticeable biases, considering the following observations:

- Audio quality assessment involves generally both sensory and affective judgments. In fact, most of the verbal expressions used in subjective assessment exhibit an affective nature, but the analysis of the results concludes generally only on the sensory judgment. Personal preference and taste, expectations, emotions and even the mood, are all influencing factors that are unfortunately not always considered in audio quality evaluation.
- The subjects task during a listening test is not only to judge the quality of an audio sequence but also implies to translate their internal judgments into some form of response, such as a numeric grade or the position of a cursor on a graphical scale. The mapping process of the internal judgment onto external response is generally not linear and eventually induce some considerable biases [130].
- The design of test interface may also lead to an additional bias. In fact, the early versions of the listening tests involved analog tape-base playback systems that presented some restrictions (randomization for instance), that were resolved by new interactive computer-based interfaces. However, as the tests complexity increases, the interfaces designed for multiple stimulus tests can get quite *confusing* if overloaded with graphical objects. Eventually, it may lead to mistakes, fatigue and ironically even to annoyance which will add substantial biases.

As mentioned above, the existing techniques lack the analysis of spatial information within multichannel 3D audio streams. In fact, the ITU standards above only mention the setup of the loudspeakers for 5.1 or 7.1 configurations but without any further information about how to measure and analyze the spatial cues within the audio signal.

Zielenski and Rumsey addressed this issue and suggest their Quality Adviser System in [132]. It consists of two softwares: Predictor A, that is used for prediction of audio quality of band-limited multichannel audio recordings and Predictor B for the quality of down-mixed multichannel audio recordings. The combination of the two aims to help audio engineers to make their decision regarding the design of a broadcast strategy. The limitation is that it is based on the MUSHRA protocol where the subjects will have to report spatial distortions on a GUI with sliders referring to the 100-point scale, which ultimately implies potential biases, as explained above.

Since the purpose of assessing audio quality is to predict and enhance the quality of experience for *human* end users, assessment techniques should be centered on the *listener* and *mimic* as close as possible the human perception and behavior. Additionally, for realistic results, the tests conditions, materials and results analysis should be as close as possible to the real conditions of the considered applications. This is actually the precept of our analysis and work to establish a new comprehensive approach for multichannel audio quality assessment, as explained later in this study.

Due to the complexity and cost of formal listening tests, several studies have been conducted to establish automatic and objective methods to evaluate audio quality. The next section present a brief summary and analysis of the most used ones.

4.2.2 Objective Assessment

As stated before, human subjective listening tests are expensive and time consuming since they require a large number of (preferably) trained human listeners and sometimes involve specific equipment. In this context, objective methods are more appealing as they offer a fast and efficient tool to predict the perceived quality of an audio system. In fact, they might be the only usable solution in the context of a time critical application (e.g. real-time quality monitoring).

The *PEAQ* method (Perceptual Evaluation of Audio Quality) or the BS.1387 is the ITU (International Telecommunications Union) standard for objective audio quality assessment [133]. It relies on a full-reference approach where audio quality changes are evaluated by comparisons between a reference audio signal and an impaired version of it (the output of the Device Under Test (DUT)), as shown in figure 4.2.

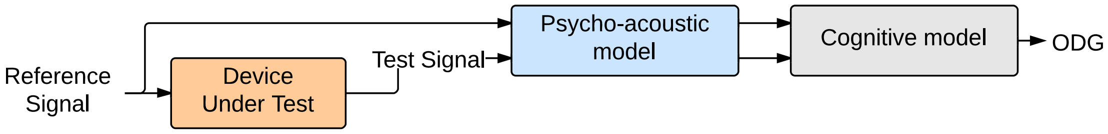


FIGURE 4.2: PEAQ principle

The PEAQ standard has been introduced in 2001 and is based on seven previously developed models:

- Disturbance Index (DIX) [Thiede and Kabot, 1996]
- Noise-to-Masked-Ratio (NMR) [Brandenburg, 1987]
- Objective Audio Signal Evaluation (OASE) [Sporer, 1997]
- Perceptual Audio Quality Measure (PAQM) [Beerends and Stemerdink, 1992]
- PERCeptual EVALuation (PERCEVAL) [Paillard et al, 1992]
- Perceptual Objective Measure (POM) [Colomes et al, 1995]
- The toolbox approach, based on [Zwicker and Feldtkeller, 1967]

As illustrated in figure 4.2, PEAQ method consists of a psychoacoustical model and a cognitive model. The first one extract some perceptual features that reflect the internal representation at the basilar membrane. They are combined then at the second stage to establish an Objective Difference Grade (ODG) that inform on how the DUT affects the perceived quality of the sound signal. PEAQ exists in two versions: the basic version which is computationally efficient with a general good accuracy and the advanced version which is more accurate but slower (four times more computationally demanding) [115].

The main structural difference between the two versions is that the basic one has only one peripheral ear model (FFT based ear model) whereas the advanced one combines two peripheral ear models (FFT based and filter bank based ear models). The Basic Version produces 11 MOVs (Model Output Variable) whereas the Advanced Version produces only 5 MOVs. The MOVs are output features based on loudness, modulation, masking and adaptation as shown in tables 4.2 and 4.3. A block diagram of the two version is shown in figure 4.3.

The Cognitive model comes next to map the MOVs to the ODG score, the quality level estimation. It is based on a one hidden layer artificial neural network and the optimization was done using the back-propagation algorithm. Considering the dataset used in the training, the ODG should reflect the subjective difference grade (SDG) which

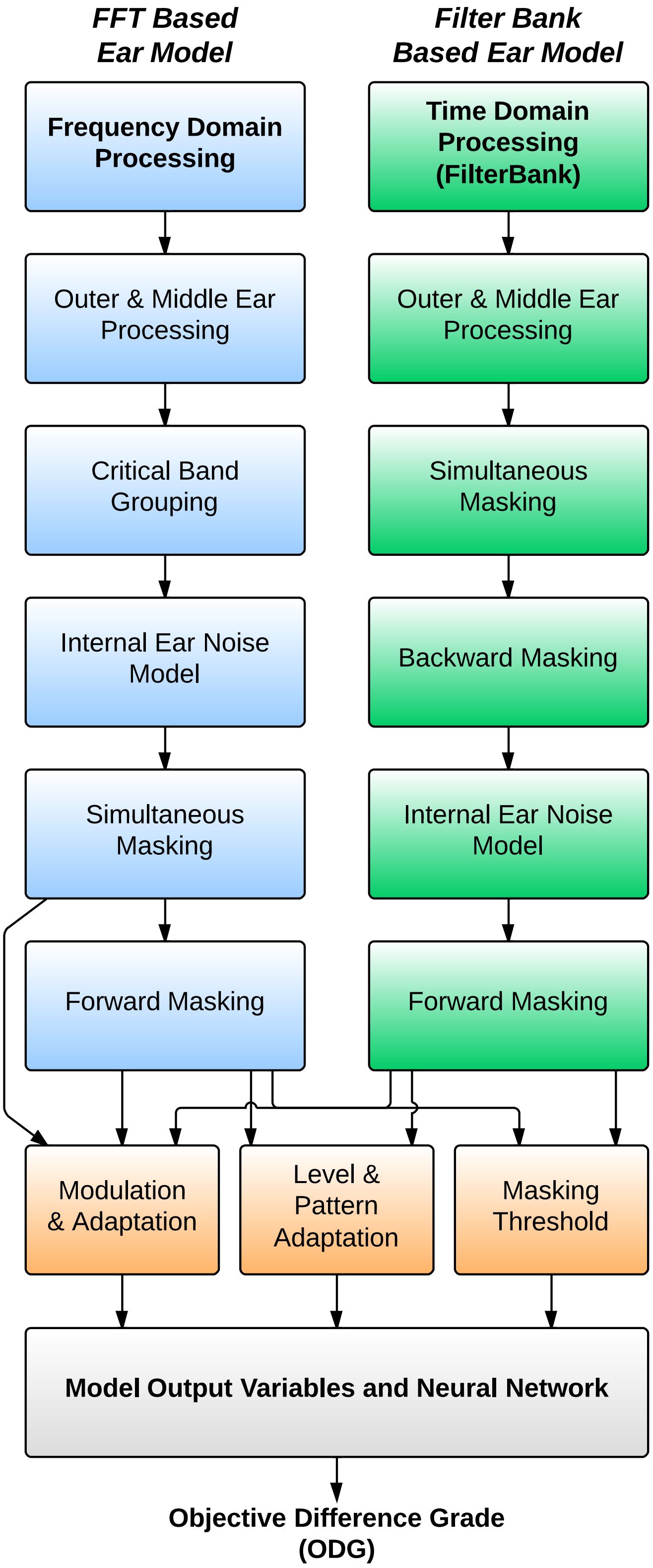

FFT Based Ear Model
Filter Bank Based Ear Model
Frequency Domain Processing
Time Domain Processing (FilterBank)
Outer & Middle Ear Processing
Outer & Middle Ear Processing
Critical Band Grouping
Simultaneous Masking
Internal Ear Noise Model
Backward Masking
Simultaneous Masking
Internal Ear Noise Model
Forward Masking
Forward Masking
Modulation & Adaptation
Level & Pattern Adaptation
Masking Threshold
Model Output Variables and Neural Network
Objective Difference Grade (ODG)

FIGURE 4.3: A block diagram of the two versions of PEAQ.

Table 4.2: MOVs of the Basic version of PEAQ

MOV	**Description**
WinModDiff	Windowed averaged difference in modulation (envelopes) between Reference Signal and Signal Under Test
AvgModDiff1	Averaged modulation difference
AvgModDiff2	Averaged modulation difference with emphasis on introduced modulations and modulation changes where the reference contains little or no modulations
RmsNoiseLoud	Rms value of the averaged noise loudness with emphasis on introduced components
BandwidthRef	Bandwidth of the Reference Signal
BandwidthTest	Bandwidth of the output signal of the device under test
TotNMR	logarithm of the averaged Total Noise to Mask Ratio
RelDistFrames	Relative fraction of frames for which at least one frequency band contains a significant noise component
AvgSegmNMR	the Segmentally Averaged logarithm of the Noise to Mask Ratio
MFPD	Maximum of the Probability of Detection after low pass filtering
ADB	Average Distorted Block (=Frame), taken as the logarithm of the ratio of the total distortion to the total number of severely distorted frames
EHS	Harmonic structure of the error over time

Table 4.3: MOVs of the Advanced version of PEAQ

MOV	**Description**
RmsModDiff	Rms value of the modulation difference
RmsMissing-Components	Rms value of the noise loudness of missing frequency components, (used in RmsNoiseLoudAsym)
RmsNoise-LoudAsym	RmsNoiseLoud + 0.5 RmsMissingComponents
AvgLinDist	A measure for the average linear distortions
Segmental NMR	It is the same as Total NMR in the Basic Version. It is the local linear average.
EHS	same as the EHS for the basic version

finally informs about the audio quality level of the system under test. The ODG score ranges from 0 to -4 where 0 represents a signal with imperceptible distortion and -4 represent a signal with a very annoying one.

Since its standardization, PEAQ has been throughly studied. Campbell et al. presented in [115] a survey of the techniques that have been introduced to remedy to the missing features of the ITU standard. Generally speaking, PEAQ suffers from the following limitations:

- The efficiency of PEAQ's perceptual model has been discussed and showed some inconsistencies with several types of impairments [115, 134].
- Since this standard has been designed for low to moderate impairments, it has been shown that it behaves poorly in the context of highly impaired audio [134, 135].
- The PEAQ standard does not handle multichannel audio and cannot analyze the spatial information within [115, 136].
- Their is no real consideration of psychological behavior in the ITU standard as the PEAQ has not been designed to cover such aspect.
- The efficiency of the cognitive model is debatable and can be noticeably enhanced [135].

In order to mitigate some of these aspects, Huber suggests in [134] a new full-reference method called PEMO-Q[1]. As a matter of fact, the auditory model used in this approach is more accurate since it is based on a more precise and realistic *Gammatone* filter bank [137]. The signal envelope is analyzed then through a linear modulation filter bank. The linear cross correlation coefficient of the assimilated *internal representations* of the two signals (reference and test) represents the perceptual similarity measure (PSM) which reflects the difference in the quality level. The study shows also a good correlation between the PSM and the measured data during the subjective validation [134].

The main advantage of Huber model is its ability to predict very small as well as more severe quality degradations for different types of audio signals and signal distortions except for linear distortions. In addition, the cognitive model is much simpler than the ITU standard and performs slightly better than it. Although and as expected, it is less computationally efficient than the advanced version of the PEAQ, which is already slow, due to the correlation computation. Also, as it includes a time and level alignment phase, it eventually loses the spatial information within the signals, as explained later.

[1] A demo version is available at: www.hoertech.de

Moreover, Cave suggests in [138] a novel auditory model based on previously developed masking models that attempt to overcome some apparent problems with these models, including the one used in PEAQ. It states that the Sound pressure level (SPL) in PEAQ should accurately reflect the level presented to the ear, independently of the frequency resolution of the auditory model. However, with PEAQ this is not the case as it normalizes the spectrum according to a single frequency component. Once the spectrum is normalized in this way, the SPL of a given frequency band is obtained by calculating the sum of all the components in that band, and is somewhat sensitive to the frequency resolution in PEAQ. The SPL should be set independently of the frequency resolution in order to give a more accurate representation of its true level.

Cave also indicates that PEAQ is one of the few auditory models to account for the additivity of masking, which is based in PEAQ, on relatively simple spreading functions. He raises questions about the accuracy of these functions. Cave suggests that noise maskers should be integrated over a complete critical band, whereas PEAQ attempts to increase its resolution by using bands that are fractions of critical bands. This is undesirable as it affects greatly the masking effects. He also claims that the forward masking model in PEAQ is inaccurate since it uses a low pass filter which fails to account for the fact that components in previous frames may also be present in the current frame, and that it is important to consider the boundaries of the maskers and the position of the masked.

To overcome these issues, Cave developed a new auditory model that was implemented for audio coding applications but not for audio quality assessment. In his model he calculates a SPL level that overcomes the problems in relation to inaccurate SPL levels. The model also accounts for tracking the temporal maskers from frame to frame and includes boundary detection to overcome the lack of accuracy in PEAQ's masking model. Thus far, Cave's model has only been used in audio coding applications but can also be applied to audio quality assessment. He tested the model by means of an audio coder test bed against the PEAQ auditory model. The PEAQ based model outperformed his model for speech coding, but not for audio coding as the novel auditory model appeared to give improvements over PEAQ according to his subjective listening tests. The model could replace the one in the FFT-based ear model, or at least some of its concepts could be considered for audio quality assessment [115].

On a different aspect, Creusere et. al. suggest in [135] to combine their previously developed Energy Equalization Algorithm (EEA) with the PEAQ. They found that by including the EEA as an additional MOV in the advanced version, a salient improvement can be obtained. The EEA approach is based on the fact that the perceived quality of an audio signal is considerably affected when an isolated segment of time-frequency energy are formed, mainly around 2 to 4 kHz. The algorithm uses the number of time-frequency segments, referred to as *islands*, as a measure of quality, grading the signal with highest number of energy islands as much lower quality compared to the one having less of them [135, 139]. The cognitive model has been also improved by using a single layer neural network (no hidden one). The correlation between subjective and objective scores suggests that this modified version of PEAQ outperforms the original one [115, 139].

Still, we noticed some limitations of this method, particularly in the case of linear amplification that will lead to false positives while it is not subjectively detected. In addition, Packets drops are not detected either by this method where they are easily spotted during a listening test (annoying clicks).

Since the methods above do not handle the spatial information within multichannel audio, more studies have been conducted to overcome this aspect. Dewhirst et. al. presented in [140], their approach for objective assessment of spatial localization attributes of surround-sound reproduction systems. Using a microphone array and an acoustical mathematical model, they were able to simulate the signal across different reproduction systems in order to analyze how a specific system affect the spatial perception. The model allows a graphical representation of three different objective measures of spatial sound quality: *directional localization*, *ensemble width* and *ensemble envelopment*. Though, it has not been validated on concrete subjective listening tests and it could only handle monochromatic time-invariant signals which is far from the case of realistic audio signals.

In the same context, Choi et. al. propose in [136] a spatial extension for PEAQ based on its advanced version. Three additional MOVs based on the interaural differences are derived from the binauralised multichannel audio: ITD (Interaural Time Difference), ILD (Interaural Level Difference) and IACC (Interaural Cross-Correlation Coefficient) and combined with the five MOVs to establish a quality metric capable of analyzing both perceptual and spatial distortions. The approach has been updated in [141] by introducing an additional cue, the EITD (Envelope Interaural Time Difference Distortion) which increases the overall accuracy. Still, the validation process is questionable since it was based only on the rating of the overall quality grade using the Basic Audio Quality scale as shown in table 4.1. Moreover, the model was trained and calibrated using a limited selection of audio process types (multichannel audio coding, bandwidth limitation and downmixes) which results of covering only low to moderate audio impairments [141].

Rumsey et al. as part of their QESTRAL project (Quality Evaluation of Spatial Transmission and Reproduction using an Artificial Listener) [142] extended the work conducted in [140] to create a more advanced *Artificial Listener*. It primarily relies on specific measurements of the reproduced sound field made at one or more listening positions, using binaural and microphone-derived signals. [119, 143]. Once again, the validation still debatable as it relies on attributes based on verbal expression. Also, the fact that it is based on actual physical measurements might be constraining for some applications.

For these reasons, the ITU initiated in 2008 a working group (Study Group 6) to investigate a possible revision of PEAQ, focusing mainly on the quality assessment of multichannel sound and intermediate quality audio. No recommendations have been published yet.

Another important aspect that has not been thoroughly covered by the previous studies is the impact of psychological factors during listening tests. The absence of such considerations may introduce a significant bias and compromise the efficiency and realism of the objective methods. In fact, we believe that such aspect should be inherited by objective measurement in order to get a more reliable and realistic assessment. For instance, the experience effect as we introduced in [144] could be modeled and taken into account during an automatic continuous quality evaluation to enhance its performance.

Considering all these aspects, we aim to establish a new comprehensive method that will provide the following features:

- An efficient assessment of the perceptual quality.
- The compatibility with multichannel sound and the analysis of spatial cues.
- An optimal computation efficiency to enable real-time implementation.
- The integration of behavioral aspects in order to establish realistic results.

Table 4.4 presents a summary of the methods mentioned above. The next section will present our subjective and objective approach to enhance the existing techniques.

Table 4.4: Summary of Objective Audio Quality Assessment Techniques

Method	**Concept**
PEAQ [133]	ITU Standard for Perceptual Evaluation of Audio Quality, combining a psychoacoustic and a cognitive model. The main limitations are that it does not handle highly impaired nor multi-channel audio.
PEMO-Q [134]	More Advanced auditory model that handle highly impaired audio but with higher complexity and fails for linear distortions.
Cave Model [138]	A novel auditory model based on PEAQ for audio coding. Attempted to address drawbacks of PEAQ's auditory model such as temporal masking model and calculation of SPL level but has not been used for Audio Quality assessment.
PEAQ with EEA [135, 139]	By adding the Energy Equalization Algorithm, this method is able to handle a wide variety of audio impairment. It does not cover though the spatial attributes nor psychological aspects.
Spatial-PEAQ [136, 141]	Introduces additional spatial MOVs to PEAQ original ones but does not handle highly impaired audio nor psychological aspects.
QESTRAL [132, 142]	Focus only on spatial distortions and uses an artificial listener model that relies on specific measurements of the reproduced sound field, which can be restrictive in some cases. Also, it does not cover psychological aspects.
ITU WG6 [133]	ITU Ongoing workgroup to extend PEAQ to multichannel audio.

4.3 Suggested approach

As highlighted before, the reliability of an objective method is determined by its correlation to a subjective reference. Since the latter should be as realistic as possible for a good prediction of the quality level, our approach was essentially based on the analysis of the *natural* process of audio quality evaluation. To handle the characteristics of multichannel sound, we also integrated additional tools for spatial quality analysis, and were mainly concerned about the convenience of the protocols and reproducibility of the results. Combining all these aspects, we were able to establish a new comprehensive approach for audio quality evaluation: AQUA (Audio QUality Assessment), as presented in the following sections.

4.3.1 Concept

The motivation of our study is to provide efficient tools for an automatic and objective multichannel audio quality assessment. In fact, and as illustrated in figure 4.4, our concept is inspired by the studies presented earlier and keeps the same base architecture, where specific features are extracted using psychoacoustical model and mapped to an objective grade through a cognitive model. In order to be able to handle multichannel audio streams, we added a binaural processor. As will be explained later, this block generates a binaural stereo signal as it is captured by the left and right ear. This way, we process only two channels while keeping the spatial information.

First, and once the binaural signals are synthesized, they go through our *Perceptual Model.* It consists of an auditory model and features extraction algorithms that provide measurements equivalent to the internal representations. It uses the MOVs of PEAQ standard in combination with additional variables.

Second, the *Spatial Model* uses the outputs of the *Binaural Model* to calculate the differences in binaural spatial cues, which will inform about any potential spatial distortions.

Finally, the extracted features and spatial cues are mapped to a single quality grade using the *Cognitive Model.* Additionally, the *Psychological Model* uses the feedback from the output to adjust the final score by simulating a realistic human judgment.

In order to do so, we first established an extended listening test protocol to include spatial cues analysis and other behavioral aspects as explained in the next section. Considering the missing aspects presented in the previous section, a new and more comprehensive listening test is actually mandatory for a proper and reliable validation.

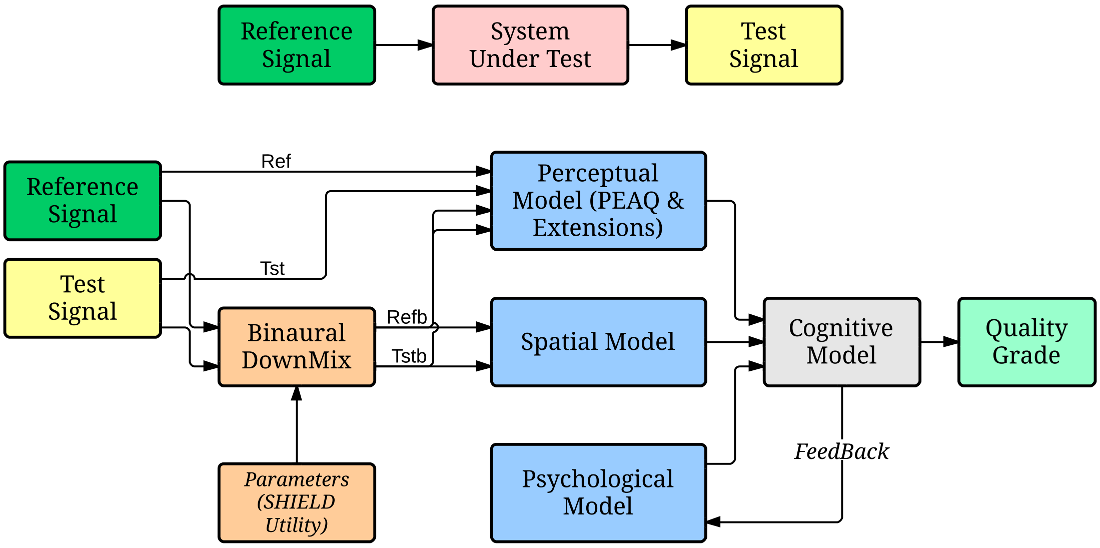


FIGURE 4.4: Concept of our objective approach

4.3.2 Subjective assessment

Thanks to our previous study of the process of subjective quality evaluation [144], we noticed that listening tests designed for audio quality assessment follow a certain generic structure as follows:

- Identify the audio feature to measure or monitor.
- Choose or create the test materials that exhibit this feature.
- Select the subjects, preferably experts, for a reliable results,
- Perform the test in a controlled environment using reliable and efficient tools,
- Analyze the results statistically while considering the potential sources of biases.

In the scope of our study, we aim to establish a method that can measure the perceptual and spatial quality of multichannel audio. Keeping in mind that two types of judgments are involved, sensory and affective, we create a protocol which includes the affective part instead of ignoring it or try to remove it. Considering that perceived audio quality is strongly influenced by cognitive aspects as mentioned in [133], we analyzed and integrated them as well to finally get a simple but yet realistic listening test that offers a more reliable assessment of 3D audio quality.

4.3.2.1 Protocol Design

In addition to the initial study of the natural quality evaluation process, the benchmarking of the existing protocols allowed us to establish a new one combining their strengths. Here is a summary of the most important aspects:

- The test is performed in two sessions: subjects rate the overall quality during the first one and the spatial quality in the second.
- A specific gesture-based protocol has been designed for spatial distortion assessment where an objective and reliable information regarding the sound localization can be measured, as detailed later.
- The test materials cover a wide variety of natural and synthetic sounds. The listener's preference for a specific type of content is taken into consideration. Also, a material duration does not exceed 10 to 15 seconds to avoid listener's fatigue or boredom.
- Unlike the existing recommendations, we believe that listening panel should not be only composed of experts. In fact, we consider that for realistic results, the panel should mainly comprise *average Joes*[2]. Besides, there is actually no exact definition for an audio expert and additional tests to evaluate such qualification will be time consuming and bothersome.
- Two types of tests have been designed: One-shot test (overall evaluation) and continuous test to cover the both possible scenarios.
- We worked on modeling the cognitive bias and monitored the subjects behavioral response. They were wearing EEG headsets that allowed us to monitor their focus level and provided a more efficient tool for post screening, as explained later.

More details on the protocol are explained in the next sections.

4.3.2.2 Perceptual assessment

On the light of the previous studies and the results of our preliminary tests, we were able to identify the most significant and easily processed attributes to evaluate audio quality as summarized in table 4.5. They are clustered in three categories: a *Perceptual* attributes that will describe timbral defects, a *Spatial* one that will describe the *spaciousness* of the multichannel sound and finally *Auxiliary* attributes that will indirectly affect the judgment.

[2] Normal non-expert people with no specific background.

Table 4.5: Most influent sound attributes on Quality perception

Category	**Attributes**
Perceptual	• Annoyance • Clarity • Realism
Spatial	• Sound Position • Listener Envelopment • Auditory Source Width
Auxiliary	• Content appreciation

Spaciousness can be defined as the amount of spatial extension of the auditory event. It is considered to be one of the most important characteristics in perceiving sound in halls and other listening spaces by many studies [145]. In fact, Beranek[3] has even stated, "it is one of the most effective indicators of the acoustical quality of concert halls" [146].

Annoyance (ANY): It is the indicator recommended by the ITU standards and refers to the psychological state of being irritated, bothered or annoyed. Although very subjective, if assisted with other attributes, it can be a reliable gauge of how the audio signal is perceived by the listener.

Clarity (CLR): A simple attribute that we introduced and allows the detection of a wide variety of distortions (clicks, hums, noise, etc.). It is complementary to previous one in a sense that the clearer a sound is, the less discomfort and annoyance it will cause to the listener.

Realism (REL): Not to be confused with synthetic sounds. This indicator should help with specific digital distortions like the clicks caused by packet loss for instance.

Sound Position: The measurement of sound position cannot be performed efficiently using verbal expression. It is measured using a specific protocol.

Listener Envelopment (LEV): is usually defined as the degree of fullness of sound images around the listener. In other words, the subjective surrounding of different sound fields around a listener (Fig. 4.5).

Auditory Source Width (ASW): can be defined as the width of a sound image fused temporally and spatially with direct sound image, or the apparent width of a sound image from a source perceived by a listener in the audience [145] (Fig. 4.5).

[3] One of the most famous acoustics expert in the world.

Content appreciation (CA): We also considered the listener's preference for the Signal Under Test (SUT). Anecdotally, a Classical music fan will most likely rate a Rap song badly and vice-versa.

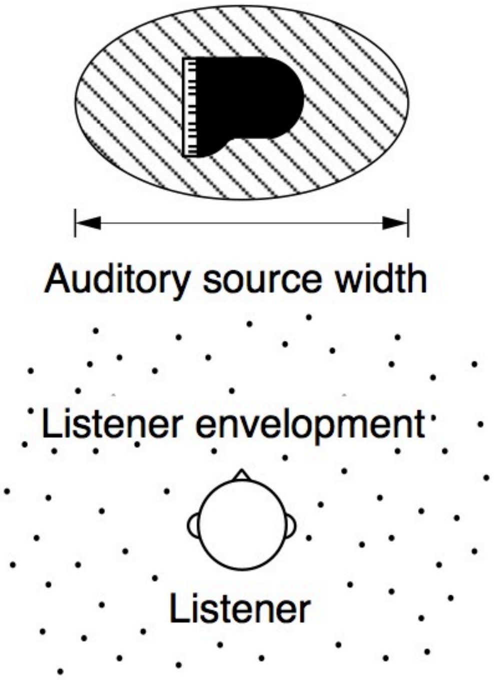


Figure 4.5: Concept of Auditory Source Width (ASW) and Listener Envelopment (LEV). Credits: [145]

Except for the sound position, the subjects were asked then to grade each attribute on a 5 points continuous scale with anchor labels. The test is computer-based and particular attention was put on the GUI design. In fact, elements of the interface can be hidden manually or automatically in order to avoid overloading subjects and focus on one feature at a time. Figure 4.7 shows a screenshot of the iPad version of the test software. It was actually appreciated by most subjects for its convenience and comfort.

According to the applications, some of these attributes can be ignored. We believe though that for a reliable subjective analysis, the spatial impression should be at least considered. As a matter of fact, figure shows the contribution of each factor and how correlated they are with the Overall Quality.

As mentioned before, verbal expression is not the best suited way for spatial quality analysis. We are suggesting though that it can be assisted with a natural gesture-based technique as explained in the following section.

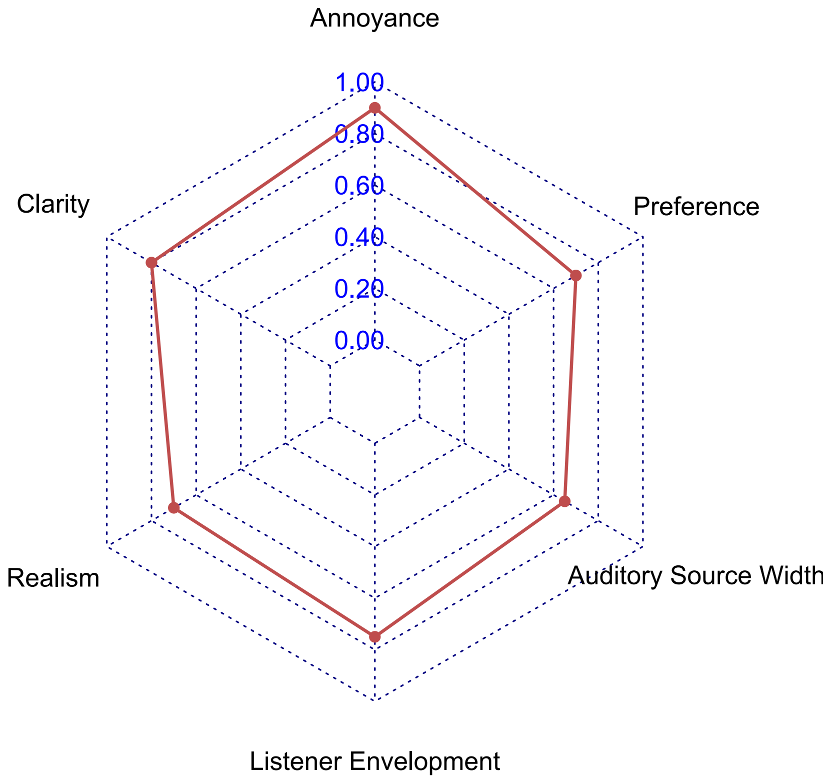
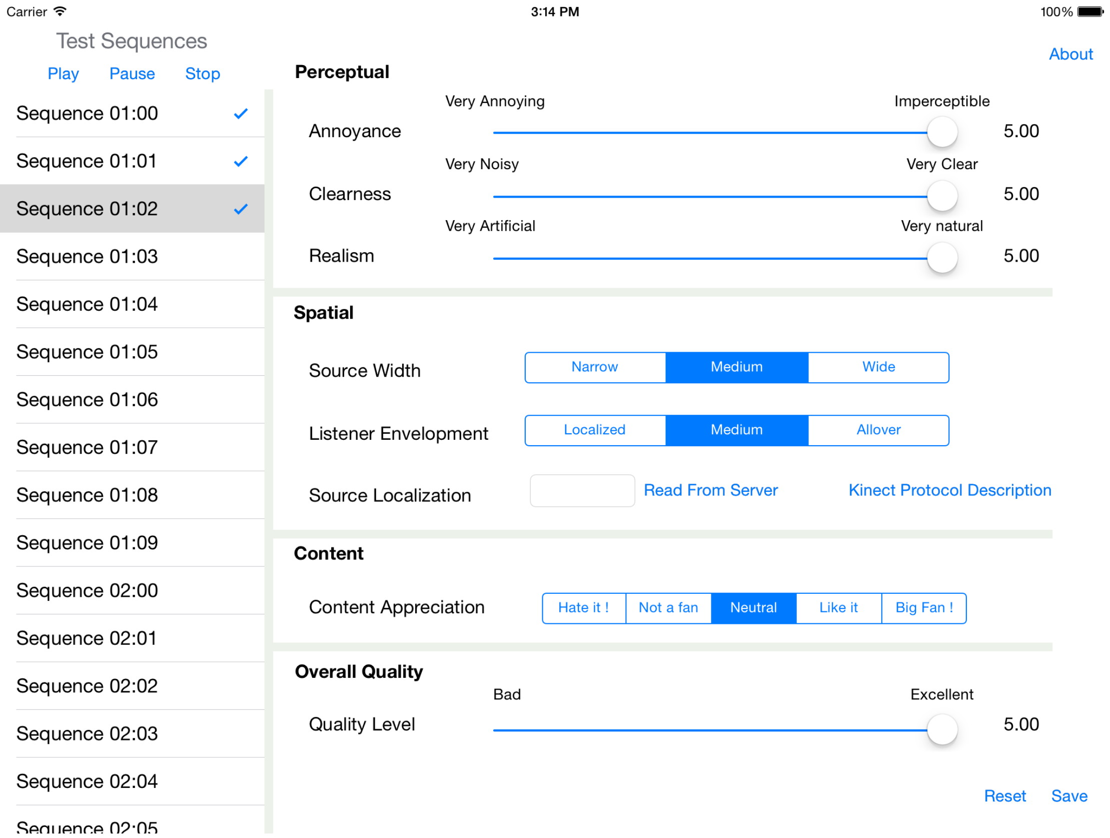


Figure 4.6: Quality attributes correlation with Overall Quality.

4.3.2.3 Analyzing spatial information

Through our study of the process of audio quality evaluation, we noticed that the most important factor for spatial quality and the first to be processed, is the localization of the audio source. Like any other forms of perception, spatial perception occurs both in the sensory organs that collect data about the environment and in the brain that process and interpret it (see section 2.1.2). For 3D sound, this relates to the auditory spatial attention, a specific form of attention, involving the focusing of auditory perception on a location in space. Auditory spatial awareness is a three-dimensional ability. In fact, hearing is the only directional human telereceptor that operates in a full 360° range and is equally effective in darkness as in bright light. Thus, the auditory system is frequently a guiding system for vision in determining the exact location and influence on a conscious and subconscious level the whole body movements [129].

As a matter of fact, auditory localization is the most critical sense to human effectiveness and personal safety. As mentioned before, the brain utilizes subtle differences in intensity, spectral, and timing cues to allow us to localize sound sources in 3D: the azimuth or horizontal angle, the elevation or vertical angle, and the distance (for static sounds) or velocity (for moving sounds). We studied accordingly how people interact



Test Sequences

Sequence 01:00 ✓
Sequence 01:01 ✓
Sequence 01:02 ✓
Sequence 01:03
Sequence 01:04
Sequence 01:05
Sequence 01:06
Sequence 01:07
Sequence 01:08
Sequence 01:09
Sequence 02:00
Sequence 02:01
Sequence 02:02
Sequence 02:03
Sequence 02:04
Sequence 02:05

Perceptual

Annoyance: Very Annoying – Imperceptible 5.00
Clearness: Very Noisy – Very Clear 5.00
Realism: Very Artificial – Very natural 5.00

Spatial

Source Width: Narrow [✓ Medium] Wide
Listener Envelopment: Localized [✓ Medium] Allover
Source Localization

Content

Content Appreciation: Hate it ! Not a fan [✓ Neutral] Like it Big Fan !

Overall Quality

Quality Level: Bad – Excellent 5.00

FIGURE 4.7: Screenshot of the iPad version of test software

with spatial audio and came to the conclusion that we naturally interact with space using our body where the most active parts are the head and the hands. As a result, our method is based on tracking the body movements.

To do so, we used the Kinect, the well-known motion sensor, and asked the subjects to simply *point the position* of the sound source, in a natural and free way. Generally speaking, it is an easy, quick and accurate test that can be performed to assess the spatial audio quality. In fact, a lot of techniques for direction pointing in the context of sound localization have been studied: Head (nose) pointing, gaze direction, laser (gun) pointing, etc. But hand pointing has been proven to be the most natural and reliable one [129]. Accordingly, while facing the kinect, subjects were simply using their hand and arm to point the sound position.

Kinect is a motion sensing input device introduced by Microsoft in 2010 for the Xbox 360 video game console and Windows PCs (Fig. 4.8). Based around a webcam-style add-on peripheral, it enables users to control and interact with the Xbox without the need to touch a game controller, through a natural user interface using gestures and spoken commands.

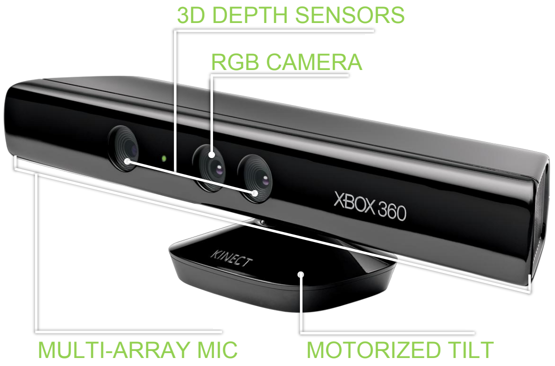


FIGURE 4.8: Kinect Sensor Description. Credits: Microsoft Website.

Kinect detection of the body position is a two-stage process: first, it constructs a depth map by analyzing a speckle pattern of infrared structured light [4], then it detects body parts and their positions using a randomized decision forest trained over a million of observations (machine learning). More details are available in [147].

One of the most interesting features of this sensor is the *Skeletal Tracking* that allows Kinect to recognize people and follow their actions as shown in figure 4.9. Using the infrared (IR) camera, Kinect can recognize up to six users in the field of view of the sensor, of which, up to two users can be tracked in details. An application can locate the joints of the tracked users in space and track their movements over time [148]. For convenience purposes, we built our software on Microsoft Kinect SDK that contains functions to directly call this feature.

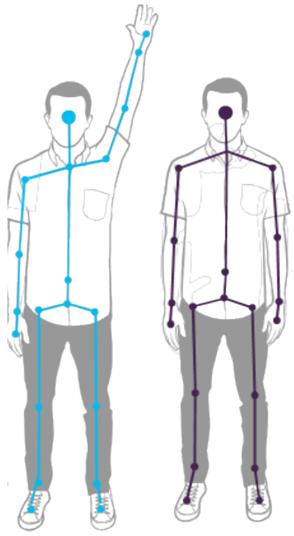


FIGURE 4.9: Kinect Skeletal Tracking

[4] The Kinect combines structured light with two classic computer vision techniques: depth from focus, and depth from stereo.

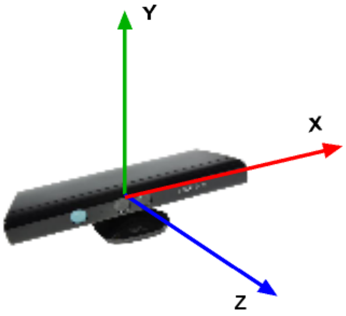


Figure 4.10: Kinect coordinate system

In the skeletal tracking mode, each joint is tracked by its (x, y, z) coordinates according to the Kinect system shown in figure 4.10. In order to estimate the pointed direction, we mainly extract the head and the active hand positions. By projecting the points, we are able to calculate the angle using the general form of *pythagorean theorem* as shown in figure 4.11:

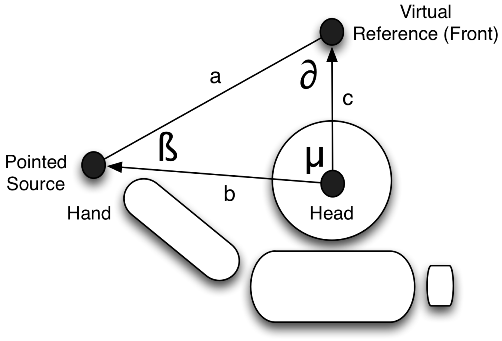


Figure 4.11: Angle Calculation

When people point a direction using their hand, they naturally consider the head as a reference. Consequently, the angles can be obtained by projecting the points on the (X, Z) plane for Azimuth and on the (Y, Z) plane for Elevation. The angles can be calculated then as described below:

$$a^2 = b^2 + c^2 - 2\ ab \cos \mu \tag{4.2}$$

$$\mu = \arccos \frac{b^2 + c^2 - a^2}{2\ ab} \tag{4.3}$$

The angle μ is the pointed direction, considering the front as a reference. Estimating the distance using this method seems to be a bit tricky and inaccurate and additional sensor might be required.

So by combining this technique with the previously discussed attributes, we can get more accurate and objectively reliable results, by slightly tweaking the protocol and adding this quick and simple test. The accuracy of the process has been validated during the preliminary tests phase where subjects were asked to point few reference points, accurately marked on the ground. Arm Tracking using kinect shows an overall high accuracy expect for the area behind the listener. This is due to the fact that the *glenohumeral joint* cannot perform rotations beyond $150 - 160°$ [149]. Figure 4.12 illustrates the results of the validation phase. Figure 4.13 demonstrates an example of direct application of this process, in the context of the evaluation of the accuracy of an HRTF database [150].

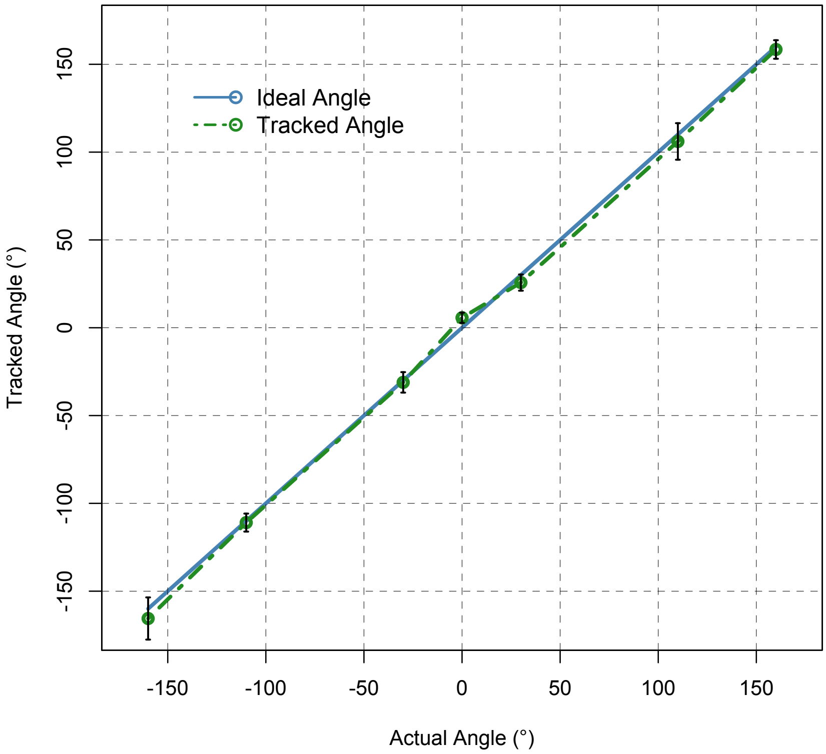


FIGURE 4.12: Kinect Tracking Performance

4.3.2.4 Studying psychological bias

In this part, we focus on how to improve the realism of quality assessment techniques by including psychological considerations. In our previous study [144], we highlighted how the traditional tests do not consider *memory bias*. As a matter of fact, if we consider the ABC test presented in section 4.2.1, and perform the test in the following order: Reference, Impaired, Reference, it is most likely to get a lower grade during the second time than the first. This is what we referred to as the *Experience Effect* where a previous experience, especially a negative one, influences the next-coming judgments for a

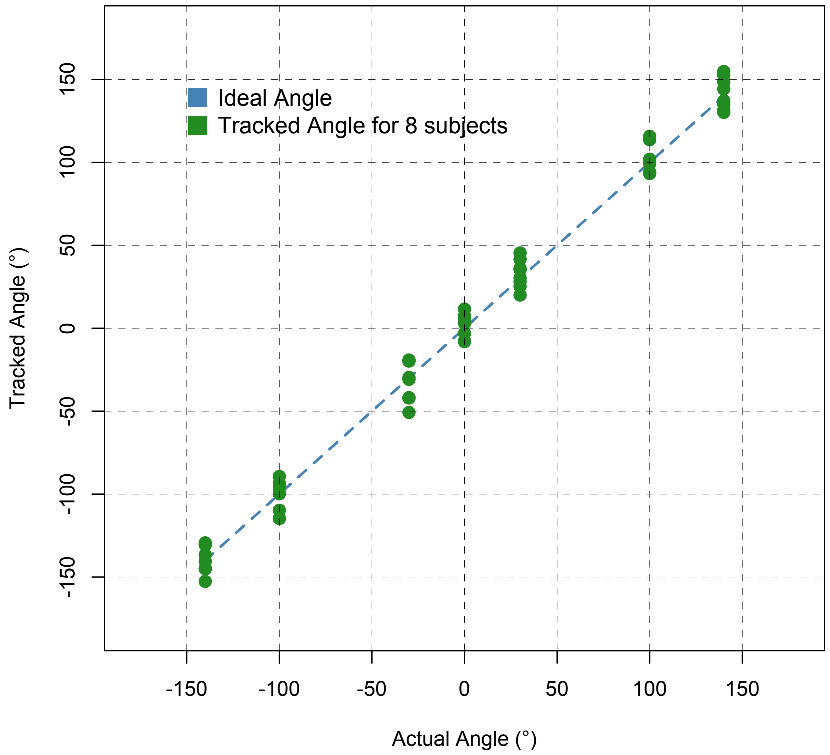


FIGURE 4.13: Example of test results using ARI DataBase with few subjects

certain amount of time. This aspect finds its theoretical roots in cognitive sciences and particularly the *Recency Effect* as studied by Ebbinghaus in [151]. It has been also highlighted by Bovik et al. in their studies of video quality assessment [152]. Accordingly, and considering that quality assessment is generally a continuous process, such feature should be integrated for a realistic evaluation.

Accordingly, we performed several tests to get a better understanding of this phenomenon. The initial and simplest ones are shown in figure 4.14, where *Ref* and *Imp* stands for the *Reference* and *Impaired* Signals.

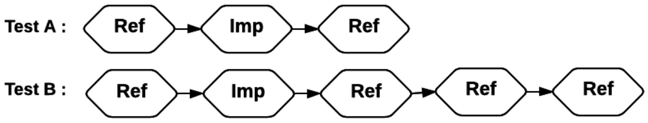
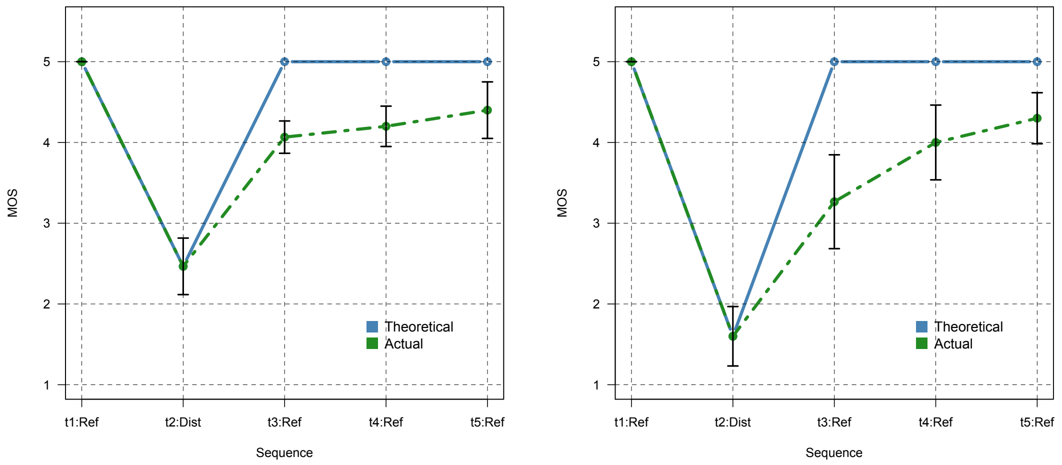


FIGURE 4.14: Basic tests for the experience effect

These tests are an extension of the ITU's ABC test. The first one can be helpful to identify the existence of the experience effect for the studied impairment. If that is the case, the second one can be performed to estimate the impact and duration of this effect. Our results show that these two factors depend mainly on the nature, duration and intensity of the distortion in the impaired signal. We also performed series of tests



(A) Recovery Cycle for an AWGN distortion (SNR=36.5 dB)

(B) Recovery Cycle for a Random Packet Loss (1%, 5 ms)

FIGURE 4.15: Illustration of the experience effect using Test B. For a more annoying distortion, the recovery cycle takes longer time.

in continuous mode (instead of rating separate and segmented sequences) for a more accurate modeling of this effect, as presented in the objective section. Depending on the test strategy, the results can be either used to smooth out data by removing this bias or modeled and integrated for a more realistic analysis. Figure 4.15 shows an illustration of this effect using the previous tests for an AWGN distortion (Additive White Gaussian Noise) and random packets drops in case of packet based transmission (like SIRIUS system). We can notice that the listeners do not instantly recover from the distortion, and that the recovery cycle gets slower as the impairments gets more severe and annoying.

A question arises then regarding the fatigue factor resulting from the continuous mode of the test, which is actually the case. This will eventually lead to additional biases resulting from tiredness or simply boredom. Since asking the subjects about such aspect will be irrelevant, we managed to monitor their focus level using an EEG (*Electro-Encephalo-Graphy*) headset.

EEG refers to the recording of the brain's spontaneous electrical activity over a short period of time. The voltage signal is non-invasively measured using one or multiple electrodes placed on or near the scalp. As noted by Mostow et al. [153], synchronized neural activity varies according to development, mental state, and cognitive activity, and the EEG signal can measurably detect such variation. For example, rhythmic fluctuations in the EEG signal occur within several particular frequency bands, and the relative level of activity within each frequency band has been associated with brain states such as focused attentional processing, engagement, and frustration [154].

Most of the EEG cerebral signal observed in the scalp falls in the range of 1−50 Hz. Activity below or above this range is most likely artifactual or pathological. Table 4.6 shows the mental state or activity associated with each [155, 156].

TABLE 4.6: EEG rhythmic activity frequency bands

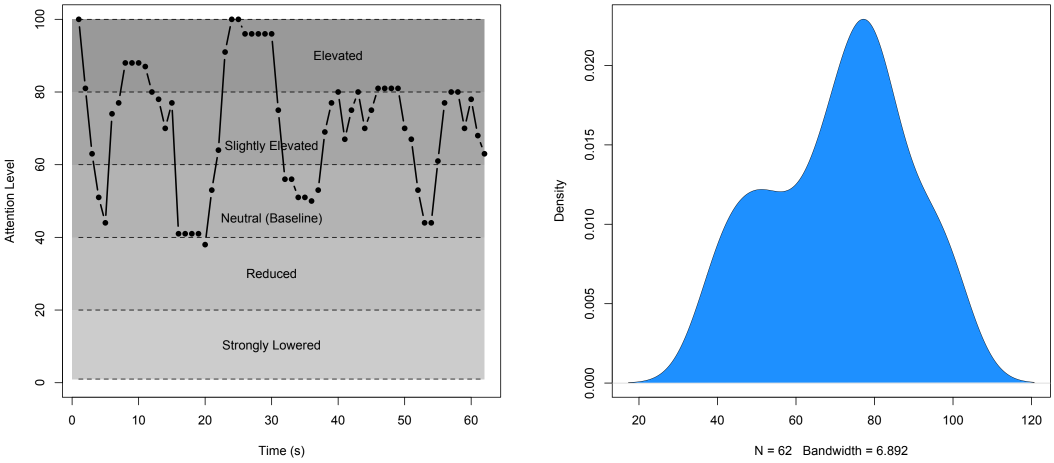
Band	**Frequency (Hz)**	**State/Activity**
Delta	$0 < \cdots < 4$	• Sleep state • During some continuous-attention tasks
Theta	$4 \leq \cdots \leq 7.5$	• Drowsiness, idling • Associated with inhibition of elicited responses (when a person is actively trying to repress a response or action)
Alpha	$8 \leq \cdots \leq 15$	• Relaxed/Reflecting • Eyes closed
Beta	$16 \leq \cdots < 31$	• Active thinking, focus, hi alert, anxious
Gamma	31 and above	• During cross-modal sensory processing (perception that combines two different senses, such as sound and sight) • Shown during short-term memory matching of recognized objects, sounds, or tactile sensations

Multi-electrode, medical grade EEG systems have long been used in hospitals and laboratories. But the recent availability of low-cost single-channel EEG devices makes it feasible to take this technology from the laboratory into informal environments such as ours to explore its capabilities.

For our application, and in order to avoid adding another bias that may result from discomfort, we used the *NeuroSky® MindWave*. It is a lightweight wireless headset designed for research grade studies and has prebuilt and validated algorithms for mental states detection. Based on the volume conduction property of EEG [157], it uses a single dry electrode contacting the user's forehead, which makes it the most appropriate headset for our study[5]. The most interesting feature in our case is the *Attention* meter [158] which indicates the intensity of a user's level of mental focus[6]. Using this feature, we are able to confidently eliminate some psychological biases and efficiently post-screen our subjects.

[5] A simple headset like the MindWave will not add stress or distraction.

[6] Based on a machine learning process.



(A) Measured Attention Level for Subject JSN

(B) Density Analysis of the Attention Level

FIGURE 4.16: Illustration of the Attention Level measurement using the NeuroSky eSense Meter.

Although it is possible to re-train the attention meter with the neural patterns of each subject, we used directly the one available with the SDK. Our preliminary tests as well as previous studies showed actually its consistency with the active state of mental concentration [159].

The eSense meter indicates the attention level on a $1-100$ scale[7] divided in 5 ranges as illustrated in figure 4.16a. Based on that, we were able to screen more efficiently the subjects and separate the psychological biases in order to focus on the experience effect. The duration of the continuous mode was limited though to one minute as we noticed signs of tiredness and boredom above it. The decision is based on the density analysis (histogram) of the measured levels. Figure 4.16b illustrates a valid one for subject JSN, who was most of the time focused on the experiment.

In the following section, we present how all these aspects can be combined to establish a new and more reliable objective audio quality measurement technique.

[7] The Zero value indicates the signal loss.

4.3.3 Objective assessment

One of our main concern is to make objective methods capable of handling multichannel sound in an efficient way. Therefore, the first step of our approach is to convert multichannel audio streams to binaural stream as presented in figure 4.4. Binaural synthesis is inspired by the natural process of sound localization as illustrated in figure 4.17 and detailed in section 2.1.2.

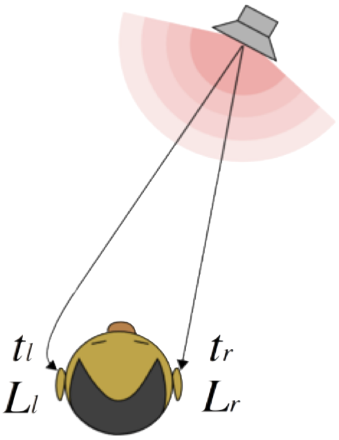


FIGURE 4.17: Binaural principle

For a considered sound source at a specific position in space, the sound wave will respectively reach the left and right ear, at time t_l and t_r, at a level L_l and L_r. The Interaural Time difference (ITD) and the Interaural Level Difference (ILD), as shown in the equations 4 and 5, are the main cues used by the brain to determine the position of the audio source [7]. This means that a signal $s(t)$ will be perceived by the left and right ear as $s_l(t)$ and $s_r(t)$ as presented in equation 6 and 7. $h_l(t)$ and $h_r(t)$ are the Binaural Room Impulse Response (BRIR) that matches the spatial position of the audio source and reflects the room acoustical characteristics, the listener morphology, etc. It is the time-domain equivalent to the infamous Head Related Transfer Functions (HRTF).

$$s_l(t) = s(t) * h_l(t) \tag{4.4}$$

$$s_r(t) = s(t) * h_r(t) \tag{4.5}$$

Consequently, binaural synthesis consists of convoluting the audio source with the BRIR that match a specific position. Accordingly, we can virtually place a sound at any position as long as the impulse responses are available.

In our case, we have been experimenting with surround audio materials using a 7.1 loudspeakers setup, according to ITU.BS.775 standard [160]. The equivalent binaural stream can be computed as described in equation 4.6.

$$\begin{pmatrix} L_B & R_B \end{pmatrix} = \begin{pmatrix} L & R & C & Ls & Rs & Bsl & Bsr \end{pmatrix} * \begin{pmatrix} h_{L,L} & h_{R,L} \\ h_{L,R} & h_{R,R} \\ h_{L,C} & h_{R,C} \\ h_{L,Ls} & h_{R,Ls} \\ h_{L,Rs} & h_{R,Rs} \\ h_{L,Bsl} & h_{R,Bsl} \\ h_{L,Bsr} & h_{R,Bsr} \end{pmatrix} \tag{4.6}$$

L_B and R_B are the left and right channel of the resulting binaural sound. L, R, C, L_s, R_s, B_{sl} and B_{sr} are respectively the Left, Right, Center, Left Surround, Right Surround, Back Surround Left and Back Surround Right channels of the multichannel sound in the 7.1 setup. The LFE channel is ignored as it has no spatial contribution. $h_{x,y}$ stands for the BRIR for the ear x at the position matching the channel y. The equation can be easily extended to more complex multichannel audio systems.

Considering the subjective nature of HRTFs [7], we have to choose the one that preserve the spatial integrity of multichannel streams and match the spatial perception of our subjects. The protocol presented in section 4.3.2.3 is an appropriate way to do so.

In fact, many studies have established different HRTF databases. The most known ones are listed in table 4.7.

Table 4.7: HRTF Database List

ID	Made By	Description
LSTN	IRCAM	51 subjects (2003)
ARI	ARI	87 subjects (2013)
CIPIC	CIPIC	45 subjects and 2 KEMAR Mannequin (2004)
PKU	Pekin University	KEMAR Mannequin and distance dependent (2008)
MIT	MIT	KEMAR Mannequin (2001)
TUB	TU Berlin	KEMAR Mannequins and distance dependent (2007-2011)

Distance dependent databases are more interesting in our case as they allow to assess the spatial quality not only at the sweet spot but virtually at any position of the listening area. In fact, as part of the parameters of the binaural down-mixer, we can specify a position within the listening space and the binaural synthesis will be done accordingly.

For the visualization convenience, let's consider a 2D context. A standard HRTF measured at a single distance will allow a binaural simulation on a circle as described by equation below, where x_{lsn} and y_{lsn} are the coordinates of the listener position.

$$x = x_{lsn} + r.cos(\theta) \tag{4.7}$$

$$y = y_{lsn} + r.sin(\theta) \tag{4.8}$$

The angle θ is generally $0 \leq \theta < 2\pi$ but in the case of an HRTF measured within a minimum and maximum Azimuth values, $Az_{min} \leq \theta \leq Az_{max}$. Accordingly, all the audio sources must be within the same distance of the listener for a proper binauralization. In other words, the listener must be at the geometrical center of the playback system (i.e., the sweet spot), which is rarely the case.

Distance dependent measurements change the previous equation and defines more of a discrete *region* around a listener as illustrated in figure 4.18 and described in equation 4.9, with $r_d = d_{min} \ldots d_{max}$.

$$Reg_{lsn} = \bigcup_{r_d = d_{min}}^{d_{max}} \begin{Bmatrix} x = x_{lsn} + r_d.cos(\theta) \\ y = y_{lsn} + r_d.sin(\theta) \end{Bmatrix} \tag{4.9}$$

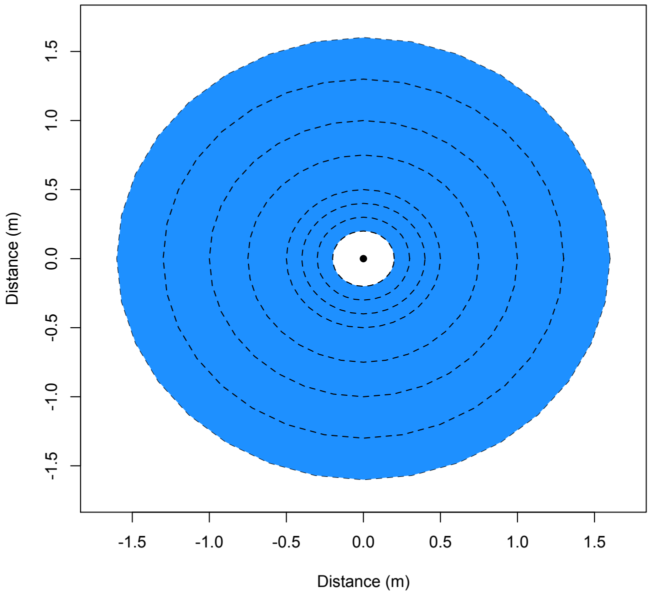


FIGURE 4.18: Possible positions for binaural simulation using a distance dependent database (PKU).

In our context, we simulate the equivalent binaural stream for a static setup. It is actually the listener who is moving within the listening area defined by the loudspeakers setup. Accordingly, each sound source will defines its own possible region $Reg_{src(i)}$. As a result, it is possible to create the equivalent binaural sound at multiple positions, instead of only at the sweet spot. For multiple N sources, the coverage area is the intersection for their respective regions as shown in equation 4.10.

$$Area = \bigcap_{i=1}^{N} Reg_{src(i)} \tag{4.10}$$

Consequently, it can be extended to a 3D context as described in equation 4.11, where θ and δ are respectively the Azimuth and Elevation angles.

$$Area = \bigcap_{i=1}^{N} \bigcup_{r_d=d_{min}}^{d_{max}} \left\{ \begin{matrix} x = x_{src(i)} + r_d.cos(\delta).cos(\theta) \\ y = y_{src(i)} + r_d.cos(\delta).sin(\theta) \\ z = z_{src(i)} + r_d.sin(\delta) \end{matrix} \right\} \tag{4.11}$$

Figure 4.19 illustrates the possible binaural simulation area for a 7.1 configuration, considering that the measurements have been done within the range of a distance dependent HRTF database.

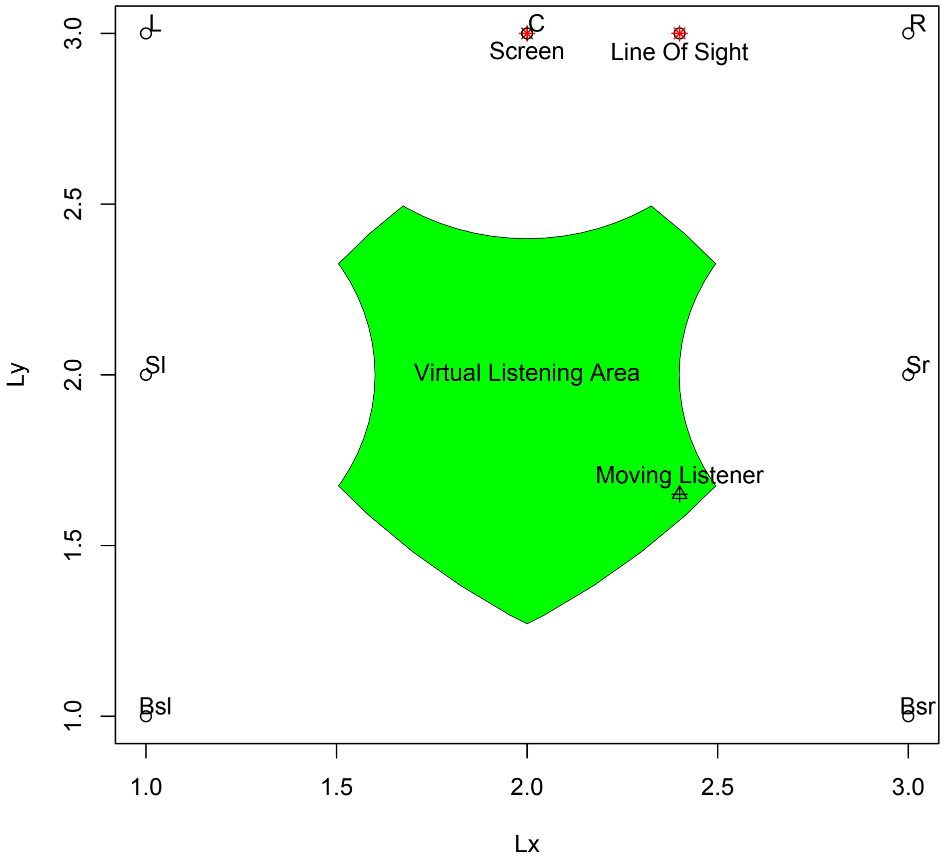


FIGURE 4.19: Snapshot from AQUA manager utility (SHIELD). The green area refers to the possible listening positions for a generic 7.1 configuration considering minimum and maximum distances from the source.

For a fair comparison, we mainly benchmarked the HRTF measured in the same conditions, namely MIT [161], PKU [162] and TUB [163] databases. By comparing these DB using the gesture-based protocol, we were able to determine that PKU DB is the most accurate for our listening panel, as illustrated in figure 4.20. Based on a 7.1 system, it scored a Root Mean Square Error (RMSE) of $\approx 29,69°$ which can be even lower with additional measurement points.

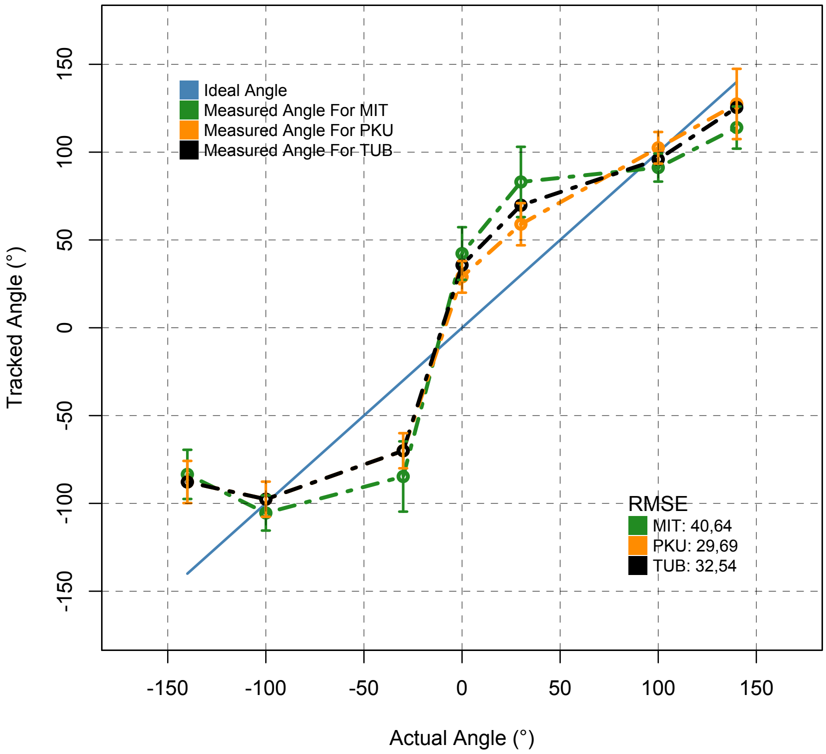


FIGURE 4.20: Accuracy comparison of MIT, PKU and TUB HRTF Databases

4.3.3.1 Perceptual model

Our objective perceptual model is based on the combination of the most reliable features of the existing methods. By separately analyzing the correlation of each MOV with the subjective reference, as well as the correlation between the MOVs, we were able to elect the best candidates as follow:

- Keeping only the MOV showing a correlation with the subjective reference higher than our considered threshold,
- If two MOV are highly correlated, we only keep the one with the higher correlation to the subjective reference to avoid redundancy.

Seo et. al. in [141] method is based on the same approach and their results are close enough to ours. Their model was trained though using primarily and only low to moderate impairments. The concept is that PEAQ advanced version MOVs generally show a high correlation with the listening tests data, which makes them our fundamental tool.

We also changed the structure of PEAQ's filterbank to 24 bands one based on fourth-order Gammatone filters, as they offer a more realistic peripheral ear model and higher implementation efficiency [137]. The width of each filter was set to the equivalent rectangular bandwidth (ERB). The outputs are weighted then to simulate the outer and middle ear as described in equation 4.12. f_c is the central frequency of each band in KHz.

$$W[k]/dB \;=\; -2.184 \times (f_c[k])^{-0.8} + 6.5 \times e^{-0.6\times(f_c[k]-3.3)^2} - 10^{-3}\,(f_c[k])^{3.6} \tag{4.12}$$

We added two additional MOVs to enhance the perceptual model reliability. The first one is based on the assumptions of EEA [135]. It has been integrated to handle high levels impairments but we optimized the implementation for computational efficiency. By using directly the filter banks outputs for each time frame, we get an equivalent time-frequency analysis with no additional operation.

For a time-frequency segment, we consider that an *island* has appeared if the energy of the segment on the test signal exceeds the initial energy on the reference signal more than the Just Noticeable Level Difference (JNLD). Although it is generally set to 1 dB, JNLD depends on the nature and pressure level of the input sound signal [164]. At low input levels, it is higher than at high input levels. For example, for 1-KHz tone, JNLD is typically 2 dB at 10 dB and drops to 0.2 dB at 80 dB [165]. Considering that we are using short windows, comparing the Maxs of each segment is a simple way to do it. Accordingly, our metric, the Island Detection Counter (IDC) is incremented as follow:

$$IDC \;=\; IDC + 1 \quad if \quad max\left(S_{tst}\left[k,n\right]\right) \;>\; max\left(S_{ref}\left[k,n\right]\right) + \delta \tag{4.13}$$

$$\delta = JNLD\left(max\left(S_{ref}\left[k,n\right], f_c[k]\right)\right) \tag{4.14}$$

The NDI (Newly Detected Islands) MOV, is finally calculated as follow, where K and N are respectively the filterbank channel number and frames numbers:

$$NDI = IDC/(K.N) \tag{4.15}$$

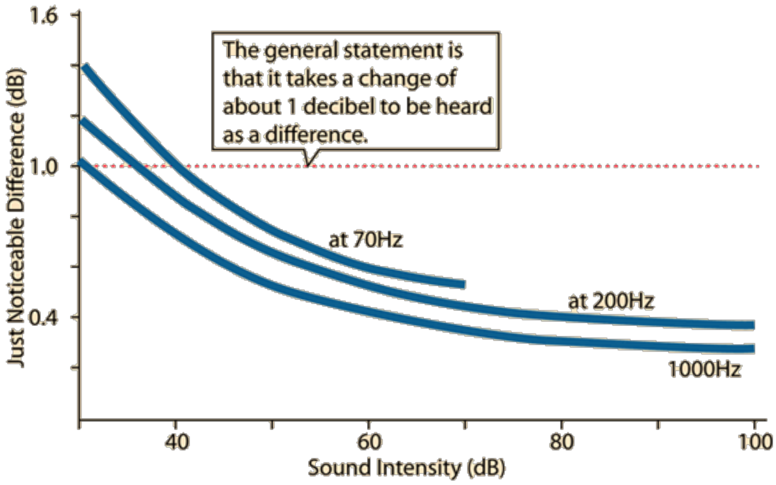


FIGURE 4.21: Just Noticeable Level Difference. Credits: [166]

Binaural down-mixing offers the advantage of always analyzing only two channels despite the complexity of the multichannel audio system under test. Still, and since the normalization may compromise the perceived intensity of the multichannel sound, the signal loudness is calculated directly from the multichannel stream as the second perceptual MOV. Multichannel loudness is calculated according to ITU-R BS.1770-3 recommendation [167] as illustrated in figure 4.22. RLB stands for Revised Low-frequency B-curve filters. As loudness measurement involves integration over time, it is performed in parallel with the perceptual model filter-bank.

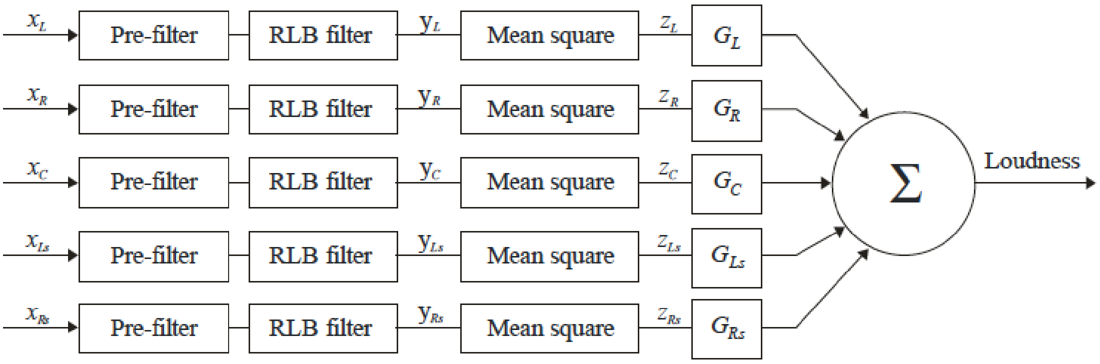


FIGURE 4.22: Block diagram of proposed multichannel loudness meter as suggested by BS1770 for a 5.1 setup. It can be extended to higher configurations.

The other MOVs are calculated according to the standards (see appendix C) and shows very high correlation with the subjective reference as illustrated in figure 4.23. Four additional MOVs have been added to analyze spatial information as presented in the next section.

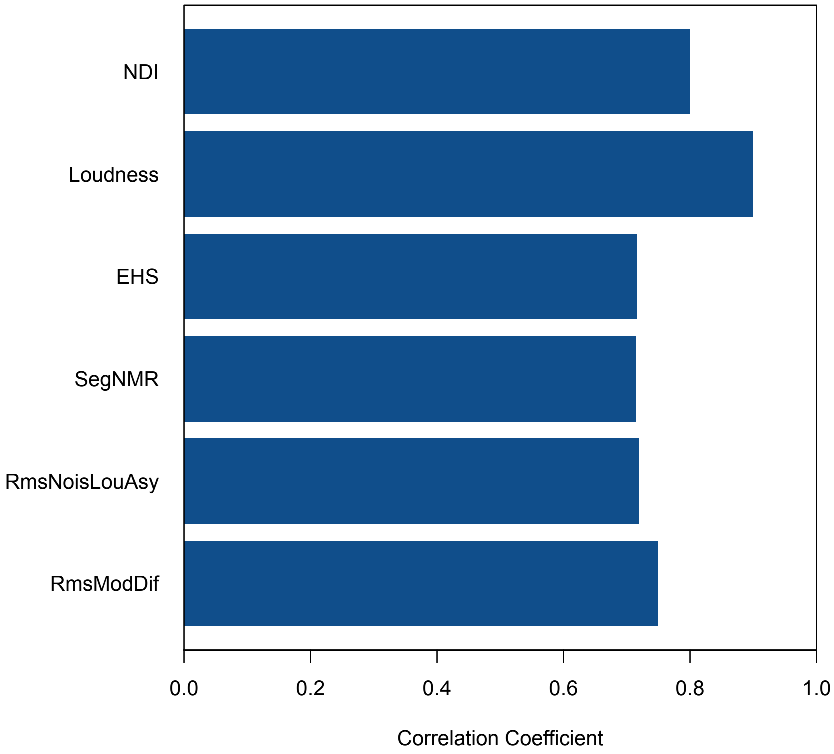


Figure 4.23: Perceptual MOVs Correlation

4.3.3.2 Spatial model

The spatial cues introduced in this section are based on the previously presented binaural principle and are as follow:

- Interaural Time Difference (ITD): difference between the times at which sounds reach the two ears. It is the most influencing cue in low frequencies, especially below 1500 Hz.
- Interaural Level Difference (ILD): difference between the levels at which sounds reach the two ears. It is the most predominant cue in high frequencies, particularly above 2500 Hz.
- InterAural Cross-correlation Coefficient (IACC): the long-term IACC is related to the perceived source width and calculated in low frequencies, below 1500 Hz.
- The difference between the estimated angle of the reference and test signal (EAD), based on ITD.

The following processing is performed for both the reference and test signals. $S_{L,k,n}$ and $S_{R,k,n}$ refer respectively to the left and right channel of signal S, where k and n are the frequency band and time frame indexes. Considering that ILD and ITD are detected in different areas of the brain, respectively the Lateral Superior Olive (LSO) and the Medial Superior Olive (MSO), which exhibit different sensitivity. The audio signal windowing (time-frame segmentation) is performed accordingly [168]. As suggested in [136], we used rectangular windows with the following configuration:

- ITD: 20 ms frame with 7/8 overlapping,
- ILD: 10 ms frame with 3/4 overlapping,
- IACC: 50 ms frame with 3/4 overlapping (the frame duration is longer as this cue is more relevant on long-term).

Windows lengths were adjusted to the closer power of 2 for computational efficiency, respectively, 1024, 512 and 32768 samples.

The ILD in dB is calculated as follow:

$$ILD\left[k,n\right] = 10\log_{10}\left(\frac{\sum\limits_{l} S^2_{L,n,k}[i]}{\sum\limits_{l} S^2_{R,n,k}[i]}\right) \tag{4.16}$$

The cross-correlation function described in equation 4.17, allows the calculation of ITD and IACC as shown in equations 4.18 and 4.19.

$$CCF_{n,k} = \sum_{-N}^{+N} S_{L,k,n}(i) S_{R,k,n}(i-\tau) \tag{4.17}$$

$$IACC\left[k,n\right] = max\left|CCF_{k,n}(\tau)\right|_{-N}^{+N} \tag{4.18}$$

$$ITD\left[k,n\right] = \underset{\tau}{argmax}\left|CCF_{n,k}(\tau)\right|_{-N}^{+N} \tag{4.19}$$

In order to improve the computational efficiency, spatial cues can be calculated in parallel with perceptual ones.

The ILD difference (ILDD) between the reference and test signal is our first MOV and can be calculated as described below. The raw ILDD is weighted by the downmixed monophonic signal as shown in equation 4.22.

$$w_M\left[k,n\right] = 10\log_{10}\left(\sum_{l} S^2{}_{M,k,n}\left[l\right]\right) \tag{4.20}$$

$$S_{M,k,n}\left[l\right] = (S_{L,k,n}\left[l\right] + S_{R,k,n}\left[l\right])/2 \tag{4.21}$$

$$ILDD_{seg}\left[k,n\right] = w_M\left[k,n\right].\left|ILDD_{ref}\left[k,n\right] - ILDD_{tst}\left[k,n\right]\right| \tag{4.22}$$

The overall ILDD is then calculated by averaging the segmental ILDD over the K frequency bands and F time frames.

$$ILDD_{tmp}\left[n\right] = \frac{1}{K}\sum_{k=1}^{K} ILDD_{seg}\left[k,n\right] \tag{4.23}$$

$$ILDD = \frac{1}{F}\sum_{n=1}^{F} ILDD_{tmp}\left[n\right] \tag{4.24}$$

For the ITD calculation and based on the method introduced in [169], the perceptual distance due to the ITD difference can be estimated by:

$$\Delta ITD\left[k,n\right] = \sqrt{2 - 2cos\pi\frac{f_s}{N}\left(ITD_{tst}\left[k,n\right] - ITD_{ref}\left[k,n\right]\right)} \tag{4.25}$$

Considering that low values of IACC can false the ITD detection, Choi introduced a logistic function[8] to model the detection probability of IACC as shown in equation 4.26, where S_{ITD} and T_{ITD} are steepness and threshold of the sigmoid logistic function [136]. S_{ITD} is set to 50 for all bands and T_{ITD} is frequency dependent as shown in table 4.8.

$$p\left[k,n\right] = \frac{1}{1 + e^{-S_{ITD}(IACC[k,n] - T_{ITD}[k])}} \tag{4.26}$$

Table 4.8: Frequency dependent Threshold for ITDD estimation

f_c	T_{ITD}	f_c	T_{ITD}
21.5	0.1	698.3	0.3
73.0	0.9	888.9	0.3
135.1	0.9	1118.8	0.1
210.0	0.4	1396.0	0.1
300.2	0.7	1560.3	0.1
409.0	0.6	1991.7	0.1
540.1	0.1		

$$ITDD_{seg}\left[k,n\right] = \Delta ITD\left[k,n\right].(p_{ref}\left[k,n\right] + p_{tst}\left[k,n\right])/2 \tag{4.27}$$

[8] Based on the neural pathway described in the 2^{nd} chapter.

The overall ITDD is then calculated by averaging the segmental ITDD over the K frequency bands and F time frames.

$$ITDD_{tmp}\left[n\right] = \frac{1}{K}\sum_{k=1}^{K} ITDD_{seg}\left[k,n\right] \tag{4.28}$$

$$ITDD = \frac{1}{F}\sum_{n=1}^{F} ITDD_{tmp}\left[n\right] \tag{4.29}$$

The IACC difference (IACCD) is calculated as follow:

$$IACCD_{seg}\left[k,n\right] = \left|IACCD_{ref}\left[k,n\right] - IACCD_{tst}\left[k,n\right]\right| \tag{4.30}$$

$$IACCD_{tmp}\left[n\right] = \frac{1}{K}\sum_{k=1}^{K} IACCD_{seg}\left[k,n\right] \tag{4.31}$$

$$IACCD = \frac{1}{F}\sum_{n=1}^{F} IACCD_{tmp}\left[n\right] \tag{4.32}$$

The traditional spatial cues calculation as described above gives very good results with low computational complexity. Although, it still can be more accurately estimated by considering the recent findings in ILD processing as suggested in [5, 141]. In fact by considering the neuronal firing pattern involved in this process, an advanced version of the calculation can be performed as explained below.

First, the loudness of each channel $c = L, R$ is computed as in eq. 4.33.

$$w_c\left[k,n\right] = 10\log_{10}\left(S^2_{c,k,n}\left[l\right]\right) \tag{4.33}$$

The spike rates of left and right LSO cells, S_L and S_R can be modeled using a logistic sigmoid function as in eq. 4.34.

$$S_L\left[k,n\right] = w_L\left[k,n\right]\frac{1}{1+e^{-S_{ILD}[k](ILD[k,n]-T_{ILD})}} \tag{4.34}$$

$$S_R\left[k,n\right] = w_R\left[k,n\right]\frac{1}{1+e^{S_{ILD}[k](ILD[k,n]-T_{ILD})}} \tag{4.35}$$

The ILD threshold, T_{ILD} was set to zero. The steepness, S_{ILD}, on the other hand, is frequency dependent as described in table 4.9.

Table 4.9: Frequency dependent Steepness for ILDD calculation

f_c	T_{ITD}	f_c	T_{ITD}
1560.3	0.01	7025.0	0.01
1991.7	0.01	8516.7	0.01
2618.6	0.01	10315	0.01
3204.2	0.01	12484	0.07
3910.2	0.01	15098	0.01
4761.4	0.01	18250	0.1

In order to include the processing occurring in the Inferior Colliculus (IC) cells, their interaction with the contralateral LSO cells can be modeled as follows:

$$i_L\left[k,n\right] = w_R\left[k,n\right]\frac{1}{1+e^{-S_{ILD}[k](w_R[k,n]-S_L[k,n])}} \tag{4.36}$$

$$i_R\left[k,n\right] = w_L\left[k,n\right]\frac{1}{1+e^{-S_{ILD}[k](w_L[k,n]-S_R[k,n])}} \tag{4.37}$$

The ILD for a given time-frequency segment is then calculated as in Eq. 4.40. The overall ILDD is then obtained by averaging the segmental ILDD over frequency bands and time frames.

$$ILD_{seg}\left[k,n\right] = i_L\left[k,n\right] - i_R\left[k,n\right] \tag{4.38}$$

$$ILDD_{seg}\left[k,n\right] = \left|ILD_{ref}\left[k,n\right] - ILD_{tst}\left[k,n\right]\right| \tag{4.39}$$

$$ILDD_{tmp}\left[n\right] = \frac{1}{K}\sum_{k=1}^{K} ILDD_{seg}\left[k,n\right] \tag{4.40}$$

$$ILDD = \frac{1}{F}\sum_{n=1}^{F} ILDD_{tmp}\left[n\right] \tag{4.41}$$

ITDD can be also weighted by the signal loudness because it is difficult to detect it when the signal is inaudible. The weighting factor is computed as in eq. 4.42:

$$w_{ITDD}\left[k,n\right] = \frac{1}{4}\left(\sum_{c=L,R}\left(\sum_{s=ref,tst} w_{c,s}\left[k,n\right]\right)\right) \tag{4.42}$$

For a given time-frequency segment, the ITDD becomes as in eq. 4.43. The overall ITDD is calculated then as in eq. 4.28.

$$ITDD_{w\_seg}\left[k,n\right] = w_{ITDD}\left[k,n\right] ITDD_{seg}\left[k,n\right] \tag{4.43}$$

The same thing applies to IACC calculation. The w_{IACCD} is equivalent to w_{ITDD} but with larger window as mentioned before.

The last cue is the Estimated Angles Difference (EAD). Based on an approach similar to the one suggested in [170], the ITD can be used to estimate the source Direction of Arrival (DOA) and mainly the azimuth. In fact, by analyzing several audio sequences, we were able to empirically approach the relationship linking the azimuth to ITD using a sine wave as shown in figure 4.24. The angle can be estimated then using the formula described in 4.44, where is the Estimated Angle is in Radians. The EAD is calculated then using the eq 4.45. Due to the π-limit of interaural differences [171], this cue is limited to $[-90, 90]°$ range. It also allows an easier matching with the subjective assessment, using the protocol in section 4.3.2.3. The *arcsin* approximation offer a simple expression and high correlation coefficient ($R^2 = 0.975$). We kept though the possibility to adjust and train a polynomial approximation by using a custom HRTF. In both cases, and to enhance the computational efficiency, ITD to azimuth matching is performed using a Look Up Table (LUT).

$$EA_{seg}\left[k,n\right] = \arcsin\left(ITD\left[k,n\right]/ITD_{max}\right) \tag{4.44}$$

$$EAD_{seg}\left[k,n\right] = \left|EA_{ref}\left[k,n\right] - EA_{tst}\left[k,n\right]\right| \tag{4.45}$$

The study in [141] suggests the Envelope Interaural Time Difference Distortion (EITDD) as an additional MOV. Due to its relatively low correlation compared to the other cues, it has not been considered in our study. The selected MOVs shows very high correlation as illustrated in figure 4.25.

4.3.3.3 Psychological model

The purpose of the psychological model is to mimic the process of human decision making. When it comes to rate an audio distortion, psychological biases such as the experience effect [144], should be considered to enhance the metric's realism.

In realistic situations, quality assessment is generally a continuous process. In fact, in a regular listening test, and while listening to the distorted test material, the subject

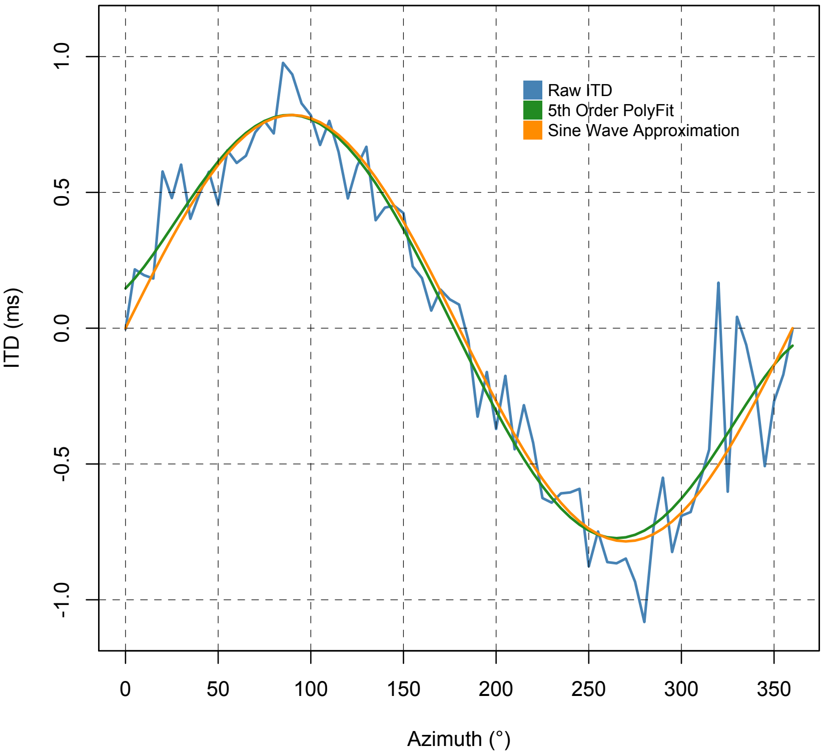


FIGURE 4.24: ITD to Azimuth Approximation

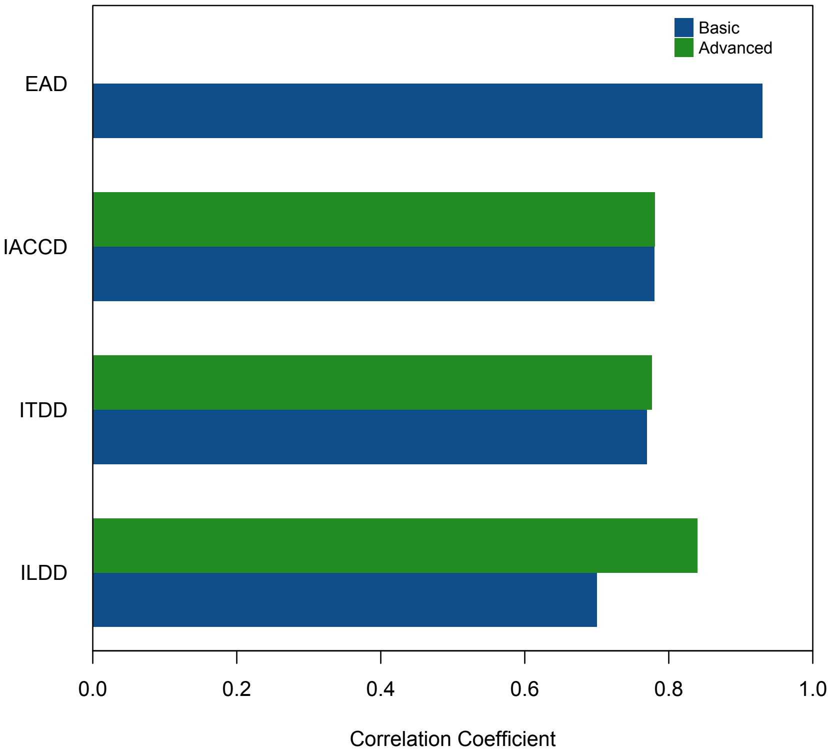


FIGURE 4.25: Spatial Cues Correlation

is continuously evaluating its quality and decides of its level when he is asked to grade it. The final grade can be considered as a rough *average* of the continuous score that he kept in his short term memory, weighted by psychological biases. In fact, we studied the rating of audio quality in continuous mode and we mainly noticed the following:

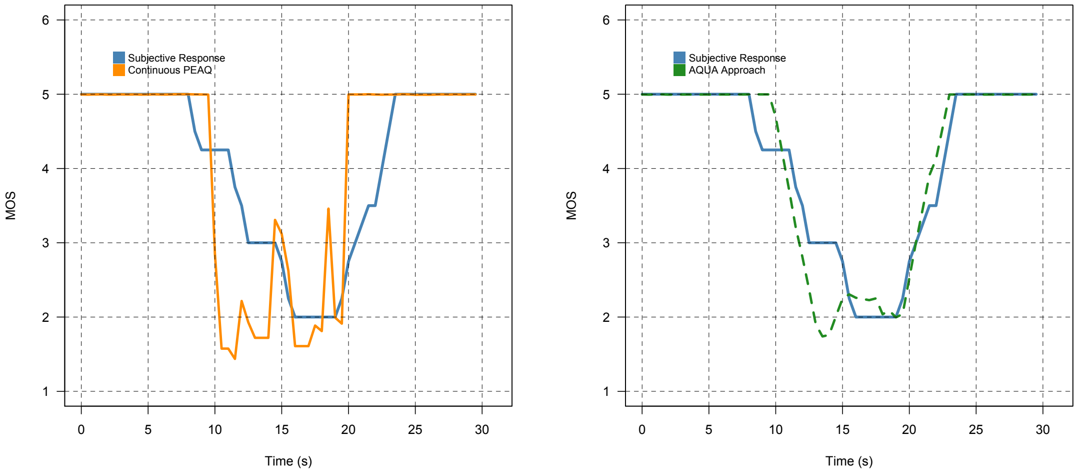
- A local distortion, and depending on its nature, affects the judgment on the next sequences differently. In fact, the less annoying is a distortion, the quicker people will forget it.
- The duration of the effect and the recovery cycle depends on the intensity and duration of the distortion,

A lot of psychological processes are involved in this context related to learning mechanism, memory recall and decision making [151, 172]. Since our purpose is to provide a simple and efficient model, we mainly focused on the *experience effect*. We conducted series of subjective tests in continuous mode, where we were able to measure preliminary but substantial data. By considering the distortion grade as the dominant factor, we train an artificial neural network (ANN) where the input is the model output grade (ODG) and the output will be the weights applied on the next ones.

Also, additional aspects have to be considered for a real time implementation of the continuous quality monitoring. For instance, we noticed that if the quality level indicator is fluctuating rapidly, it can be confusing for the end-user and eventually lead to a misevaluation of the system. Figure 4.26 illustrates this phenomenon using a continuous implementation of PEAQ, and how our approach is closer to the subjective reference. Although the smoothing function is more efficient in offline mode, it can be integrated in an online approach with a tolerable latency, by using a moving average.

4.3.3.4 Cognitive model

Once all MOVs are calculated, they are mapped through an Artificial Neural Network to a single number matching the subjective differential grade to predict the overall quality level. It also provides the feedback for the psychological model, as explained in the previous section. This cognitive model is based on a one hidden layer ANN using the asymmetric sigmoid as an activation function with eight hidden neurons. Cross-validation has been performed first and the model was trained after with the all dataset, using the resilient backpropagation (Rprop+) algorithm. The overall performances were quite satisfying: MAE=0.12, RMSE=0.16, and presented in the next section. The training and analysis has been performed using Matlab *nftool* utility and R *neuralnet* tool.



(A) PEAQ versus Subjective reference

(B) The AQUA model output

FIGURE 4.26: Illustration of the model output with and without the integration of the experience effect and the moving average.

4.4 Validation & Model Discussion

The training was performed using a wide variety of audio signals combining speech, music and movies dialogue. As mentioned earlier, the content preference was also monitored and the rating was considered accordingly. Several audio codecs and distortions have been tested but we focused mainly on the usual and realistic ones, namely the Additive White Gaussian Noise (AWGN), random Packet Loss during a network transmission as studied in [71] and Psychoacoustical Compression (lossy). Most of the PEAQ materials have been included in addition to new ones. Table 4.10 enumerates the main subsets of the test materials that was included in the training in order to cover most of the acoustical properties of audio signals. For the spatial study, we did not experiment much of musical sounds as localisation performance has been known to be better for non-musical ones [173]. Figure 4.27 illustrates the scoring for the 279 signals, where sequences with packet loss were rated with the lowest grades.

As illustrated, the grades cover the ITU 5-points scale which provides a more realistic and reliable training set for the ANN. In fact, the distortions levels were adjusted in order to fulfill this requisite.

The model has been trained with and without the multichannel loudness measurement as it adds considerable latency. The overall performance are quite satisfying as illustrated in figure 4.28, with a MAE=0.1273 and RMSE=0.1607[9]. We noticed that the

[9] Mean Average Error and Root Mean Square Error

Table 4.10: Test Materials subsets and examples

ID	Items	Description
1	Vocal (Speech)	English, French and German (M/F)
2	Vocal (Opera)	Tenor, Soprano
3	Inst. 1	Piano, Violins, Cello
4	Inst. 2	Saxophone, Flute, Harmonica
5	Inst. 3	Bongos, Drums, Toms, Cymbals
6	Inst. 4	Bag Pipe, Glockenspiel, Organ
7	Ac. Guitar	strummed, finger style
8	Elec. Guitar	Clean, Distorted
9	Country music	e.g. Wilie Nelson, Blake Shelton
10	Rock music	e.g. Led Zeppelin, AC/DC
11	Rap music	e.g. 2pac, Eminem
12	Electro	e.g. Daft Punk, various dubstep

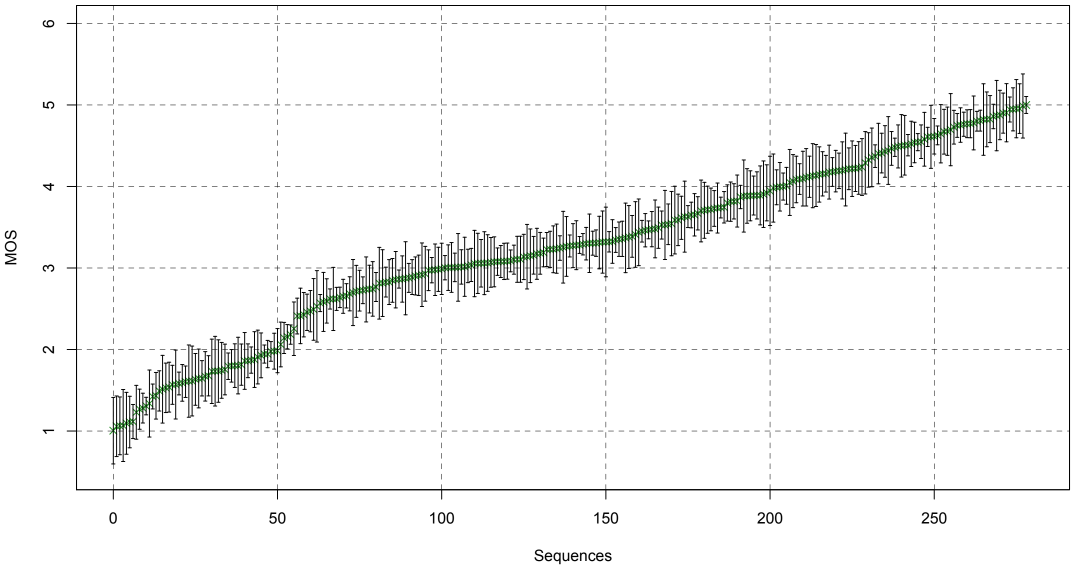


Figure 4.27: Subjective Assessment of the test sequences using the averaged MOS over the perceptual and spatial attributes with the standard deviation. The grades were sorted for better visualisation.

model is very efficient for linear distortions, thanks to NDI metric, but fails in the case of digital impairments such as packet loss. In fact, specific metrics has yet to be established for such impairment and we believe that integrating a VAD (Voice Activity Detector) should be helpful. The individual MOVs are also available as outputs for expert users.

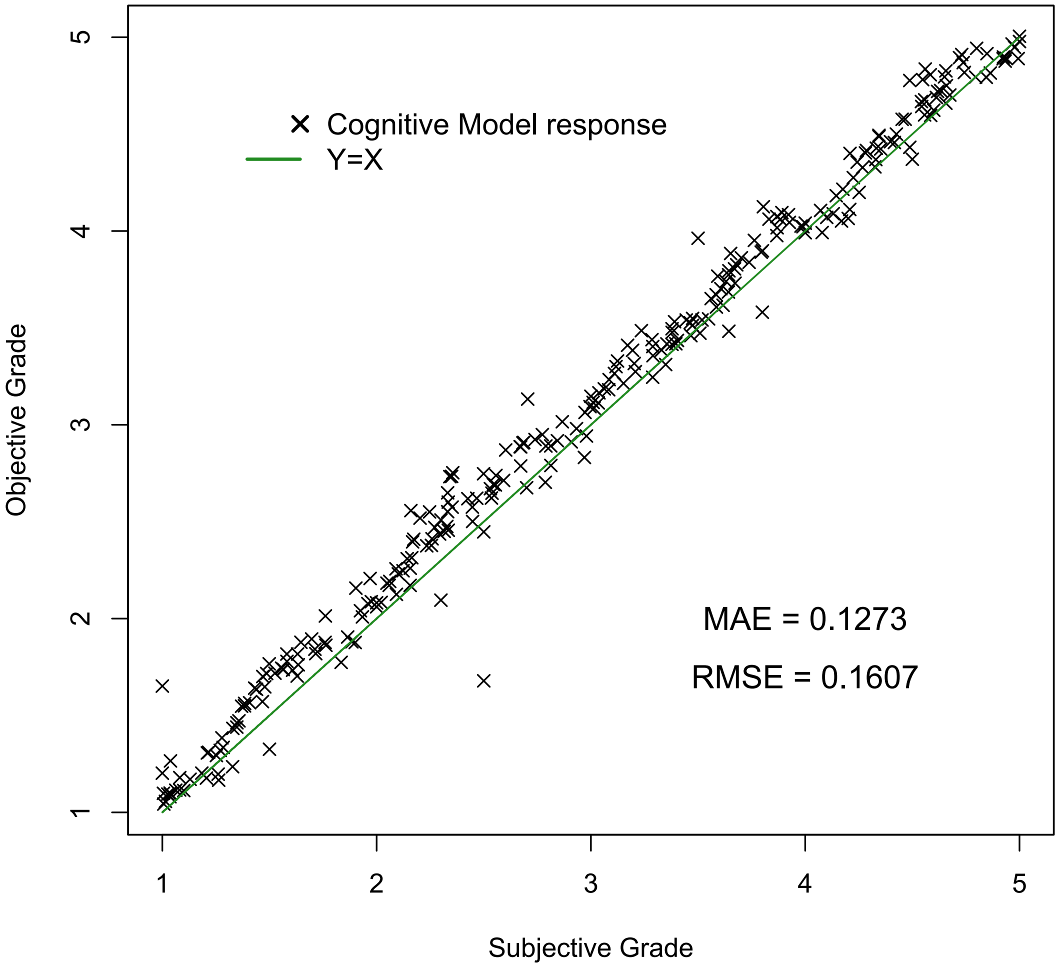


Figure 4.28: AQUA Model response

Although it may seem accessory, the design of test Interface plays a major role. Through our study, and as we improved the GUI, we noticed that people get more comfortable with the test as the interface become lighter and simpler. For instance, the ipad version was really appreciated and the focus level was more stable using it. In a related manner, we asked subjects to express their content appreciation. Although it can be applied to weight the rating, it has been essentially used for screening.

The spatial cues are one of the most interesting additions. Thanks to the binaural approach and the gesture-based protocol, the analysis is independent of the reproduction format. ITD calculation still presents a limitation though: Since it is sample-wise, and with signals sampled at 48 KHz, the minimal detected variation is $20,83$ ns, which refer to approximately 2° variation that can be detected using EAD. It still though one of the slowest operation in our model as it is based on intercorrelation on short windows. The psychological model introduces additional latency also is it requires previous grades.

4.5 Conclusion

As the users become more concerned about multimedia quality, new techniques need to be established to match the characteristics of the new audiovisual contents. Multichannel audio for instance add a real spatial dimension and offers with 3D video, a fully immersive experience for users. Our approach offers a simple and efficient way to assess the quality of the spatial rendering.

For the subjective assessment, our gesture-based protocol using Kinect offers a low-cost, fast and efficient tool for spatial quality evaluation. Although, it still limited to DOA estimation, it can be easily combined with additional low-cost sensor to include distance too.

The inherent problem with audio quality assessment is that audio distortion can be voluntary introduced for artistic purposes, which makes the evaluation even more difficult. One good example will be rock music involving electric guitar. An efficient method should integrate such aspects.

Objective methods are a good way to avoid those issues but cannot resolve them. With that in mind, one of our future direction is deeper study and modeling of the psychological and behavioral response to multimedia distortions. EEG for instance, has been efficiently used for screening, and we believe that it still hold more potential for quality analysis.

These finding allowed us to validate the multichannel audio transport mechanism. The word truncation using Dithering was generally imperceptible ($4{,}5-5$) as well as Packet Loss Rate below 0,3%. For values above, 0,5%, the packet drops become very perceptible (clicks) and annoying. The synchronization performance using our software-only approach was satisfying as well, since no spatial distortion has been detected.

Matlab, R and other tools source codes, as well as the sequences database, will be available soon on L2TI website.

Chapter 5

Surround Sound Optimization

5.1 Context

A three-dimensional audio system function is to render sound images around the listener by using either headphones or loudspeakers. In the case of a headphone-based 3D audio systems, the 3D cues to localize a virtual source can be perfectly reproduced at the listener's ear drums, because the headphone isolates the listener from external sounds and room reverberations [174]. The spatial reproduction of sound in a conventional surround system works only in a small area which is located on the symmetry axis between the loudspeakers, the so called *Sweet Spot*.

The sweet spot describes the focal point between the loudspeakers, where a listener is fully capable of hearing the stereo audio mix the way it was intended to be heard by the mixer. Beyond this area, the spatial perception collapses and the stereo image moves to the nearest loudspeaker since the signal arrives both sooner and louder [175].

Although they can sense it, most people are not really aware of this phenomenon. Through this chapter, we analyze how the sweet spot affects the spatial impression and how it can be improved.

5.2 Literature Review

Stereophonic and surround multichannel reproduction systems are widely used today. One major disadvantage of such playback systems is the narrow sweet spot in which correct audio localization is possible [176]. This is actually a result of the panning mixing technique, as presented in section 2.2. It results from the fact that the mixer (i.e., sound engineer) adjusts the level and time differences between each channel to create the desired spatial impression by simulating virtual or *phantom* audio sources around a specific listening position (Fig. 5.1). As a reminder, a virtual source denotes an auditory object that is perceived in a location that does not correspond to any physical sound source.

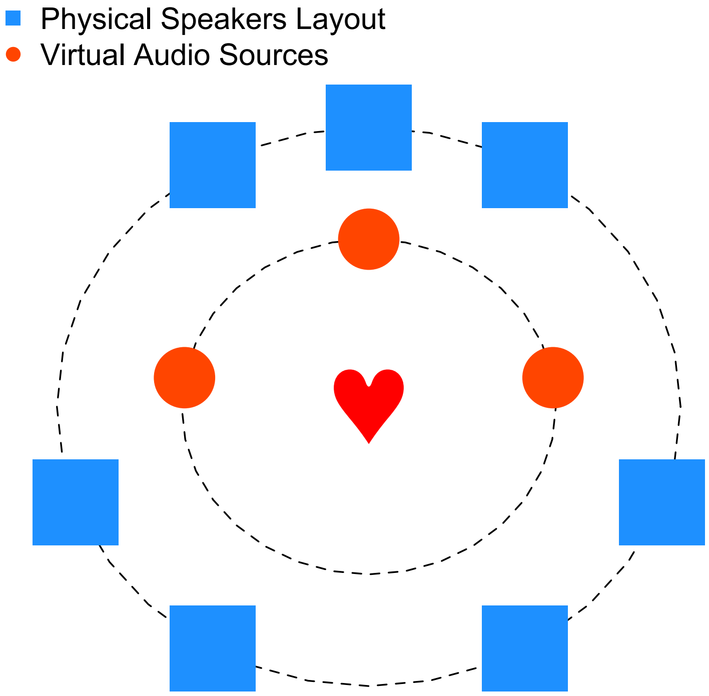


FIGURE 5.1: The Sweet Spot position for ITU 7.1 setup.

In fact, releasing the listener from this static hearing position was an issue since the beginning of stereophony. The Sweet Spot is actually more of a stretched area rather than a spot. Several studies have been conducted to determine the stereophonic localization within this area, and explain that the optimal listening zone depends on the maximum tolerable shift of the phantom source [177]. The sweet spot limitation is also tied to the concept of *Precedence effect*, which is a binaural psychoacoustic effect. When a sound is followed by another one separated by a relatively short time delay[1], the listener perceives a single fused auditory image. The perceived spatial location is affected by the second sound but dominated by the location of the first arriving one. As a result, and in the context of a stereophonic playback system, the stereo image is completely located at the nearest loudspeaker.

In an attempt to reproduce a better auditive localization over a larger listening area, different playback methods have been developed. Ambisonics and WFS are good examples of systems that do not suffer from such limitation. Unfortunately, and in

[1] below the listener's echo threshold.

most cases, new recording techniques are needed and the complexity of the reproduction systems increases consistently and rapidly. On the contrary, stereophony is widely used and many stereophonic and surround recordings techniques and materials are available. Consequently, releasing the listener from the sweet spot while keeping the same recordings and playback systems would be a the most appreciated and advantageous solution.

For that matter, two approaches have been suggested: either to widen the sweet spot or to adjust the audio signals in order to dynamically move it.

The existing techniques to *widen* or *broaden* the sweet spot can be classified into two categories. Those who try to adjust the radiation pattern of the loudspeakers and those who directly adjust the signals of the loudspeakers.

The first attempt to widen the area of stereophonic perception by adjusting the loudspeakers characteristics goes back to 1960 [177]. As a matter of fact, Bauer explains that the level difference between two loudspeakers in a listening point depends on room characteristics and the directivity of the loudspeakers. For monopoles there is only a small area close to the symmetry axis, in which the level difference is smaller than 3 dB. He suggests a system, where the angel between the loudspeaker axes is approximately 120 to 130°. Frequencies above 250 Hz should be radiated using dipoles. In such a system, the level difference between the loudspeakers remains almost constant over a *wider* area as can be seen in figure 5.2. Further studies dealing with the measurement of the optimal loudspeaker directivity and its use for sweet spot broadening are available in [178, 179].

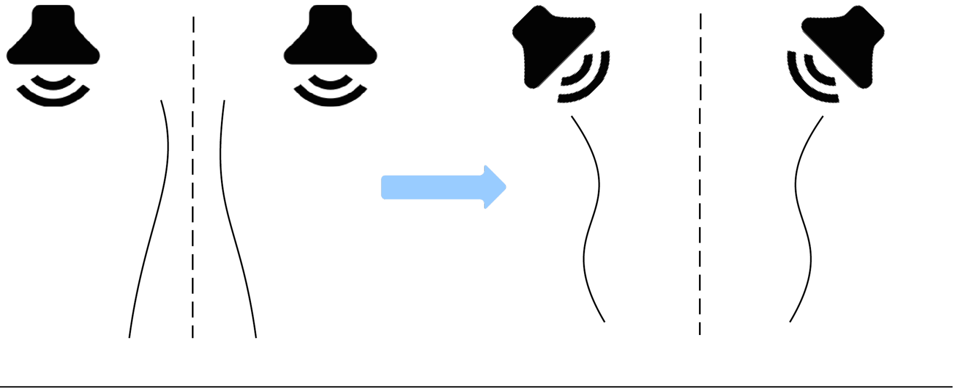


FIGURE 5.2: Sweet Spot Widening using Bauer technique.

The issue with such techniques is that the loudspeaker radiation pattern is frequency-dependent, which cannot be arbitrarily adjusted. Furthermore, all those works do not address the real problem of the precedence effect. This is especially dominant for transient signals like speech or music. In addition, these methods introduce contradicting localization cues from interaural level and time differences, leading to localization blur for eccentric listening positions [177]. Besides, changing the loudspeakers orientation is not always possible (e.g., laptops, handheld devices, etc.)

Another approach consists of directly adjusting the signals sent to the loudspeakers. Beside the level differences, the delay between the loudspeaker signals resulting from different distances to the listener's position, is responsible for localization shift (i.e., ILDs and ITDs). The following methods try to adjust the delay for a specific listening position so that the speaker signals reach the head of the listener at the same time. In many High Fidelity systems, the delay between loudspeakers can be adjusted manually. Others systems have automated this process using microphone measurements. For instance, the study in [180] describes an interesting system that uses groups of delayed directional loudspeakers. The main limitation is that such systems are generally designed for static positions and do not adjust to listener movement.

Additionally, adaptive room equalization can be used to widen the sweet spot. The Trinnov Optimizer is a good example of such system as presented in [181] and discussed in [182]. Unfortunately, it is time consuming since it requires specific measurements using multiple microphones as well as some knowledge of the rooms acoustics.

The first concept for a system that adjusts to the listener's position was suggested by Kyriakakis in [183]. He described a system which uses head tracking and time delay between the loudspeaker signals to *move the sweet spot*. There was no publications though about how it can be performed or on how to handle the artifacts that may result from it. Another use of the listener's location estimation is presented in [184]. In the context of a virtual reality presentation system, the position of the remote control is used to reproduce binaurally rendered surround signals via loudspeakers while using a crosstalk cancellation system. The system however is not designed for stereophonic reproduction and continuous adjustment of the sweet spot.

In this context, several techniques have been suggested to virtually put the sweet spot on the listener instead of the classical approach. The most comprehensive one is the *SweetSpotter* approach [175, 176]. Merchel et. al. suggest a two loudspeakers system which includes an optical face tracker providing information about the listener's x, y position. Accordingly, the loudspeakers signals are manipulated in real-time in order to move the sweet spot. The delay is calculated in such a manner that the signals of both loudspeakers arrive at the center of the listener's head at the exact same time. Moreover, the amplitudes of the loudspeaker signals are adjusted to reduce the level difference at the listening position and maintain the stereo image location.

As illustrated in Figure 5.3, for *off-center* listening positions, the signal paths from loudspeakers to ears become asymmetrical. For this example, the signal at the left ear

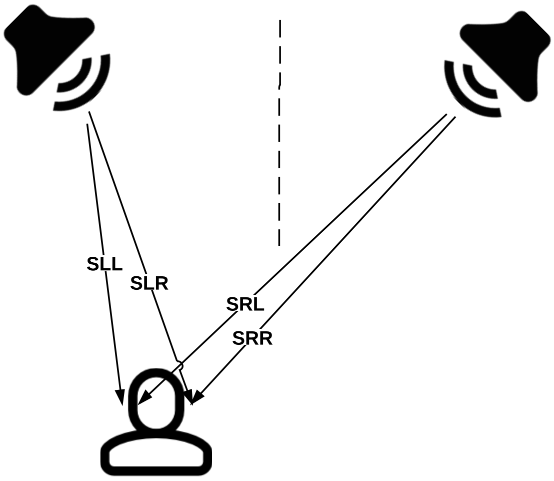


FIGURE 5.3: The effect of an off-center listening position.

originating from the right loudspeaker S_{RL} will be more attenuated due to stronger head shadowing than the signal at the right ear originating from the left loudspeaker S_{LR}. Consequently, the arrival time difference τ_R between the signal at the right ear originating from the right loudspeaker S_{RR} and the signal at the left ear originating from the right loudspeaker S_{RL} will be bigger than the arrival time difference τ_L, between S_{LL} and S_{LR} ($\tau_L < \tau_R$). This asymmetry is important for correct off-center localization.

The effect can be illustrated using an impulse phantom source located at the center between the loudspeakers. The two loudspeaker emit identical impulses, which are adjusted to reach the center of the head at the exact same time with the same amplitude. Using Lipshitz's analytical approach [185], the superimposed signals at the listener's ears can be analyzed, under low-frequency sine waves assumption. The resulting phase difference is converted into an ITD and finally in a corresponding azimuth angle.

Accordingly, the sound pressure at the ears can be mathematically described, under the assumption of sinusoidal signals with low frequencies. Thus head shadowing effects can be neglected. The signals at the loudspeakers have an amplitude ratio of $\frac{L}{R}$ and a time difference τ_d. It is also assumed that delay and level differences due to different distances between loudspeakers and listener position are compensated for, using the previous method of signal adjustment. The resulting signals from left and right loudspeaker at the left ear, according to Figure 5.3, are:

$$S_{LL} = L \;.\; exp\left(j\omega\frac{\tau_d + \tau_L}{2}\right) \tag{5.1}$$

$$S_{RL} = R \;.\; exp\left(-j\omega\frac{\tau_d + \tau_R}{2}\right) \tag{5.2}$$

The resulting signal at the left ear, S_L, is the superimposed sum of the two signals:

$$\begin{aligned} S_L = S_{LL} + S_{RL} = L\cos\left(j\omega\frac{\tau_d+\tau_L}{2}\right) + jL\sin\left(j\omega\frac{\tau_d+\tau_L}{2}\right) \\ + R\cos\left(j\omega\frac{\tau_d+\tau_R}{2}\right) - jR\sin\left(j\omega\frac{\tau_d+\tau_R}{2}\right) \end{aligned} \quad (5.3)$$

Similarly, the resulting signal at the right ear, S_R is described by equation 5.4:

$$\begin{aligned} S_R = S_{LR} + S_{RR} = L\cos\left(j\omega\frac{\tau_d-\tau_L}{2}\right) + jL\sin\left(j\omega\frac{\tau_d-\tau_L}{2}\right) \\ + R\cos\left(j\omega\frac{\tau_d-\tau_R}{2}\right) - jR\sin\left(j\omega\frac{\tau_d-\tau_R}{2}\right) \end{aligned} \quad (5.4)$$

The resulting equivalent interaural transfer function is described by equation 5.5.

$$\underline{A}(\omega) = \frac{S_L}{S_R} \quad (5.5)$$

The absolute value and phase of $\underline{A}(\omega)$ are related to ILD and ITD, which can be converted into a localization angle. In order to simulate a center phantom source, the left and right loudspeaker should get the same input: $L = R$ and $\tau_d = 0$. Accordingly, we get the following equations:

$$|\underline{A}(\omega)| = 1 \quad (5.6)$$

$$arg\left[\underline{A}(\omega)\right] = \frac{\tau_L - \tau_R}{2}\omega \quad (5.7)$$

It should be noted that under the previous assumptions, the ILD at both ears is independent of the listening position. The ITD depends on τ_L and τ_R and hence on the listening position. Thus, using ITD, a modeled localization angle can be estimated [175].

This technique uses also Braasch's binaural model [186] as an advanced approach. This model simulates outer ear, inner ear and a decision device. The pathway to the ear is simulated by measured HRTFs for several angles. The decision device performs a cross-correlation analysis in several frequency bands and finally provides the ITD with maximum likelihood which can be converted into an azimuth angle as well. This model can also handle broadband input signals like bandpass noise. More details on the model can be found in [175, 187]. The two approaches show a considerable improvement of localization over the whole off-center listening area. Although, *cross-talk* artifacts have not been discussed in the scope of this study.

The study conducted by Microsoft research in [174] suggests a binaural-based personal 3D audio systems using only two loudspeakers, that has "unlimited" sweet spots. It relies on a webcam-based 3D face tracker to provide accurate head position and orientation information to the binaural audio system.

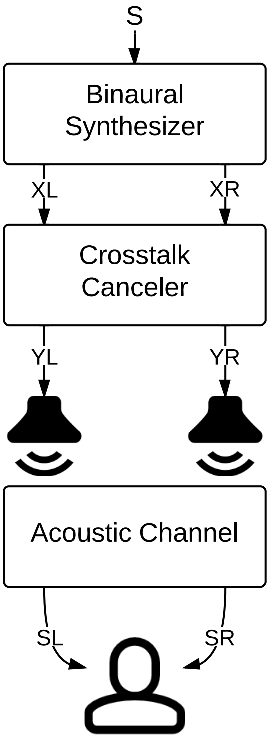


FIGURE 5.4: Block diagram of binaural audio system with loudspeakers.

The block diagram of a typical binaural audio playback system with two loudspeakers is illustrated in Figure 5.4. It consists of two main blocks: *binaural synthesizer*, and *crosstalk canceler*. The goal of the binaural synthesizer is to create spatial sounds that should be heard by the listener's ear drums, as explained in the 2^{nd} chapter. The crosstalk canceler aims to equalize the effect of *acoustic channel* during the transmission path, so that S_L and S_R match X_L and X_R. The crosstalk components are the signals S_{LR} and S_{RL} of Figure 5.3.

The acoustic paths between the loudspeakers and the listener's ears are described by an acoustic transfer matrix C, as shown in equation 5.8, where C_{LL} and C_{LR} are the transfer functions between the left loudspeaker and the left and right ears. C_{RL} and C_{RR} are the transfer functions between the right loudspeakers and the two ears.

$$C = \begin{bmatrix} C_{LL} & C_{RL} \\ C_{LR} & C_{RR} \end{bmatrix} \tag{5.8}$$

For headphones applications, the acoustic channels are completely separated, because each sound signal goes only to a specific ear. Therefore, the listener feels perfect 3D auditory experience. However, in loudspeakers applications, the paths from the *contralateral* speakers, C_{LR} and C_{RL}, also referred to as the *CrossTalk*, can destroy the 3D cues of binaural signals. The crosstalk canceler plays an essential role in this context by equalizing the transmission path between the loudspeakers and the listener. The crosstalk cancelation matrix H can be calculated by taking the inverse of the acoustic transfer matrix C (Eq. 5.9), where D denotes the determinant of the matrix C.

$$H = C^{-1} = \begin{bmatrix} C_{LL} & C_{RL} \\ C_{LR} & C_{RR} \end{bmatrix}^{-1} = \begin{bmatrix} C_{RR} & -C_{RL} \\ -C_{LR} & C_{LL} \end{bmatrix} \frac{1}{D} \tag{5.9}$$

Since the acoustic transfer functions including HRTFs are generally non-minimum phase filters, it is not always easy to calculate the inverse filter due to its instability [174]. In practice, the crosstalk canceler is adaptively obtained using the Least Mean Square algorithm (LMS) [188].

The conventional binaural audio system works well if the listener stays at the theoretical position (in-center) corresponding to the presumed binaural synthesizer HRTFs and the acoustic transfer matrix. Once the listener moves away from the sweet spot, the system performance collapses rapidly. If the system intends to keep the virtual sound sources at the same location when the head moves, the binaural synthesizer should update its HRTFs to reflect the new relative position. Consequently, the acoustic transfer matrix C needs to be updated too, which leads to a varying crosstalk canceler matrix H as well.

This system uses a 3D model based head tracker which relies on the video frames of a webcam [189]. First, the position and orientation of the listener's head is detected and tracked. The HRTF filters are then updated using the tracking information. Delays and level attenuation from the speakers to the ears are also calculated to model the new acoustic transmission channel. Finally, the filters for both binaural synthesis and crosstalk cancellation are updated. The listening tests conducted in [174] show an improvement of the virtual image localization over a wide area.

A more recent study is presented in [190]. A new stereophonic playback system is suggested, where the cross-talk signals would be reasonably cancelled at any arbitrary listener position. The system consists of two main parts: the listener position tracking system and the sound rendering block. The position of the listener was estimated using acoustic signals from the listener (i.e. voice or hand-clapping signals). A direction of arrival (DOA) algorithm was adopted to estimate the directions of acoustic sources where the room reverberation effects were taken into consideration. A Crosstalk cancellation filter was designed using a free-field model. To determine the maximum tolerable shift of the listener position, a quantitative analysis of the channel separation ratio according to the displacement of the listener position was performed. The results showed that the average mean square error between the true direction of a listener and the estimated direction was about 5°. Most of the tested subjects indicated that better stereo images were obtained by the proposed system, compared with the non-processed signals.

It should be noted that the use of CrossTalk cancellation filters can introduce severe spectral coloration to the signal [191]. Moreover, the subjective nature of HRTFs can introduce a considerable bias. One of the most successful filters is BACCH, where the spectral coloration is reduced at the cost of lowering the level of the cancellation [192].

The results of the previous methods indicate that the adjustment of audio signals to match the listener's head position does improve the localization over the whole listening area and the overall listening experience. Their main and common limitation is that they mainly and only deal with stereophonic systems. As far as we know, no extension or discussion of a potential use of such technique in the context of surround sound systems have been established. This is actually the main motivation of this part of the study where we address this sweet spot problem primarily in the context of cinemas and large listening spaces. An extension to home-cinema application is also suggested.

5.3 Suggested Approach

To enhance the listening experience, the sweet spot has to be put at the listener's position even when he/she is moving around. This is achieved by identifying a listener, tracking the identified listener actual position and then adjust the audio signals. In the following sections, we present how these steps are performed in our study.

5.3.1 Position detection

Listener tracking is the main component of our system. The previously presented techniques rely either on a head tracking model using a webcam or a sensor that the listener has to carry around. Keeping in mind the cinemas and home-cinemas context of our application, these approaches can be constraining considering the following aspects:

- Bearing in mind that such environments are designed for multiple listeners, providing each one with a portable sensor is not practical nor cost-effective (e.g. the issues with 3D glasses in movie theaters).

- Unless the cinema is under construction, equipping each seat with a sensor (e.g. piezoelectric transducer) is not practical either.
- Using standard cameras for listener tracking is not possible in a movie theater. This is due to the optical sensor sensitivity to lighting conditions while the ambient light in cinemas is generally low (almost dark) and varying depending on the movie.

For these reasons, a non-contact and non-intrusive approach is the most appropriate way to listeners tracking in such environments. It also should be computationally efficient to allow real-time tracking, independent of lighting conditions and cost-effective.

Accordingly, we established our tracking system on the basis of *Thermal Imaging* or *Thermography*. A thermal imaging camera records the intensity of radiation in the infrared part of the electromagnetic spectrum and converts it to a visible image. The existence of infrared was discovered in 1800 by astronomer Sir F. W. Herschel. Curious to the thermal difference between different light colors, he directed sunlight through a glass prism to create a spectrum and then measured the temperature of each color. He found that the temperatures of the colors increased from the violet to the red part of the spectrum. After noticing this pattern Herschel decided to measure the temperature just beyond the red portion of the spectrum in a region where no sunlight was visible. To his surprise, he found that this region had the highest temperature of all [193].

Infrared radiation lies between the visible and microwave portions of the electromagnetic spectrum, as illustrated in Figure 5.5. The primary source of infrared radiation is heat or thermal radiation. Any object that has a temperature above the absolute zero (-273.15 $°C$ or 0 $°K$) emits radiation in the infrared region. Even objects that we think of as being very cold, such as ice cubes, emit infrared radiation. Thermography can be simply seen as an imaging technique where each pixel is actually a temperature measurement.

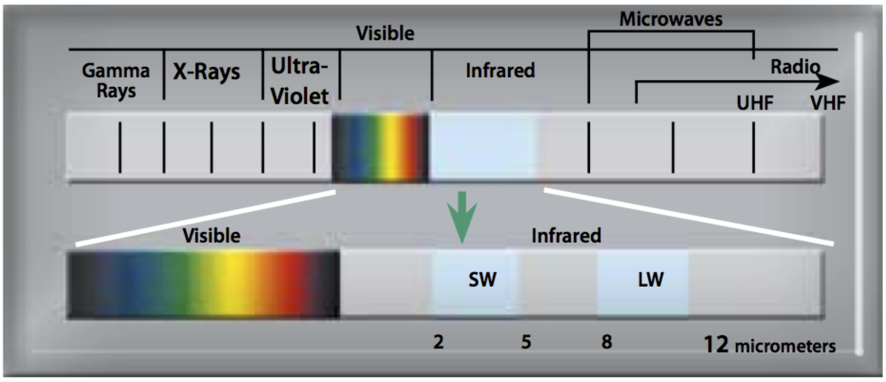


FIGURE 5.5: The infrared portion of the electromagnetic spectrum. Credits: [193].

Two temperatures are important in the human body: core temperature and skin temperature. Core temperature is the temperature of the brain and inner body, and this temperature extends to about 2 cm beneath the body's surface. Skin temperature is the temperature at the skin's surface [194]. Extra heat can come to the body from internal or external sources. Internal sources raise core temperature first and skin temperature second. An example of an internal source is the muscular activity that may result from an exercise. The sun, hot air, hot surfaces, or other external sources raise skin temperature first and core temperature second. Because heat flows from areas of warmth to areas of coolness, core temperature and skin temperature also affect each other.

Average core temperature, which is 37 °C, can rise to 38°C or 40°C during physical activity. During sleep, core temperature usually drops by 1 °C. The average comfort zone of skin temperature is 33 °C to 34 °C. If skin temperature drops to 32 °C, the person feels cold and a person will lose all sensation in skin with a temperature of 15 °C or less. When skin temperature reaches a range of 35 °C to 39 °C, individuals will feel warm or hot. At 39 °C to 41 °C, people feel pain. When skin temperature is greater than 41°C, burning pain begins. Rapid tissue damage to skin occurs when skin temperatures is of 45 °C and above [194].

Human skin is an almost perfect emitter of infrared radiation in the spectral region beyond 3 microns. This energy may be recorded as a thermogram to yield a quantitative temperature map of the skin [195].

Using these properties, our thermal-imaging-based approach has the advantage of being able to identify the listeners from other objects and track their position on the same time, in a non-intrusive way. Figures 5.6 and 5.7 illustrates how we used a thermal camera to identify the listener relative position in a classroom.

The approach consists of identifying the listeners positions on the thermal image and then calculate the geometric center of their respective locations. Accordingly, the new sweet spot position will be moved toward the highest spatial concentration of the listeners.

However, during the preliminary phase of this study, we encountered few constraints related to thermal imaging use in the context of cinemas: camera placement, thermal noise and scaling issues.

Audio adjustment requires to localize the listeners heads positions. Fortunately, the hottest point of the human skin is on the face, mainly around the center of the *T-zone*[2] $(33.5-35.5\ ^\circ C)$ [196]. Accordingly, the camera position should be adjusted to capture all

[2] Infrared frontal thermometer are based on this aspect.

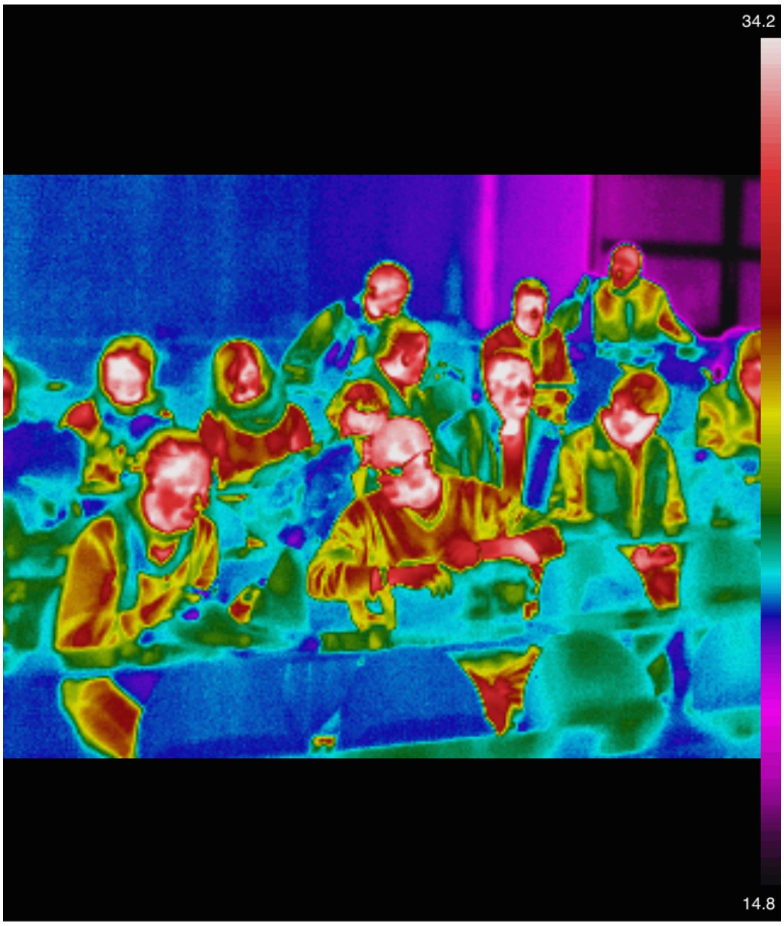


FIGURE 5.6: A thermal image taken on a classroom during our tests.

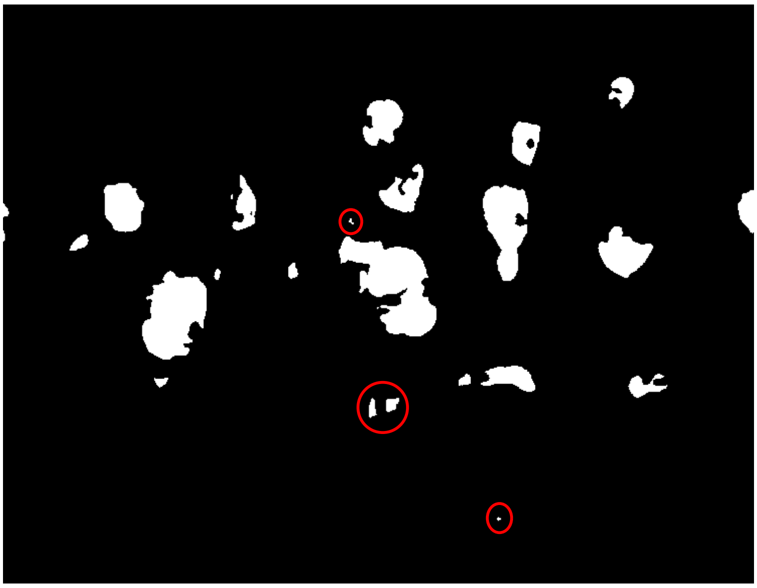


FIGURE 5.7: People identification by image segmentation using a simple *thresholding*.

the listeners faces. Considering the descending seats rows configuration of most movie theaters (as illustrated in Figure 5.8), we found that the most optimal position will be at the top of the screen with a slight downward orientation to guarantee a better coverage of the audience[3] (Fig. 5.9). Additionally, this location reduces the potential thermal noise as it will hide some body parts that may compromise the detection. With a proper calibration, the whole audience can be fit within the Depth Of Field (DOF) of the thermal sensor. Otherwise, placing the camera at a different position might involve using multiple sensors which will increase the system cost and the processing complexity.

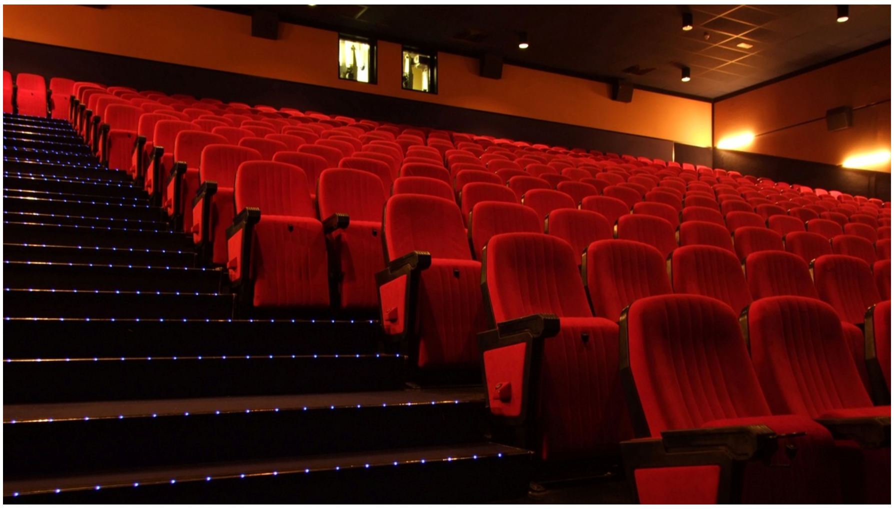


FIGURE 5.8: Movie Theater seats organization. Credits: electricmeat.net.

Thermal noise should also be considered. In fact, even with the previous optimal positioning, other body parts or hot objects can be detected and compromise the head position identification. For computational efficiency, we used a simple two-points thresholding technique to maintain only the listeners faces on the image, assuming that the camera has been properly calibrated. For the resulting gray-scaled image, $I(i,j)$, the resulting binary thresholded image, $T(i,j)$, is described by the following equations:

$$T(i,j) = 1 \quad if \quad th_{Low} \leq I(i,j) \leq th_{High} \tag{5.10}$$

$$= 0 \quad otherwise. \tag{5.11}$$

For our 8-bit encoded grayscale image, the results illustrated in Figure 5.7 were obtained with $th_{Low} = 201$ and $th_{High} = 235$.

Still, few noisy pixels can pass through the thresholding phase (red circles in Fig. 5.7). To resolve this issue, we use a morphological *area opening* filter to remove the small objects [197]. The filter removes from the binary image all connected components

[3] Although it may seem tempting, placing the camera at the top of the audience is not practical.

(objects) that have fewer than the minimum defined pixels number. The area opening algorithm operates in three steps: it determines the connected components[4] (8-connected neighborhood in our context), computes the area of each component and finally remove small objects which area is below the threshold. For Matlab or OpenCV environments, the `bwareaopen` or `removeSmallBlobs` functions can be used to perform this task.

Finally, it should be noted that due to the depth difference between the row seats, the computed area of a front listener's face will seem bigger than a listener's one at the back of the room, since the face is closer to the camera. This scaling aspect should be taken into consideration during the area opening in order to avoid removing actual listeners. Moreover, it is even more important for the centroid calculation in order to perform a more reliable geometric averaging that takes into account the area differences.

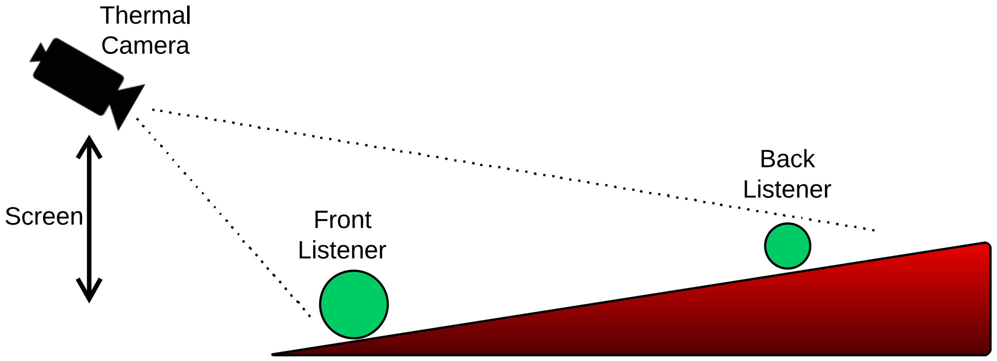


Figure 5.9: Area reduction resulting from the depth difference. With a proper calibration, the whole audience can be fit within the sensor's Depth Of Field.

Assuming that the audience is within the sensor's FOV, the surface ratio between the nearest and farthest listener face can be analytically estimated using Lens and Magnification equations. If the sensors characteristics are not available, it can be easily obtained during the calibration phase. Considering the distance d_F from the front listener to the sensor and the lens *focal length* f, we can establish the following equations, where d_{Fi} is the image distance:

$$\frac{1}{f} = \frac{1}{d_F} + \frac{1}{d_{Fi}} \tag{5.12}$$

$$M = \frac{h_{Fi}}{h_o} = -\frac{d_{Fi}}{d_F} \tag{5.13}$$

where M is the magnification ratio linking the initial object (face) size, h_o, to the image size, h_i. The image size can be calculated then using equation 5.14.

[4] Connected-component labeling.

$$h_{Fi} = -h_o \frac{d_{Fi}}{d_F} = -h_o \frac{1}{d_F} \frac{f.d_F}{d_F - f} = -h_o \frac{f}{d_F - f} \tag{5.14}$$

Considering the distance $L = d_B - d_F$ separating the Front and Back listeners, we can calculate the the image size for the farthest listener, h_{Bi}, as follows:

$$h_{Bi} = -h_o \frac{d_{Bi}}{d_B} = -h_o \frac{1}{d_B} \frac{f.d_B}{d_B - f} = -h_o \frac{f}{d_B - f} \tag{5.15}$$

Assuming that the two listeners have approximately the same face height, the ratio between the two faces sizes can be calculated as follow:

$$\omega_L = \frac{h_{Fi}}{h_{Bi}} = \frac{d_B - f}{d_F - f} = \frac{d_F + L - f}{d_F - f} \tag{5.16}$$

As described by the equation, the ω_L is independent of the head size and requires only some basic knowledge of the room configuration and the thermal camera characteristics. This factor is used as a weight for the area calculation of the listeners on the back, so that the audience spatial concentration can be realistically calculated.

To summarize, the optimal position is calculated as follows:

- The grayscale image is thresholded to keep only the listeners faces,
- Small objects (thermal noise) are removed using an area opening filter,
- Morphological transformations (dilatations and erosions) are applied [197] to compensate for potential image inconsistencies that may occurs mainly at the front row (e.g. listener's glasses in Fig. 5.7),
- The resulting zones are labeled and their centroids are calculated,
- The centroids are weighted according to Y-axis to reflect the depth difference, using equation 5.16.
- Finally the new position is calculated as described in equation 5.17.

$$\overline{p}_{opt} = \frac{\sum_{i=1}^{n} \omega_i p_i}{\sum_{i=1}^{n} \omega_i} \tag{5.17}$$

In a smaller listening environment such as home-cinema, the Kinect can be directly used for head tracking using the functionalities presented in the previous chapter.

5.3.2 Audio Processing

Once the new listeners location is calculated, the multichannel audio signals are processed to adjust to the new sweet spot position. Our approach consists of considering each channel of the multichannel sound as a virtual source that need to be placed around the listener's new position (Fig. 5.10).

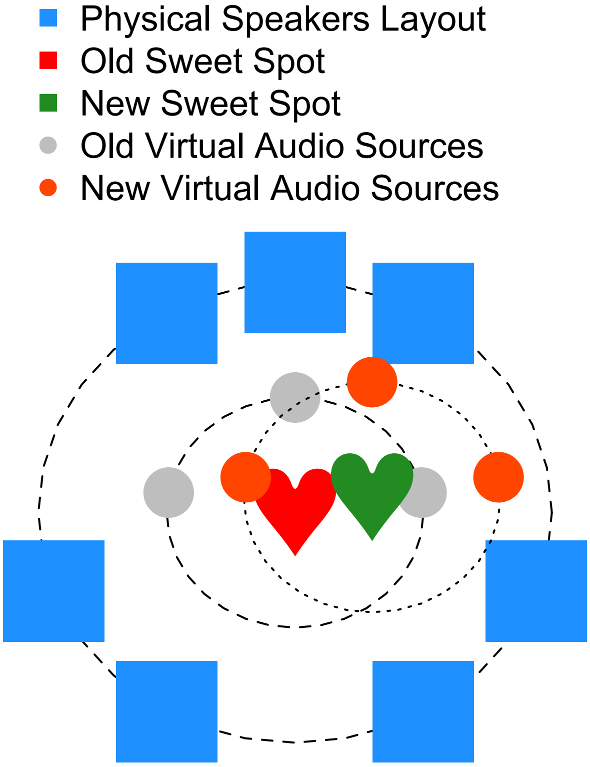


FIGURE 5.10: The system virtually moves the sound sources to adjust to the new sweet spot position.

For this purpose, we use the Vector Base Amplitude Panning (VBAP) technique introduced by V. Pulkki in [18], as presented in the 2^{nd} chapter.

VBAP is an amplitude panning method to position virtual sources in any arbitrary 2D or 3D loudspeaker setups. Amplitude panning consists of applying the same sound signal to a number of loudspeakers with appropriate non-zero amplitudes. In a 2D setup, VBAP is a reformulation of the well-known pair-wise panning method. Its advantage is that it can be generalized to 3D loudspeakers setups using a triplet-wise panning formulation. The sound signal is then applied to one, two, or three loudspeakers simultaneously. VBAP has the advantage over the previous methods to adjust to any arbitrary speakers layout [17].

VBAP is a method to calculate gain factors for pair-wise or triplet-wise amplitude panning. In a 2D context, as illustrated in Figure 5.11, it is a vector reformulation of the tangent law [198].

The base is defined by unit-length vectors $l_1 = \begin{bmatrix} l_{11} & l_{12} \end{bmatrix}^T$ and $l_2 = \begin{bmatrix} l_{21} & l_{22} \end{bmatrix}^T$, which are pointing toward the two loudspeakers. T denotes the matrix transpose. The unit-length vector $p = \begin{bmatrix} p_1 & p_2 \end{bmatrix}^T$ that refers to the virtual source position can be considered as a linear combination of the loudspeakers vectors, as described in equation 5.18.

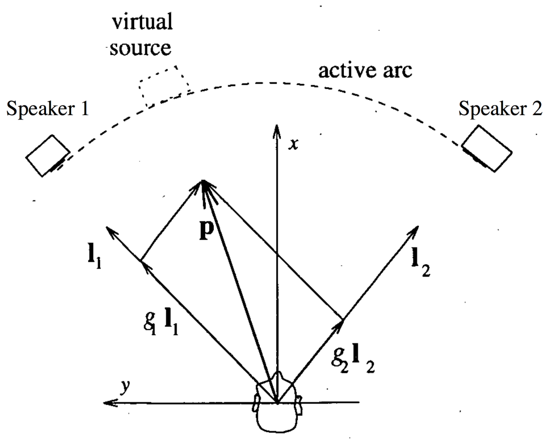


FIGURE 5.11: Virtual sources positioning using VBAP in a 2D context. Based on [18]

$$p = g_1 l_1 + g_2 l_2 \Rightarrow p^T = g.L_{12} \tag{5.18}$$

where g_1 and g_2 are non-negative scalar variables representing the gain factors, $g = \begin{bmatrix} g_1 & g_2 \end{bmatrix}$ and $L_{12} = \begin{bmatrix} l_1 & l_2 \end{bmatrix}^T$. Finally, if L_{12}^{-1} exists, the gain factors can be obtained using the equation 5.19.

$$g = p^T L_{12}^{-1} = \begin{bmatrix} p_1 & p_2 \end{bmatrix} \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{bmatrix}^{-1} \tag{5.19}$$

The loudspeakers layout in most surround sound systems is generally 2D. In other words, they are at same level with no height channels. Accordingly, when a listener is moving around, we simply move the virtual source positions to maintain them around the listener. Each source is generated using the closest loudspeakers pair to its virtual position. For each movement, the loudspeakers selection as well as the signals gains are dynamically updated to fit the new sweet spot position.

For cinemas with the loudspeakers at different heights (height channels), the 3D model is applicable. In a 3D context, VBAP uses a triplet-wise amplitude panning, as illustrated in figure 5.12.

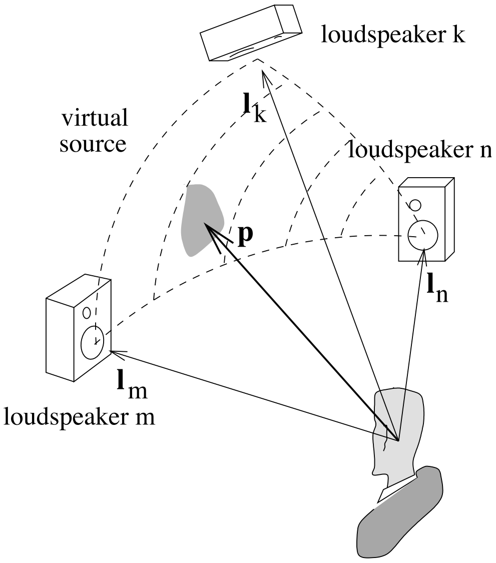


FIGURE 5.12: Virtual sources positioning using VBAP in a 3D context. Credits: [17]

The 2D formulation can now be generalized. Similarly, the listening configuration can be formulated with vectors: from a listening position, a cartesian unit vector $l_\alpha = \begin{bmatrix} l_{\alpha 1} & l_{\alpha 2} & l_{\alpha 3} \end{bmatrix}^T$ points to the direction of loudspeaker $\alpha = n, m, k$. Consequently, the panning direction of a virtual source is defined as a 3D unit vector $p = \begin{bmatrix} p_n & p_m & p_k \end{bmatrix}^T$ which is expressed as the linear combination of vectors l_n, l_m and l_k as described in equation 5.20.

$$p = g_n l_n + g_m l_m + g_k l_k \Rightarrow p^T = g.L_{nmk} \tag{5.20}$$

Finally, the gain factors vector $g = \begin{bmatrix} g_1 & g_2 & g_3 \end{bmatrix}$ can be solved as described in equation 5.21.

$$g = p^T L_{nmk}^{-1} = \begin{bmatrix} p_n & p_m & p_k \end{bmatrix} \begin{bmatrix} l_{n1} & l_{n2} & l_{n3} \\ l_{m1} & l_{m2} & l_{m3} \\ l_{k1} & l_{k2} & l_{k3} \end{bmatrix}^{-1} \tag{5.21}$$

For loudspeakers-triplet selection, we use the automated method presented in [199]. The chosen loudspeakers should meet the following requirements: the loudspeakers of a triplet cannot be all in the same plane with the listener and the triangles should not be overlapping and should have sides as short as possible. When the loudspeakers are placed in a unit-length orthogonal grid, the gain factors are simply equivalent to the cartesian coordinates, which enhances the computation efficiency.

5.4 Results & Discussion

Since the thermal camera was rented for a short period, most of this part has been studied theoretically using a Google SketchUp[5] simplified 3D model of the listening environment. The advantage of this approach is the possibility to easily simulate the listeners and camera positioning for different scenarios (Fig. 5.13).

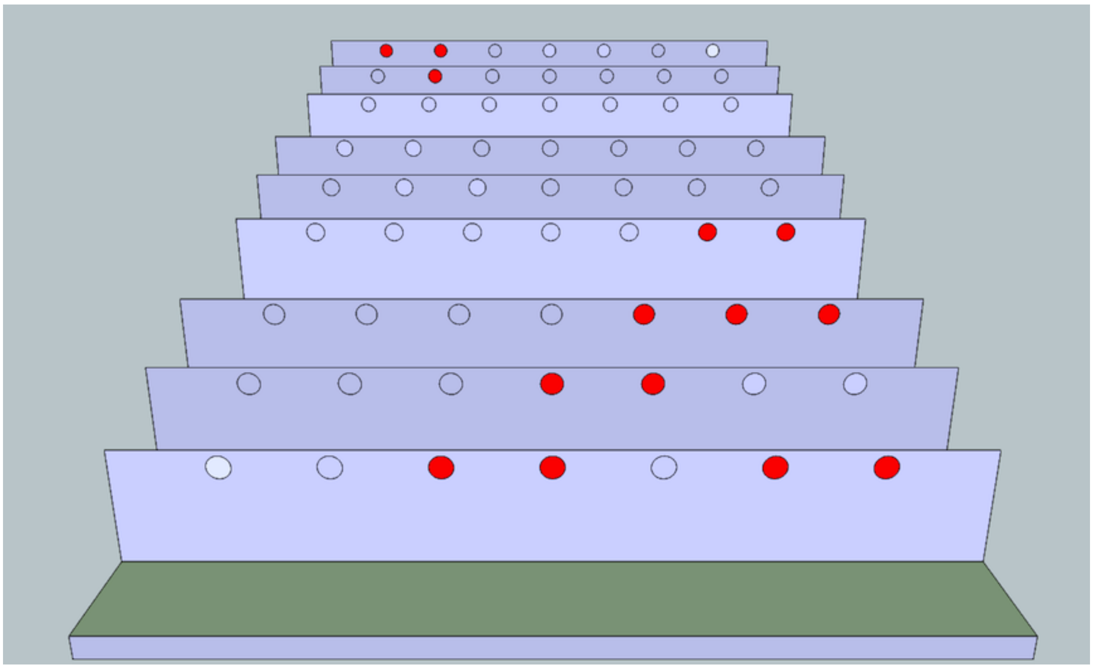


FIGURE 5.13: A simplified 3D model of the listening environment. We can notice the size difference between the front and back rows.

The sensor that we experimented with was the FLIR E60 [6]. It has a 320x240 pixels resolution with a ±2% temperature measurement accuracy. Despite its relatively low resolution, such sensor can be used in our context since we are interested in locating the audience spatial concentration instead of a listener identification. We recommend to use the camera vertical so that the depth information can be captured more efficiently (e.g. for the presented model, the 320px side should be on the Y-axis). The tracking model has been simulated using Matlab and openCV.

Using the 3D room model, we were able to assess the localization performance under different conditions. Figures 5.14, 5.15 and 5.16 show how the model is still capable of tracking the audience position despite the noisy environment.

[5] www.sketchup.com
[6] www.flir.com

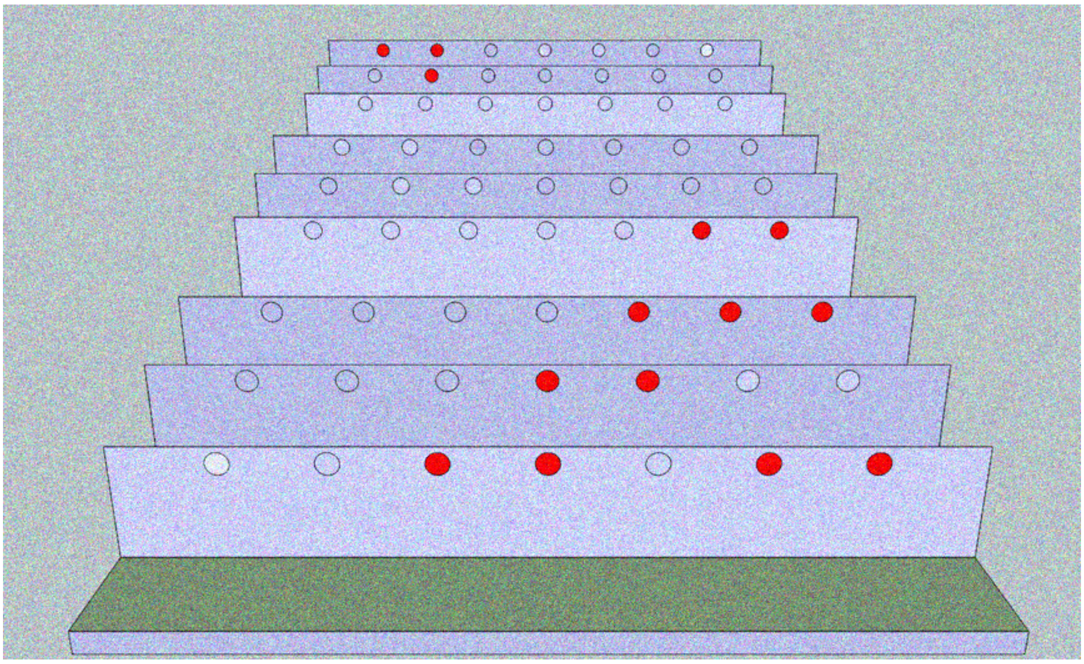


Figure 5.14: Simulation of a noisy thermal image

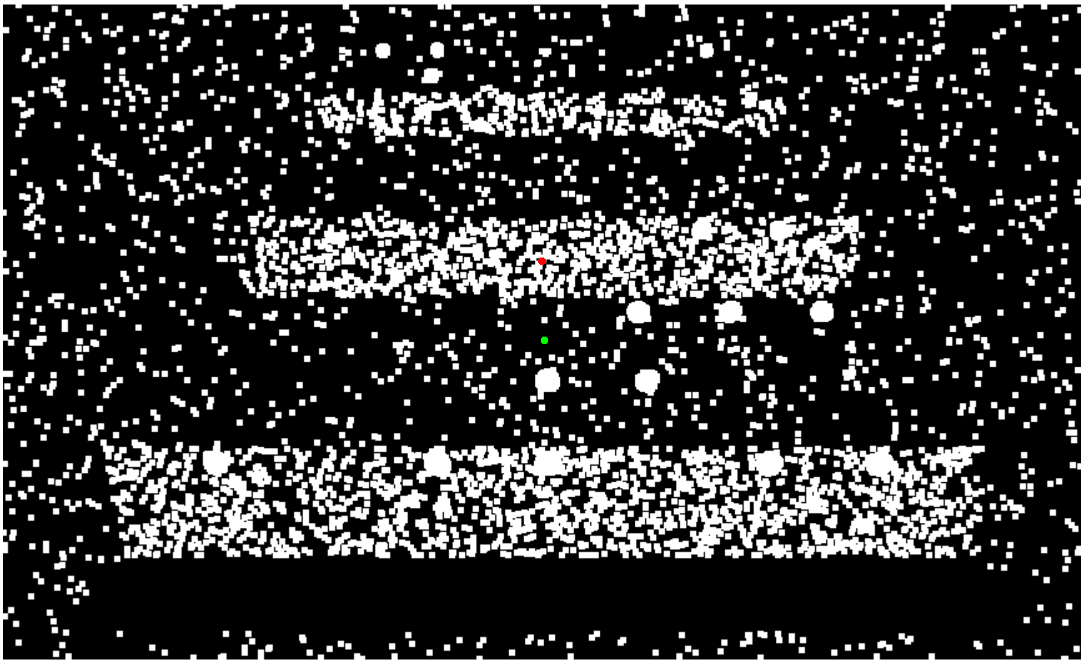


Figure 5.15: Thresholding result of the noisy image.

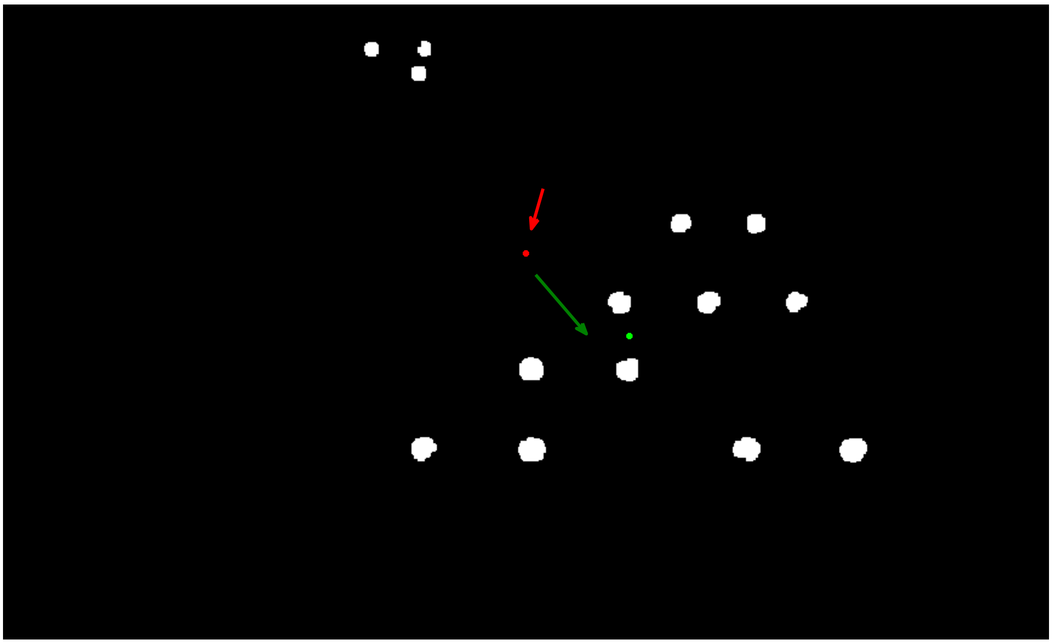


FIGURE 5.16: Listener position tracking. The model is still capable of tracking the listener location even in a noisy environment. The red dot is the initial Sweet Spot and the green one is the new one.

It should be noted that this approach is only applicable within the geometrical boundaries of the loudspeakers layout. In VBAP, as in all amplitude panning methods, the virtual source can not be positioned outside the active arc or region. If such property is needed, a more advanced technique like WFS is then required.

The current model limitation is that it does not handle head orientation which affects the localization performance. In the context of cinemas, we can just assume that the listeners are looking at the screen, and therefore facing the thermal camera. In a home cinema application, the Kinect RGB camera can be used to track such information. The audio adjustment will be better performed using HRTFs and cross-talk cancellation filters.

Using the gesture-based protocol presented in the previous chapter, we were able to assess the localization accuracy of this approach. The preliminary results on a small room show an improvement of the localization performance.

Chapter 6

Conclusion & Outlook

In this work, several constraints regarding the use of spatial sound in cinema and pro audio applications were presented and potential solutions were investigated.

Three complementary topics concerning multichannel audio use in professional applications have been presented. *SIRIUS*, an *audio transport* mechanism was introduced in the third chapter. It is designed to convey multiple professional-grade audio channels over a regular LAN while maintaining their synchronization. The system reliability is also enhanced by using a FEC mechanism and a selective redundancy, without causing any important network overload. Finally, the system offers a low latency that meet the professional applications requirements. Our architecture is based on IPv6, and unlike many of other audio transport technologies, can work on any of the existing infrastructures and coexist with other IT traffic. The system relies on standard protocols and offers a high level of interoperability with equivalent technologies. The system performances have been objectively and subjectively validated and comply with professional applications requirements.

The second contribution was *AQUA*, a novel framework for multichannel audio *quality assessment* that provides efficient tools for both subjective and objective quality analysis. The subjective layer consists of a new approach for designing reliable listening tests for multichannel sound that analyze the perceptual as well as the spatial information. Audio localization accuracy can be reliably evaluated using our *gesture-based* protocol build around the Kinect. Additionally, the method relies on EEG signals analysis for psychological biases monitoring and efficient subjects screening.

The objective method uses a subjectively-validated *binaural model* to down-mix the multichannel audio signal into a 2-channels binaural mix that keeps the spatial cues and provides a simple and scalable analysis. The binaural stream is then analyzed by a perceptual and spatial model that calculate several cues. Their combination is equivalent to the internal representation. The resulting variables are combined through an ANN that finally estimates the objective quality grade. In parallel, the psychological model simulate the human behavior by adjusting the output grades according to the previous ones (i.e., the *experience effect*). The overall performance shows that AQUA model can accurately predict the perceptual and spatial quality of a multichannel audio in a very realistic manner.

The third focus of our study was to optimize the listening experience for surround sound systems. Considering the *sweet spot* issue with these systems and the complexity of broadening solutions that imply specific measurements or the control of loudspeakers directivity, we introduced a tracking technique that virtually moves the sweet spot location to the actual position of listener(s). Our approach is non-intrusive and relies on thermal imaging for listeners identification and tracking. The channels of the original stream are considered as virtual sources and remixed using the VBAP technique. Accordingly, the audio system virtually *follows* the listener actual position. For home-cinema application, the kinect can be used for the tracking part and the audio adjustment can be done using HRTFs and cross-talk cancellation filters. The overall performance shows an increase of localization accuracy and improvement of the listening experience.

For each part, there is several directions that can be investigated in the future:

- Audio Transport
  - The synchronization accuracy using PTP can be enhanced to meet the AES11 requirements. This will imply hardware modifications and changes at the lower layers.
  - Power Over Ethernet (PoE) and Power-Line Communications (PLC) will be interesting features for professional applications. A further investigation of how they affect the network performances must be established to guarantee that they satisfy the pro audio requirements.
  - Wireless Transmission is also a very appealing feature. The new generation of the WiFi standard (802.11ac) offers a higher stability and a larger bandwidth and its use should be investigated for latency-tolerant audio applications.
- Quality Assessment
  - To enhance the system performances, the training set should be increased and include more audio signals with more distortions types.

  - More perceptual and spatial cues can be investigated and integrated to enhance the overall performance. The analytic expression of LEV and ASW would be a good start.
  - The gesture-based protocol can be enhanced by adding a new sensor (slider for instance) to integrate the distance information.
  - A deeper study of the EEG signals can offer a new approach for quality assessment. A more accurate and less noise-sensitive headset should be used though for a reliable analysis.
  - More subjective data will provide a better psychological biases modeling. A good start will be to consider additional bio-feedback signal to monitor such aspects in a more reliable manner.
  - The current technique can be combined with L2TI video and image quality metrics to establish a more comprehensive multimedia quality assessment technique. The interaction between audio and video distortions should also be investigated.

- Sound Optimization
  - The tracking performance was mainly validating in a simulation environment. A validation in realistic conditions is still required and will be performed as soon as the thermal cameras are available.
  - The potential use of Ambisonics or higher order systems can be investigated for a better optimization.
  - The system will benefit from the use of MDA format. The current system should be adjusted to render MDA audio objects.

The code sources, simulation algorithms and the subjective data measured during this study will be available soon on L2TI website.

Appendix A

Existing Audio Transport Technologies

Thanks to their numerous advantages, digital audio Networks are more appealing for professional applications. Table A.1 presents a non-exhaustive list of the most used ones. More details are available in [33] and their respective websites.

N.A: Details on this aspect are not available.

(*) 802.1 can enhance the system reliability by using STP/LACP.

Table A.1: A non-exhaustive list of digital audio transport technologies

Technology	**Transport**	**Reliability Enhancement**	**Synchronization**	**Latency**	**Control**	**Mixed Use**
AES47/AES51	ATM/Ethernet	Based on ATM	N.A	125 μs/hop	ATM or IP-based protocol	Coexist with ATM traffic
AES50, SuperMAC, HyperMAC	Ethernet	Redundant link	Master-Slave WordClock	63 μs	Ethernet-based	Dedicated Network
Aviom Pro64	Ethernet	Redundant link	WordClock	322 μs+1.34 μs/hop	Proprietary	Dedicated
CobraNet	Ethernet	Provided by 802.1	Synch Frames	1.33-5.33 ms	Ethernet, SNMP	Partial coexistence
EtherSound	Ethernet	None	PLL	84-125 μs+0.5 μs/hop	Proprietary	Dedicated
AVB	Ethernet	802.1Qat, 802.1Qav	802.1AS	2ms/50ms	IP tunneling	Dedicated
Dante	IP network	QoS, redundant link	PTP	84 μs	IP-based proprietary	Coexists with other trafic
Ravenna (AES67)	IP network	Redundancy	PTPv2	5 ms	N.A	Coexists
LiveWire (Axia)	IP network	Provided by 802.1	N.A	750 μs	HTTP, XML	Coexists
Wheatnet	IP network	N.A	WordClock	0.5 ms	SNMP, proprietary	Dedicated

Appendix B

Details on SIRIUS packet format

As presented in section 3.3, the SIRIUS packet format is described in Figure B.1:

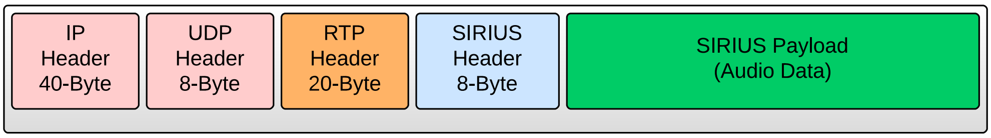


Figure B.1: General Sirius Packet Format

SIRIUS offers two packet types: standard (Fig. B.2) and time-critical (Fig. B.3):

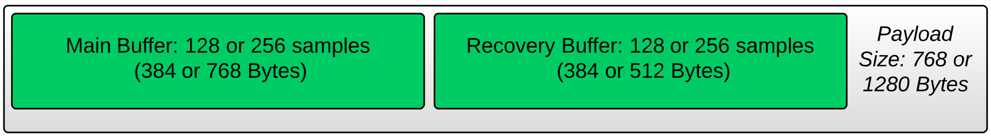

Main Buffer: 128 or 256 samples (384 or 768 Bytes)
Recovery Buffer: 128 or 256 samples (384 or 512 Bytes)
Payload Size: 768 or 1280 Bytes

Figure B.2: Sirius Packet Format: standard mode

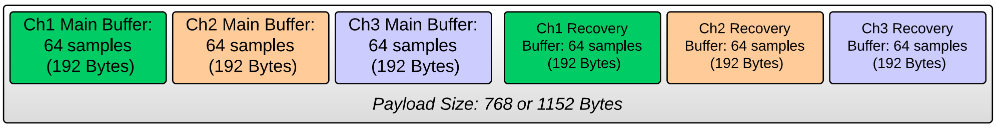

Ch1 Main Buffer: 64 samples (192 Bytes)
Ch2 Main Buffer: 64 samples (192 Bytes)
Ch3 Main Buffer: 64 samples (192 Bytes)
Ch1 Recovery Buffer: 64 samples (192 Bytes)
Ch2 Recovery Buffer: 64 samples (192 Bytes)
Ch3 Recovery Buffer: 64 samples (192 Bytes)
Payload Size: 768 or 1152 Bytes

Figure B.3: Sirius Packet Format: time-critical mode

Each packet contains the audio data samples and the recovery information for the previous packet. If packet loss is detected (using sequence number), the audio data are extracted from the recovery information. Additionally, time-critical mode can convey up to three channels within the same packet, so that they get delivered almost at the same time using IP multicast.

SIRIUS header format is illustrated in Figure B.4:

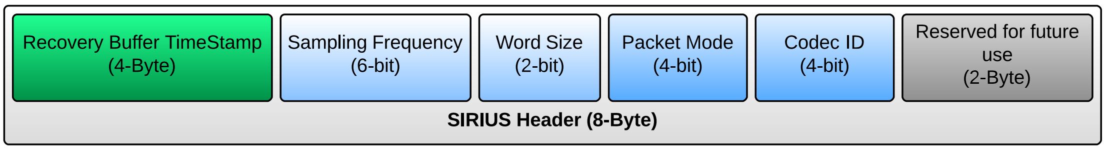

Recovery Buffer TimeStamp (4-Byte)
Sampling Frequency (6-bit)
Word Size (2-bit)
Packet Mode (4-bit)
Codec ID (4-bit)
Reserved for future use (2-Byte)
SIRIUS Header (8-Byte)

FIGURE B.4: Sirius Packet Format: Header Format

TimeStamp of the recovery data is integrated so that the restored samples are played on the right time. The signal sampling frequency (Table B.1) and word size (audio resolution) are included as well in order to allow an automatic configuration of the playback sound card.

TABLE B.1: Sampling Frequencies Coding

Frequency (KHz)	**Used in**	**Value**
8	Telephone, wireless microphone	0x05
11,025	Low quality PCM, MPEG audio and audio analysis of subwoofer	0x04
16	Modern VoIP and VVoIP	0x03
32	HQ wireless microphones, Digital FM radio	0x02
44,1	Audio CD, MPEG-1	0x01
48	Standard Professional Audio	0x00 (default)
88,2	Professional recording equipment intended for CD	0x0A
96	DVD Audio, HD DVD, Blu-Ray	0x0B
192	DVD-Audio, some LPCM DVD tracks, Blu-ray, HD DVD	0x0C
352,8	Digital eXtreme Definition, used for recording and editing Super Audio CDs	0x0D

The 2-LSB of Packet Mode field describe whether standard or critical mode is used and the channels number within the packet. The 2-MSB indicate the codec used to generate the recovery data. Codec-ID describes the used codec and 2-Byte are reserved for future use.

TABLE B.2: Packet Mode Coding

Rec. Codec	**Value**	**Mode**	**Value**
Truncation with Dithering	0x00	Standard	0x00
OPUS	0x01	Time-Critical (2-channels)	0x01
PCM	0x02	Time-Critical (3-channels)	0x02
Custom	0x03	Custom	0x03

Appendix C

Notes on PEAQ MOVs calculation

Here is a brief description of how the PEAQ's MOVs are calculated. More details are available in [133].

At the input of the Perceptual Model, the Reference Signal and the Signal under Test are adjusted to the assumed playback level and sent through a high pass filter in order to remove DC and subsonic components of the signals. The two signals are then decomposed into band pass signals using our Gammatone filter-bank and a frequency dependent weighting is applied in order to model the spectral characteristics of the outer and middle ear. The level dependent spectral resolution of the auditory filters is modeled by a frequency domain convolution of the outputs with a level dependent spreading function.

The signals envelopes of the filter-bank outputs are calculated using the Hilbert-transform and a convolution with a window function is applied in order to model the *backward masking*. A frequency dependent offset is then added and accounts for internal noise in the auditory system and models the threshold in quiet. Finally, a second convolution is carried out using an exponential spreading function that accounts for *forward masking*. At this stage, we obtained the *excitation patterns* which are used to compute specific *loudness patterns* and *modulation patterns*. These three types of patterns are the basis on which the MOVs are calculated.

Considering, $E_2[k,n]$, the energies at the filter outputs combined with the internal noise and the masking effects, for n^{th} time index and k^{th} band, the excitation patterns can be described as follow:

$$
\begin{aligned}
\overline{E}_{der}[k,n] &= a.\overline{E}_{der}[k,n-1] + (1-a).\frac{f_s}{192}.\left|E_2[k,n]^{0.3} - E_2[k,n-1]^{0.3}\right| && \text{(C.1)}\\
\overline{E}[k,n] &= a.\overline{E}[k,n-1] + (1-a).E_2[k,n]^{0.3} && \text{(C.2)}\\
Mod[k,n] &= \frac{\overline{E}_{der}[k,n]}{1+\overline{E}[k,n]/0.3} && \text{(C.3)}\\
ModDif[k,n] &= w.\frac{|Mod_{test}[k,n] - Mod_{ref}[k,n]|}{offset + Mod_{ref}[k,n]} && \text{(C.4)}
\end{aligned}
$$

By averaging the Modulation Difference, $ModDif$, over time and frequency bands, and then calculate the RMS (Root Mean Square) value, we get the `RmsModeDif` as our first MOV. Variables w and a are frequency-dependent and their calculation is available in [133].

The `SegNMR` MOV is the linear average of the local NMR described in equation C.5:

$$
\text{NMR} = 10 * log_{10}\frac{1}{N}\sum_{n}\left(\frac{1}{Z}\sum_{k=0}^{Z-1}\frac{P_{noise}[k,n]}{M[k,n]}\right) \qquad \text{(C.5)}
$$

$$
\begin{aligned}
M[k,n] &= \frac{E[k,n]}{10^{\frac{m[k]}{10}}} && \text{(C.6)}\\
E[k,n] &= a.E[k,n-1] + (1-a).E_2[k,n] && \text{(C.7)}\\
F_{noise}[k_f,n] &= ||F_{eref}[k_f,n]| - |F_{etst}[k_f,n]|| && \text{(C.8)}
\end{aligned}
$$

$$
z/Bark = 7.\ \text{arsinh}\left(\frac{f/Hz}{650}\right) \qquad \text{(C.9)}
$$

$M[k,n]$ is the *masking pattern* obtained by weighting the *excitation pattern* $E[k,n]$ with $m[k]$, a frequency-dependent weighting function. $P_{noise}[k,n]$, the *noise pattern*, is obtained by calculating the the auditory pitch scale (Bark) of F_{noise}, the error signal, using the approximation given by Schroeder et al. (Eq. C.9).

The Model Output Variable `RmsNoisLouAsy` is the weighted sum of the squared averages of the noise loudness and the loudness of lost signal components, as described by equation C.10.

$$
RmsNoisLouAsy = RmsNoiseLoud + 0,5.RmsMissingComponents \qquad \text{(C.10)}
$$

The MOV *RmsNoiseLoud* is the squared average of the noise loudness. The MOV *RmsMissingComponents* is the squared average of the noise loudness calculated from the spectrally adapted excitation patterns of test and Reference Signals interchanged in order to yield the loudness of components in the Reference Signal that are lost in the test signal.

An audio signal containing strong harmonics (e.g. clarinet) has a spectrum characterized by a number of regularly spaced peaks separated by deep valleys. Under some conditions, the error signal may inherit that structure. For example, noise mixed with such signal is more likely to remain unmasked if the signal is low in the spectral valleys. The resulting error spectrum would then contain a structure similar to the original spectrum but offset in frequency to correspond to the locations of the valleys. This structure may result in a distortion with tonal qualities that could increase the salience of the error. The error is defined as the difference in the log spectra of the reference and processed signals, each weighted by the frequency response of the outer and middle ear.

For the last MOV, `EHS` (Error's Harmonic Structure), the harmonic structure magnitude is obtained by identifying and measuring the largest peak in the spectrum of the autocorrelation function. Each correlation is calculated as the cosine of the angle between two vectors according to equation C.11, where $\overrightarrow{F}_0$ is the error vector and $\overrightarrow{F}_t$ is the same vector lagged by a certain amount.

$$C = \frac{\overrightarrow{F}_0 . \overrightarrow{F}_t}{\left|\overrightarrow{F}_0\right| \cdot \left|\overrightarrow{F}_t\right|} \tag{C.11}$$

Bibliography

[1] The physics of hearing, 2009. URL http://www.unc.edu/~beachum/.

[2] Auditory system, wikibook. URL http://en.wikibooks.org/wiki/Sensory_Systems/Auditory_System.

[3] Mathilde Beudin. Influence de la surdité sur la perception de la mélodie. Master's thesis, Ecole d'audioprothèse J.E. Bertin - Fougères, 2008.

[4] David Heeger. Auditory pathways and sound localization. Perception Lecture Notes, 2006.

[5] Benedikt Grothe, Michael Pecka, and David McAlpine. Mechanisms of sound localization in mammals. *Physiological Reviews*, 90(3):983–1012, 2010. doi: 10.1152/physrev.00026.2009.

[6] Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT Press, 1996.

[7] Brian C. J. Moore. *An Introduction to the Psychology of Hearing, Fifth Edition.* Academic Press, April 2003. ISBN 0125056281.

[8] Curtis Roads. *The Computer Music Tutorial.* MIT Press, 1996.

[9] BG Shinn-Cunningham, S Santarelli, and N Kopco. Tori of confusion: binaural localization cues for sources within reach of a listener. *The Journal of the Acoustical Society of America*, 2000.

[10] D.S. Brungart and B.D. Simpson. Auditory localization of nearby sources in a virtual audio display. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 107–110, 2001. doi: 10.1109/ASPAA.2001.969554.

[11] L R Bernstein. Auditory processing of interaural timing information: new insights. *Journal of Neuroscience Research*, 66(6):1035–1046, Dec 2001. ISSN 0360-4012 (Print); 0360-4012 (Linking).

[12] Leslie R Bernstein and Constantine Trahiotis. Enhancing sensitivity to interaural delays at high frequencies by using "transposed stimuli". *J Acoust Soc Am*, 112(3 Pt 1):1026–1036, Sep 2002. ISSN 0001-4966 (Print); 0001-4966 (Linking).

[13] Dan FM Goodman, Victor Benichoux, and Romain Brette. Decoding neural responses to temporal cues for sound localization. *eLife*, 2, 2013. doi: 10.7554/eLife.01312.

[14] Mono vs stereo. URL http://www.mcsquared.com/mono-stereo.htm.

[15] Bobby Owsinski. *The Mixing Engineer's Handbook*. Delmar Cengage Learning, 2013.

[16] Günther Theile. Multichannel natural recording based on psychoacoustic principles. In *Audio Engineering Society Convention 108*, Feb 2000. URL http://www.aes.org/e-lib/browse.cfm?elib=9182.

[17] Ville Pulkki. *Spatial Sound Generation and Perception by Amplitude Panning Techniques*. PhD thesis, Helsinki University of Technology, 2001.

[18] Ville Pulkki. Virtual sound source positioning using vector base amplitude panning. *J. Audio Eng. Soc*, 45(6):456–466, 1997. URL http://www.aes.org/e-lib/browse.cfm?elib=7853.

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

[20] Dylan Menzies and Marwan Al-Akaidi. Nearfield binaural synthesis and ambisonics. *J Acoust Soc Am*, 121(3):1559–1563, Mar 2007. ISSN 0001-4966 (Print); 0001-4966 (Linking).

[21] Florian Hollerweger. An introduction to higher order ambisonic. Tutorial, 10 2008.

[22] Michael A. Gerzon. Panpot laws for multispeaker stereo. In *Audio Engineering Society Convention 92*, Mar 1992. URL http://www.aes.org/e-lib/browse.cfm?elib=6824.

[23] D G Malham and A Myatt. 3-d sound spatialization using ambisonic techniques. *Computer music journal*, 19(4):58–70, 1995. ISSN 0148-9267.

[24] Jérome Daniel, Jean-Bernard Rault, and Jean-Dominique Polack. Ambisonics encoding of other audio formats for multiple listening conditions. In *Audio Engineering Society Convention 105*, Sep 1998. URL http://www.aes.org/e-lib/browse.cfm?elib=8385.

[25] A J Berkhout, D Vries, and P Vogel. Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America*, 1993.

[26] Günther Theile and Helmut Wittek. Wave field synthesis – a promising spatial audio rendering concept. 2007.

[27] Alois Sontacchi and Robert Höldrich. " getting mixed up with wfs, vbap, hoa, trm ..." from acronymic cacophony to a generalized rendering toolbox. *DEGA Wave Field Synthesis Work*, 2007. URL http://decoy.iki.fi/dsound/ambisonic/motherlode/source/getting.pdf.

[28] Henrik Møller, Michael Friis Sørensen, Dorte Hammershøi, and Clemen Boje Jensen. Head-related transfer functions of human subjects. *J. Audio Eng. Soc*, 43(5):300–321, 1995. URL http://www.aes.org/e-lib/browse.cfm?elib=7949.

[29] Jeroen Breebaart, Jonas Engdegård, Cornelia Falch, Oliver Hellmuth, Johannes Hilpert, Andreas Hoelzer, Jeroen Koppens, Werner Oomen, Barbara Resch, Erik Schuijers, and Leonid Terentiev. Spatial audio object coding (saoc) - the upcoming mpeg standard on parametric object based audio coding. In *Audio Engineering Society Convention 124*, May 2008. URL http://www.aes.org/e-lib/browse.cfm?elib=14507.

[30] Frank Melchior and Sascha Spors. Spatial audio reproduction: From theory to production. AES Convention Tutorial, 2010.

[31] Al Keltz. Balanced versus unbalanced lines. Pro Sound Web, 2012.

[32] Unbalanced vs balanced i/o and how to work with them. The HUB: musicians-friend.com, 2014.

[33] Best practices in network audio. AES WHITE PAPER, 2009.

[34] Stefan Schmitt and Jochen Cronemeyer. *Audio over Ethernet: There are many solutions – but which one is best for you?* DSPECIALISTS GmbH, 2011.

[35] Ethernet. Wikipedia, 2013.

[36] *IEEE 802: Generic Framing Procedure.* American National Standard for Telecommunications, 2001.

[37] *Real Time Audio Distribution Via Ethernet: CobraNet datasheet.* CIRRUS LOGIC, 2003.

[38] *CobraNet: Programmer's Reference.* CIRRUS LOGIC, 2006.

[39] *Working with CobraNet.* MediaMatrix, 2012.

[40] *Guide d'utilisation des réseaux EtherSound*. NEXO, 2004.

[41] *EtherSound Setup Guide*. Yamaha CO, 2010.

[42] IEEE-802.1. Time-sensitive networking task group. IEEE Standards, 2011.

[43] Rick Kreifeldt. Avb for professional a/v use. White Paper, 2009.

[44] Axel Holzinger and Andreas Hildebrand. Realtime linear audio distribution over networks: A comparison of layer 2 and 3 solutions using the example of ethernet avb and ravenna. In *Audio Engineering Society Conference: 44th International Conference: Audio Networking*, Nov 2011. URL http://www.aes.org/e-lib/browse.cfm?elib=16147.

[45] IEEE.Std.802.1BA-2011. *IEEE Standard for Local and metropolitan area networks — Audio Video Bridging (AVB) Systems*. IEEE 802.1, New York, USA, 2011.

[46] IEEE.Std.802.1AS-2011. *IEEE Standard for Local and metropolitan area networks — Timing and Synchronization for Time-Sensitive Applications in Bridged Local Area Networks*. IEEE 802.1, New York, USA, 2011.

[47] IEEE.Standard. *Precision Time Protocol: IEEE 1588 Version 2*. IEEE, 2008.

[48] IEEE.Std.802.1Qat-2011. *IEEE Standard for Local and metropolitan area networks — Stream Reservation Protocol*. IEEE 802.1, New York, USA, 2011.

[49] IEEE.Std.802.1Qav-2011. *IEEE Standard for Local and metropolitan area networks — Traffic Shaping*. IEEE 802.1, New York, USA, 2011.

[50] Hyung-Taek Lim, Daniel Herrscher, Martin Johannes Waltl, and Firas Chaari. Performance analysis of the ieee 802.1 ethernet audio/video bridging standard. In *Proceedings of the 5th International ICST Conference on Simulation Tools and Techniques*, SIMUTOOLS '12, pages 27–36, ICST, Brussels, Belgium, Belgium, 2012. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). ISBN 978-1-4503-1510-4. URL http://dl.acm.org/citation.cfm?id=2263019.2263024.

[51] *HiQnet: Guide to audio networking*. Harman, October 2012.

[52] *RFC 791 Internet Protocol - DARPA Inernet Programm, Protocol Specification*. Internet Engineering Task Force, September 1981. URL http://tools.ietf.org/html/rfc791.

[53] ITU-T. *A Handbook on Internet Protocol (IP)-Based Networks and Related Topics and Issues*. International Telecommunication Union, 2005.

[54] Dante: Networking related questions, 2014. URL http://www.audinate.com/index.php?option=com_content&view=article&id=197&Itemid=190.

[55] Farhan Siddiqui, Sherali Zeadally, Thabet Kacem, and Scott Fowler. Zero configuration networking: Implementation, performance and security. *Computers and Electrical Engineering*, 38(5):1129 – 1145, 2012. ISSN 0045-7906. doi: http://dx.doi.org/10.1016/j.compeleceng.2012.02.011. URL http://www.sciencedirect.com/science/article/pii/S004579061200033X. Special issue on Recent Advances in Security and Privacy in Distributed Communications and Image processing.

[56] S. Cheshire, B. Aboba, and E. Guttman. *RFC 3927: Dynamic Configuration of IPv4 Link-Local Addresses*. IETF, May 2005.

[57] *RAVENNA Operating Principles*. ALC NetworX GmbH, Munich, Germany, draft 1.0 edition, June 2011.

[58] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. *RTP: A Transport Protocol for Real-Time Applications (RFC 3550)*. IETF, July 2003.

[59] K. Nichols, S. Blake, F. Baker, and D. Black. *Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers (rfc2474)*. IETF, December 1998.

[60] Adrian Freed and Andy Schmeder. Features and future of open sound control. In *NIME*, June 2009. URL http://cnmat.berkeley.edu/node/7002.

[61] Andrew Schmeder, Adrian Freed, and David Wessel. Best practices for open sound control. In *Linux Audio Conference*, Utrecht, NL, 01/05/2010 2010.

[62] Richard Foss, Robby Gurdan, Bradley Klinkradt, and Nyasha Chigwamba. The xfn connection management and control protocol. In *Audio Engineering Society Conference: 44th International Conference: Audio Networking*, Nov 2011. URL http://www.aes.org/e-lib/browse.cfm?elib=16143.

[63] Osedum P Igumbor and Richard J Foss. A solution for integrating layer 2 controllable audio devices into a layer 3 network. In *Southern Africa Telecommunication Networks and Applications Conference (SATNAC) proceedings*, 2012.

[64] J.D. Case, M. Fedor, M.L. Schoffstall, and J. Davin. Simple Network Management Protocol (SNMP). RFC 1157 (Historic), May 1990. URL http://www.ietf.org/rfc/rfc1157.txt.

[65] Andrew Eales and Richard Foss. Towards a standard model for networked audio devices. In *Audio Engineering Society Conference: 44th International Conference: Audio Networking*. Wellington Institute of Technology, Wellington, New Zealand and Department of Computer Science, Rhodes University, Grahamstown, South Africa, Nov 2011. URL http://www.aes.org/e-lib/browse.cfm?elib=16144.

[66] Jeff Berryman. *Open Control Architecture: Technical Overview*. OCA Alliance, November 2011.

[67] *Bonjour Overview: Developer Documentation*. Apple Inc., April 2013. URL https://developer.apple.com/librarY/mac/documentation/Cocoa/Conceptual/NetServices/NetServices.pdf.

[68] Jeffrey Berryman. Technical criteria for professional media networks. In *Audio Engineering Society Conference: 44th International Conference: Audio Networking*. Bosch Communication Systems, Burnsville, MN, USA, Nov 2011. URL http://www.aes.org/e-lib/browse.cfm?elib=16145.

[69] *AES11-2009: AES recommended practice for digital audio engineering - Synchronization of digital audio equipment in studio operations*. Audio Engineering Society, Inc., February 2010.

[70] Christoph Sladeczek, Thomas Reussner, Michael Rath, Kar Preidl, Hermann Scheck, and Sandra Brix. Audio network based massive multichannel loudspeaker system for flexible use in spatial audio research. In *Audio Engineering Society Conference: 44th International Conference: Audio Networking*, Nov 2011. URL http://www.aes.org/e-lib/browse.cfm?elib=16135.

[71] A. Smimite, Ken Chen, and A. Beghdadi. Next-generation audio networking engineering for professional applications. In *Telecommunications Forum (TELFOR), 2012 20th*, pages 1252–1255, 2012. doi: 10.1109/TELFOR.2012.6419443.

[72] EBU-UER. *EBU-TECH 3326: Audio contribution over IP*. Geneva, April 2008.

[73] Mathias Coinchon and Lars Jonsson. Ebu tech.doc. 3326 for interoperability between audio over ip units. In *Audio Engineering Society Convention 124*, May 2008. URL http://www.aes.org/e-lib/browse.cfm?elib=14452.

[74] Lars Jonsson. Ebu standardisation of audio over ip for contribution systems. In *Audio Engineering Society Conference: 44th International Conference: Audio Networking*. Swedish Radio R&D Stockholm, Sweden, Nov 2011. URL http://www.aes.org/e-lib/browse.cfm?elib=16134.

[75] *AES67-2013: AES standard for audio applications of networks - High-performance streaming audio-over-IP interoperability.* Audio Engineering Society, March 2014.

[76] S. Deering and R. Hinden. *RFC 2460 Internet Protocol, Version 6 (IPv6) Specification.* Internet Engineering Task Force, December 1998. URL http://tools.ietf.org/html/rfc2460.

[77] S. Amante, B. Carpenter, S. Jiang, and J. Rajahalme. *RFC 6437: IPv6 Flow Label Specification.* IETF, November 2011.

[78] S. Thomson, T. Narten, and T. Jinmei. *RFC4862: IPv6 Stateless Address Autoconfiguration.* IETF, September 2007.

[79] Colin Perkins. *RTP: Audio and Video for the Internet.* Addison-Wesley Professional, 2003.

[80] D. Singer and H. Desineni. *RFC 5285: A General Mechanism for RTP Header Extensions.* IETF, July 2008.

[81] Hans Weibel. *IEEE 1588 Standard for a Precision Clock Synchronization Protocol.* Zurich University of Applied Sciences, November 2012.

[82] *Precision Time Protocol: White Paper.* EndRun technologies, July 2011.

[83] Kendall Correll, Nick Barendt, and Michael Branicky. Design considerations for software only implementations of the IEEE 1588 precision time protocol. In *Conference on IEEE 1588 Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems*, Zurich, Switzerland, 2005. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.67.4565&rep=rep1&type=pdf.

[84] Hans Weibel and Stefan Heinzmann. Media clock synchronization based on ptp. In *Audio Engineering Society Conference: 44th International Conference: Audio Networking.* ALC NetworX GmbH, Munich, Germany; Zurich University of Applied Sciences, Winterthur, Switzerland, Nov 2011. URL http://www.aes.org/e-lib/browse.cfm?elib=16146.

[85] M. Handley, V. Jacobson, and C. Perkins. *SDP: Session Description Protocol (RFC 4566).* IETF, July 2006.

[86] A. Williams, K. Gross, R. van Brandenburg, and H. Stokking. *RTP Clock Source Signalling (draft 11).* IETF, March 2014. URL http://tools.ietf.org/html/draft-ietf-avtcore-clksrc-11.

[87] Shamima Kabir, Mohamad Khazani Abdullah, Sabira Khatun, and Mohamad Adzir Mahdi. Buffered csma/cd based ethernet optical lan avoiding packet loss. *Suranaree Journal of Science and Technology*, 2007.

[88] James F. Kurose and Keith W. Ross. *Computer Networking: A Top-Down Approach*. Pearson Education, May 2012.

[89] C. Perkins, O. Hodson, and V. Hardman. A survey of packet loss recovery techniques for streaming audio. *Network, IEEE*, 12(5):40–48, Sept 1998. ISSN 0890-8044. doi: 10.1109/65.730750.

[90] Gerhard Hasslinger and Oliver Hohlfeld. The gilbert-elliott model for packet loss in real time services on the internet. In *Measuring, Modelling and Evaluation of Computer and Communication Systems (MMB), 2008 14th GI/ITG Conference -*, pages 1–15, March 2008.

[91] Rishi Sinha, Christos Papadopoulos, and Chris Kyriakakis. Loss concealment for multi-channel streaming audio. In *Proceedings of the 13th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, NOSSDAV '03, pages 100–109, New York, NY, USA, 2003. ACM. ISBN 1-58113-694-3. doi: 10.1145/776322.776339. URL http://doi.acm.org/10.1145/776322.776339.

[92] H. Sanneck, A. Stenger, K. Ben Younes, and B. Girod. A new technique for audio packet loss concealment. In *Global Telecommunications Conference, 1996. GLOBECOM '96. 'Communications: The Key to Global Prosperity*, pages 48–52, Nov 1996. doi: 10.1109/GLOCOM.1996.586117.

[93] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin. *Resource ReSerVation Protocol (RSVP)*. IETF, September 1997.

[94] M. Cotton, L. Vegoda, and D. Meyer. *IANA Guidelines for IPv4 Multicast Address Assignments*. IETF, March 2010.

[95] Nika Aldrich. *Dither Explained: An explanation and proof of the benefit of dither for the audio engineer*, April 2002.

[96] JM. Valin, K. Vos, and T. Terriberry. *Definition of the Opus Audio Codec*. IETF, September 2012.

[97] Alexey Lukin. *Sonically Optimized Noise Shaping Techniques*, 2 edition, 2002.

[98] J. Valin and K. Vos. *Requirements for an Internet Audio Codec*. IETF, August 2011.

[99] Jean François Frigon and Vladislav Teplitsky. *Implementation of Linear Predictive Coding (LPC) of Speech*, 2000.

[100] Ye Wang and Miikka Vilermo. The modified discrete cosine transform: Its implications for audio coding and error concealment. In *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, Jun 2002. URL http://www.aes.org/e-lib/browse.cfm?elib=11125.

[101] K. Vos, S. Jensen, and K. Soerensen. *SILK Speech Codec*. Skype Technologies S.A., September 2010.

[102] Jean-Marc Valin, Gregory Maxwell, and Timothy B. Terriberry. *CELT: A Low-latency, High-quality Audio Codec*. The Xiph.Org Foundation, 2011.

[103] Daniel Steinberg and Stuart Cheshire. *Zero Configuration Networking: The Definitive Guide*. O'Reilly Media, Inc., 1st edition, 2005. ISBN 0596101007.

[104] S. Cheshire, B. Aboba, and E. Guttman. *Dynamic Configuration of IPv4 Link-Local Addresses*. IETF, May 2005.

[105] S. Cheshire and M. Krochmal. *Multicast DNS*. Apple Inc., February 2013.

[106] S. Cheshire and M. Krochmal. *RFC 6763: DNS-Based Service Discovery*. Apple Inc., February 2013.

[107] OSEDUM P. IGUMBOR. A proxy approach to protocol interoperability within digital audio networks. Master's thesis, Rhodes University, September 2009.

[108] *ADVANCED AUDIO DISTRIBUTION PROFILE SPECIFICATION*. Bluetooth Special Interest Group, July 2012.

[109] *802.11ac In-Depth (WhitePaper)*. Aruba Networks, 2014.

[110] Gordon Kelly. *802.11ac vs 802.11n - What's the difference?* Trusted Reviews, June 2013.

[111] Seppo Nikkilä. Introducing wireless organic digital audio: A multichannel streaming audio network based on ieee 802.11 standards. In *Audio Engineering Society Conference: 44th International Conference: Audio Networking*. ANT - Advanded Network Technologies Oy, Helsinki, Finland, 2011.

[112] *Taxicab Geometry*. Wikipedia, 2014. URL http://en.wikipedia.org/wiki/Taxicab_geometry.

[113] J. MacQueen. Some methods for classification and analysis of multivariate observations, 1967. URL http://projecteuclid.org/euclid.bsmsp/1200512992.

[114] *K-means clustering*. Wikipedia, 2014. URL http://en.wikipedia.org/wiki/K-means_clustering.

[115] Dermot Campbell, Edward Jones, and Martin Glavin. Audio quality assessment techniques: A review, and recent developments. *Signal Processing*, 89(8):1489 – 1500, March 2009. ISSN 0165-1684. doi: 10.1016/j.sigpro.2009.02.015.

[116] NHK Features Report. Objective perceptual audio quality measurement methods. *Broadcast Technology No 35*, 2008-2009. URL http://www.nhk.or.jp/strl/publica/bt/en/fe0035-2.pdf.

[117] Kalpana Seshadrinathan and AlanConrad Bovik. Automatic prediction of perceptual quality of multimedia signals: a survey. *Multimedia Tools and Applications*, 51(1):163–186, 2011. ISSN 1380-7501. doi: 10.1007/s11042-010-0625-9. URL http://dx.doi.org/10.1007/s11042-010-0625-9.

[118] Junyong You, Ulrich Reiter, Miska M. Hannuksela, Moncef Gabbouj, and Andrew Perkis. Perceptual-based quality assessment for audio,Äöàìvisual services: A survey. *Signal Processing: Image Communication*, 25(7):482 – 501, 2010. ISSN 0923-5965. doi: http://dx.doi.org/10.1016/j.image.2010.02.002. Special Issue on Image and Video Quality Assessment.

[119] R Conetta. *Towards the automatic assessment of spatial quality in the reproduced sound environment*. PhD thesis, University of Surrey, 2011. URL http://epubs.surrey.ac.uk/39628/. Copyright 2011 The Author. The author confirms that he has permission for any third party copyrighted material used in the thesis.

[120] ITU-R. General methods for the subjective assessment of sound quality. ITU Tech Report, 12 2003.

[121] ITU-R. Method for the subjective assessment of intermediate quality levels of coding systems. ITU Tech Report, 01 2003.

[122] ITU-R. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. ITU Tech Report, 06 2001.

[123] Joe Palca and Flora Lichtman. *Annoying: The Science of What Bugs Us*. Wiley, 2011.

[124] Tomasz Letowski. Sound quality assessment: Concepts and criteria. In *Audio Engineering Society Convention 87*, Oct 1989. URL http://www.aes.org/e-lib/browse.cfm?elib=5869.

[125] Jan Berg and Francis Rumsey. Systematic evaluation of perceived spatial quality. University of Surrey, 2003.

[126] Jan BERG. Evaluation of perceived spatial audio quality. *SYSTEMICS, CYBERNETICS AND INFORMATICS*, 4(2), 2006.

[127] S. Le Bagousse, M. Paquier, and C. Colomes. Assessment of spatial audio quality based on sound attributes. In *Acoustics*, 2012.

[128] Abderrahmane SMIMITE, Azeddine BEGHDADI, and Ken CHEN. Vers une nouvelle approche de l'évaluation de la qualité audio multicanal. In *Compression et Représentation des Signaux Audiovisuels*, Novembre 2013.

[129] Tomasz R. Letowski and Szymon T. Letowski. Auditory spatial perception: Auditory localization. Technical report, Army Research Lab, May 2012.

[130] S/lawomir Zielinski, Francis Rumsey, and Søren Bech. On some biases encountered in modern audio quality listening tests-a review. *J. Audio Eng. Soc*, 56(6):427–451, 2008.

[131] Slawomir Zielinski, Philip Hardisty, Christopher Hummersone, and Francis Rumsey. Potential biases in mushra listening tests. In *Audio Engineering Society Convention 123*, Oct 2007. URL http://www.aes.org/e-lib/browse.cfm?elib=14237.

[132] Slawomir K. Zielinski, Francis Rumsey, Søren Bech, and Rafael Kassier. Quality adviser: A multichannel audio quality expert system. In *Audio Engineering Society Convention 116*, May 2004. URL http://www.aes.org/e-lib/browse.cfm?elib=12626.

[133] ITU-R. Method for objective measurements of perceived audio quality. Technical report, ITU-R Standards, Jun 2002.

[134] R. Huber and B. Kollmeier. Pemo-q-a new method for objective audio quality assessment using a model of auditory perception. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(6):1902–1911, 2006. ISSN 1558-7916. doi: 10.1109/TASL.2006.883259.

[135] C.D. Creusere, K.D. Kallakuri, and R. Vanam. An objective metric of human subjective audio quality optimized for a wide range of audio fidelities. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1):129–136, Jan 2008. ISSN 1558-7916. doi: 10.1109/TASL.2007.907571.

[136] Inyong Choi, Barbara G. Shinn-Cunningham, Sang Bae Chon, and Koeng-Mo Sung. Objective measurement of perceived auditory quality in multichannel audio compression coding systems. *J. Audio Eng. Soc*, 56(1/2):3–17, 2008.

[137] R.F. Lyon, A.G. Katsiamis, and E.M. Drakakis. History and future of auditory filter models. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 3809–3812, May 2010. doi: 10.1109/ISCAS.2010.5537724.

[138] Christopher R. Cave. Perceptual modelling for low-rate audio coding. Master's thesis, McGill University, Jun 2002.

[139] R. Vanam and C.D. Creusere. Evaluating low bitrate scalable audio quality using advanced version of peaq and energy equalization approach. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 3, pages iii/189–iii/192 Vol. 3, 2005. doi: 10.1109/ICASSP.2005.1415678.

[140] Martin Dewhirst, Philip Jackson, Francis Rumsey, and Slawomir K. Zielinski. Objective assessement of spatial localisation attributes of surroung-sound reproduction systems. In *Audio Engineering Society Convention 118*, May 2005. URL http://www.aes.org/e-lib/browse.cfm?elib=13157.

[141] Jeong-Hun Seo, Sang Bae Chon, Keong-Mo Sung, and Inyong Choi. Perceptual objective quality evaluation method for high quality multichannel audio codecs. *J. Audio Eng. Soc*, 61(7/8):535–545, 2013.

[142] Francis Rumsey, Slawomir Zielinski, Philip Jackson, Martin Dewhirst, Robert Conetta, Sunish George, Søren Bech, and David Meares. Qestral (part 1): Quality evaluation of spatial transmission and reproduction using an artificial listener. In *Audio Engineering Society Convention 125*, Oct 2008. URL http://www.aes.org/e-lib/browse.cfm?elib=14746.

[143] Philip Jackson, Martin Dewhirst, Robert Conetta, Slawomir Zielinski, Francis Rumsey, David Meares, Søren Bech, and Sunish George. Qestral (part 3): System and metrics for spatial quality prediction. In *Audio Engineering Society Convention 125*, Oct 2008. URL http://www.aes.org/e-lib/browse.cfm?elib=14748.

[144] Abderrahmane Smimite, Azeddine Beghdadi, and Ken Chen. Investigating "the experience effect" in audio quality assessment. In *Telecommunications Forum (TELFOR), 2013 21st*, pages 769–772, 2013. doi: 10.1109/TELFOR.2013.6716343.

[145] Bruno Cardenas, Andrew Schmitt, and Mandi Vance. A study of auditory source width and listener envelopment. Technical report, Rensselaer Polytechnic Institute, 2012.

[146] Leo L. Beranek. *Concert Halls and Opera Houses: Music, Acoustics, and Architecture*. Springer-Verlag, New York, USA, 2nd edition, 2004.

[147] John MacCormick. *How does the Kinect work?* Dickinson College, Carlisle, Pennsylvania, 2012.

[148] MSDN Team. Microsoft developer network: Skeletal tracking.

[149] Richard S. Snell. *Clinical Anatomy by Systems.* Lippincott Williams and Wilkins, 2007.

[150] (Several). Ari hrtf database. URL http://www.kfs.oeaw.ac.at/.

[151] Hermann Ebbinghaus. *Memory; a contribution to experimental psychology.* New York city, Teachers college, Columbia university, 1913.

[152] K. Seshadrinathan and Alan C. Bovik. Temporal hysteresis model of time varying subjective video quality. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 1153–1156, 2011. doi: 10.1109/ICASSP.2011.5946613.

[153] Jack Mostow, Kai-Min Chang, and Jessica Nelson. Toward exploiting eeg input in a reading tutor. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, AIED'11, pages 230–237, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-21868-2. URL http://dl.acm.org/citation.cfm?id=2026506.2026539.

[154] Bao Hong Tan. Using a low-cost eeg sensor to detect mental states. Master's thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, August 2012.

[155] Fernando Lopes da Silva. *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields.* Lippincott Williams & Wilkins, 5th edition, 2012.

[156] Elif Kirmizi-Alsan, Zubeyir Bayraktaroglu, Hakan Gurvit, Yasemin H. Keskin, Murat Emre, and Tamer Demiralp. Comparative analysis of event-related potentials during go/nogo and cpt: Decomposition of electrophysiological markers of response inhibition and sustained attention. *Brain Research*, 1104(1):114 – 128, 2006. ISSN 0006-8993. doi: http://dx.doi.org/10.1016/j.brainres.2006.03.010. URL http://www.sciencedirect.com/science/article/pii/S0006899306007244.

[157] Fernando Lopes da Silva. Functional localization of brain sources using eeg and/or meg data: volume conductor and source models. *Magnetic resonance imaging*, 22 (10):1533–1538, 2004. ISSN 0730-725X.

[158] *NeuroSky: eSense Meters.* NeuroSky, May 2014. URL http://developer.neurosky.com/docs/doku.php?id=esenses_tm.

[159] Genaro Rebolledo-Mendez, Ian Dunwell, ErikaA. Martínez-Mirón, MaríaDolores Vargas-Cerdán, Sara Freitas, Fotis Liarokapis, and AlmaR García-Gaona. Assessing neurosky's usability to detect attention levels in an assessment exercise. In JulieA. Jacko, editor, *Human-Computer Interaction. New Trends*, volume 5610, pages 149–158. Springer Berlin Heidelberg, 2009.

[160] ITU-R. Bs.775 : Multichannel stereophonic sound system with and without accompanying picture. ITU Tech Report, March 2013.

[161] Bill Gardner and Keith Martin. Hrtf measurements of a kemar dummy head microphone. MIT Media Lab, May 1994.

[162] Tianshu Qu, Zheng Xiao, Mei Gong, Ying Huang, Xiaodong Li, and Xihong Wu. Distance dependent head-related transfer function database of kemar. In *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on*, pages 466–470, 2008. doi: 10.1109/ICALIP.2008.4590089.

[163] Hagen Wierstorf, Matthias Geier, Alexander Raake, and Sascha Spors. A free database of head-related impulse response measurements in the horizontal plane with multiple distances. In *130th Convention of the Audio Engineering Society*, May 2011.

[164] Thilo Thiede, William C. Treurniet, Roland Bitto, Christian Schmidmer, Thomas Sporer, John G. Beerends, and Catherine Colomes. Peaq - the itu standard for objective measurement of perceived audio quality. *J. Audio Eng. Soc*, 48(1/2): 3–29, 2000. URL http://www.aes.org/e-lib/browse.cfm?elib=12078.

[165] Eberhard Zwicker and Hugo Fastl. *Psychoacoustics: Facts and Models*. Springer, June 2007.

[166] R Nave. *HyperPhysics: Sound*. Georgia State University, 2013.

[167] ITU-R. Recommendation itu-r bs.1770-3: Algorithms to measure audio programme loudness and true-peak audio level, 08 2012.

[168] Wake Forest University. Sensation and perception. PSY 329: Lecture 10, 2012.

[169] B. Shinn-Cunningham and K. Kawakyu. Neural representation of source direction in reverberant space. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 79–82, Oct 2003. doi: 10.1109/ASPAA.2003.1285824.

[170] Mathias Dietz, Stephan D. Ewert, and Volker Hohmann. Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication*, 53(5):592 – 605, 2011. ISSN 0167-6393. doi: http://dx.doi.org/10.1016/j.specom.2010.05.006. Perceptual and Statistical Audition.

[171] Andrew Brughera, Larisa Dunai, and William M. Hartmann. Human interaural time difference thresholds for sine tones: The high-frequency limit. *The Journal of the Acoustical Society of America*, 133(5):2839–2855, 2013. doi: http://dx.doi.org/10.1121/1.4795778.

[172] Rick P. Thomas Michael R. Dougherty. *Encyclopedia of Medical Decision Making: Minerva-DM*, pages 768–770. SAGE Publications, Inc., 2009.

[173] Abderrahmane Smimite, Azeddine Beghdadi, Ouadie Jafjaf, and Ken Chen. A new approach for spatial audio quality assessment. In *2014 International Conference on Telecommunications and Multimedia (TEMU2014)*, Heraklion, Greece, July 2014.

[174] Myung-Suk Song, Cha Zhang, D. Florencio, and Hong-Goo Kang. An interactive 3-d audio system with loudspeakers. *Multimedia, IEEE Transactions on*, 13(5):844–855, Oct 2011. ISSN 1520-9210. doi: 10.1109/TMM.2011.2162581.

[175] Sebastian Merchel and Stephan Groth. Analysis and implementation of a stereophonic play back system for adjusting the "sweet spot" to the listener's position. In *Audio Engineering Society Convention 126*, May 2009. URL http://www.aes.org/e-lib/browse.cfm?elib=14922.

[176] SEBASTIAN MERCHEL and STEPHAN GROTH. Evaluation of a new stereophonic reproduction method with moving "sweet spot" using a binaural localization model. In *Proceedings of the ISAAR*, Copenhagen, Denmark, 2009.

[177] Benjamin B. Bauer. Broadening the area of stereophonic perception. *Journal of Audio Engineering Society*, 8(2):91–94, 1960. URL http://www.aes.org/e-lib/browse.cfm?elib=525.

[178] Josep A Rodenas, Ronald M Aarts, and A J E M Janssen. Derivation of an optimal directivity pattern for sweet spot widening in stereo sound reproduction. *J Acoust Soc Am*, 113(1):267–278, Jan 2003. ISSN 0001-4966 (Print); 0001-4966 (Linking).

[179] T. Takeuchi and P.A. Nelson. Optimal source distribution, 2013. URL http://www.southampton.ac.uk/engineering/research/groups/fluid_dynamics/electroacoustics/optimal_source_distribution.page.

[180] Shigeaki Aoki, Hiroyuki Miyata, and Kiyoshi Sugiyama. Stereo reproduction with good localization over a wide listening area. *J. Audio Eng. Soc*, 38(6):433–439, 1990. URL http://www.aes.org/e-lib/browse.cfm?elib=6026.

[181] Trinnov room optimization concept, 2014. URL http://www.trinnov.com/technologies/loudspeaker-room-optimization/.

[182] Joachim Olsen Wille. Performance of a multichannel audio correction system outside the sweetspot. Master's thesis, Norwegian University of Science and Technology, June 2008.

[183] Chris Kyriakakis, Tomlinson Holman, Jong-Soong Lim, Hai Hong, and Hartmut Neven. Signal processing, acoustics, and psychoacoustics for high quality desktop audio. *Journal of Visual Communication and Image Representation*, 9(1):51 – 61, 1998. ISSN 1047-3203. doi: http://dx.doi.org/10.1006/jvci.1998.0379. URL http://www.sciencedirect.com/science/article/pii/S1047320398903790.

[184] Sunmin Kim, Donggeon Kong, and Seongcheol Jang. Adaptive virtual surround sound rendering system for an arbitrary listening position. *Journal of Audio Engineering Society*, 56(4):243–254, April 2008.

[185] Stanley P. Lipshitz. Stereo microphone techniques: Are the purists wrong? In *Audio Engineering Society Convention 78*, May 1985. URL http://www.aes.org/e-lib/browse.cfm?elib=11494.

[186] Jonas Braasch. Modelling of binaural hearing. In Jens Blauert, editor, *Communication Acoustics*, pages 75–108. Springer Berlin Heidelberg, 2005.

[187] S. Groth. Untersuchung eines stereo-systems mit signalanpassung an die hörposition, 2008.

[188] Jong soong Lim and Chris Kyriakakis. Multirate adaptive filtering for immersive audio. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, volume 5, pages 3357–3360 vol.5, 2001. doi: 10.1109/ICASSP.2001.940378.

[189] Qiang Wang, Weiwei Zhang, Xiaoou Tang, and Heung-Yeung Shum. Real-time bayesian 3-d pose tracking. *Circuits and Systems for Video Technology, IEEE Transactions on*, 16(12):1533–1541, Dec 2006. ISSN 1051-8215. doi: 10.1109/TCSVT.2006.885727.

[190] Ki-Seung Lee and Seok pil Lee. A real-time audio system for adjusting the sweet spot to the listener's position. *Consumer Electronics, IEEE Transactions on*, 56 (2):835–843, May 2010. ISSN 0098-3063. doi: 10.1109/TCE.2010.5506009.

[191] Ramin Anushiravani and Douglas L. Jones. 3d audio playback through two loudspeakers, January 2014.

[192] Edgar Y. Choueiri. Optimal crosstalk cancellation for binaural audio with two loudspeakers. Online, 2008. URL http://www.princeton.edu/3D3A/Publications/BACCHPaperV4d.pdf.

[193] *Thermal imaging guidebook for indusTrial applicaTions*. FLIR Systems, 2011.

[194] Rinerhart Holt and Winston. Integrating health: Skin temperature. Online, 2008. URL http://go.hrw.com/resources/go_sc/ssp/HK1IE066.PDF.

[195] R B BARNES. Thermography of the human body. *Science*, 140(3569):870–877, May 1963. ISSN 0036-8075 (Print); 0036-8075 (Linking).

[196] Healthy heating: Skin temperatures. Online, 2012. URL http://www.healthyheating.com/Definitions/facts_about_skin.htm.

[197] Luc Vincent. Grayscale area openings and closings, their efficient implementation and applications. In *Proc. EURASIP Workshop on Mathematical Morphology and its Applications to Signal Processing*, pages 22–27, Barcelona, Spain, May 1993.

[198] John C. Bennett, Keith Barker, and Frederick O. Edeko. A new approach to the assessment of stereophonic sound system performance. *Journal of Audio Engineering Society*, 33(5):314–321, 1985. URL http://www.aes.org/e-lib/browse.cfm?elib=4449.

[199] Ville Pulkki and Tapio Lokki. Creating auditory displays with multiple loudspeakers using vbap: A case study with diva project. In *Proceedings of the 1998 International Conference on Auditory Display*, ICAD'98, pages 23–23, Swinton, UK, 1998. British Computer Society. URL http://dl.acm.org/citation.cfm?id=2227605.2227628.