

# POLITECNICO DI TORINO

Master's Degree in Ing. Del Cinema e dei Mezzi di  
Comunicazione



Master's Degree Thesis

## Spatial Comparison of Full Sphere Panning Methods

Supervisors

Prof. Marco Carlo MASOERO

Prof. Enzo DE SENA

Candidate

Enrico FODDE

s280135

a.a. 2021/2022



"...Totu torrat e at a torrare cando nois no amus a èssere pius bios,  
Ma b'at a èssere s'arvure chi amus piantadu.

Sa cosa giusta est de lassare bonos ammentos,  
comente unu chi at fatu carchi cosa pro mezorare su mundu."

Salvatore Cambosu

A nonna Marianna, nonnu Chìrigu e tia Tomasina  
A tiu Chìrigu e tia Giuanna





# Table of Contents

<b>List of Figures</b>	VII
<b>1 Introduction</b>	1
1.1 Aims of the Thesis . . . . .	1
1.2 Organization of the Thesis . . . . .	3
<b>2 Stereophony Methods for Horizontal and Vertical Plane - Background</b>	4
2.1 Introduction . . . . .	4
2.2 Human Auditory System . . . . .	5
2.2.1 Auditory System Anatomy . . . . .	5
2.2.2 The Theories of Hearing . . . . .	8
2.3 Perception on the Horizontal and Vertical Plane . . . . .	10
2.3.1 Perception on the Horizontal Plane . . . . .	10
2.3.2 Perception on the Vertical Plane . . . . .	16
2.3.3 Perception of Distance . . . . .	17
2.3.4 The Parameters of Perception- Position . . . . .	18
2.3.5 Other Parameters of Perception . . . . .	18
2.4 Stereophony . . . . .	20
2.4.1 Two-Channels Stereophonic Setup . . . . .	20
2.4.2 Microphones and Characteristics . . . . .	24
2.4.3 Microphone Arrays for 2-Channel Stereophony . . . . .	26
2.5 From 2-Channels to Multichannel Stereophony . . . . .	33
2.5.1 5.1 Multichannel Configuration . . . . .	34
2.6 Perceptual Sound Field Reconstruction- PSR . . . . .	36
2.6.1 How Many Loudspeakers Render a Single Sound Source? . .	36
2.6.2 Design of the Microphone Directivity . . . . .	39
2.6.3 MAX-MSP Implementation . . . . .	42
2.7 Panning Techniques for Vertical Plane . . . . .	43
2.8 VBAP- Vector Based Amplitude Panning . . . . .	44
2.8.1 MAX MSP Implementation of VBAP . . . . .	48

2.9	Other Options . . . . .	49
2.9.1	Ambisonics . . . . .	49
2.9.2	Wave-Field Synthesis - WFS . . . . .	53
2.10	Microphone Arrays for Multichannel . . . . .	55
2.10.1	Physical Reconstruction Methods: Soundfield and Eigenmike . . . . .	55
2.10.2	Tree-Structure Arrays . . . . .	57
2.10.3	Arrays for a Higher Number of Channels . . . . .	59
2.11	Chapter Conclusions . . . . .	59
<b>3</b>	<b>Hybrid Time Amplitude Approach</b>	<b>60</b>
3.1	Why Use Time Differences? . . . . .	60
3.2	Explanation of the Hybrid Approach . . . . .	61
3.2.1	Introduction . . . . .	61
3.2.2	Horizontal Panning . . . . .	61
3.3	Extension to the Vertical Panning . . . . .	62
3.3.1	First Faulty Approach: the $\alpha$ Parameter . . . . .	62
3.3.2	Panning for 3 Loudspeakers . . . . .	63
3.3.3	Final Values of the Gains $g_n(\theta, \varphi)$ . . . . .	67
3.4	ORTF3D Microphone Array . . . . .	68
3.5	Chapter Conclusions . . . . .	69
<b>4</b>	<b>Experiment Methodology and Setup</b>	<b>70</b>
4.1	Introduction . . . . .	70
4.2	Experiment Setup . . . . .	70
4.2.1	Room . . . . .	70
4.2.2	Loudspeakers . . . . .	71
4.2.3	Curtain . . . . .	74
4.2.4	Listening Positions . . . . .	74
4.2.5	Interface . . . . .	76
4.3	Methodology and Stimuli . . . . .	76
4.3.1	Methodology . . . . .	76
4.3.2	Stimuli . . . . .	78
4.3.3	Experiment Execution . . . . .	78
4.3.4	Subjects . . . . .	79
4.4	Chapter Conclusions . . . . .	79
<b>5</b>	<b>Experimental Results and Discussion</b>	<b>80</b>
5.1	Locatedness . . . . .	81
5.1.1	Center + Off-Center . . . . .	81
5.1.2	Center Position . . . . .	81
5.1.3	Off-Center Position . . . . .	82

5.2	Localization . . . . .	84
5.2.1	Conversion of the Localization Data . . . . .	84
5.2.2	Angular Error . . . . .	85
5.3	Chapter Conclusions . . . . .	91
<b>6</b>	<b>Conclusions</b>	<b>93</b>
<b>A</b>	<b>Appendix</b>	<b>96</b>
A.1	Consent Form . . . . .	97
A.2	Listening Test Instruction . . . . .	98
A.3	Experiment Guide . . . . .	100
	<b>Bibliography</b>	<b>102</b>

# List of Figures

2.1	Representation of the Horizontal and Vertical Angles of Azimuth and Elevation, and of the Distance from Listener to the Sound Source. Image courtesy of [6] . . . . .	5
2.2	Human Auditory System Anatomy. Image courtesy of Encyclopædia Britannica, Inc. . . . .	6
2.3	Structure of the Cochlea. Image courtesy of [9] . . . . .	8
2.4	Resonance Zones of the Cochlea in Place Theory of Hearing. Image courtesy of [10] . . . . .	8
2.5	Overlapping Zones of Resonance. Image courtesy of [10] . . . . .	9
2.6	Sound-Wave Path for the Human Head. Image courtesy of [17] . . .	13
2.7	Accuracy of Horizontal Localisation for Frequencies Ranges. Image courtesy of [6] . . . . .	16
2.8	Accuracy of Vertical Localisation for Frequencies Groups. Image courtesy of [6] . . . . .	17
2.9	ISO 226:2003 Standard Curves. Image courtesy of [29] . . . . .	19
2.10	Standard 2-Channel Stereophonic Setup . . . . .	21
2.11	Williams Time-intensity Psychoacoustic Curves. Image courtesy of [33] . . . . .	23
2.12	Contour Plot of Localisation Uncertainty for a Centered Position. Picture courtesy of [33] . . . . .	23
2.13	Frequency Response of a Shure Microphone. Image courtesy of [34] . . . . .	25
2.14	Polar Patterns of a Microphone. Image courtesy of [35] . . . . .	26
2.15	Blumlein Stereophonic Configuration. Image courtesy of [36] . . . . .	27
2.16	X-Y Configuration. Image courtesy of [37] . . . . .	28
2.17	Mid-Side Configuration. Image courtesy of [38] . . . . .	28
2.18	ORTF Configuration. Image courtesy of [37] . . . . .	29
2.19	NOS Configuration. Image courtesy of [37] . . . . .	29
2.20	A-B Configuration. Image courtesy of [37] . . . . .	30
2.21	A-B Configuration Curves. Image courtesy of [37] . . . . .	31

2.22	Dummy Head System Neumann KU100. Image courtesy of Georg Neumann GmbH, Berlin . . . . .	32
2.23	Decca Tree Configuration. Image courtesy of [42] . . . . .	34
2.24	5.1 Surround Configuration. Picture courtesy of [44] . . . . .	35
2.25	PSR Configuration for 5.1. Image courtesy of [27] . . . . .	36
2.26	Recording (a) and Reproduction (b) System for a Multichannel Array. Image courtesy of [2] . . . . .	38
2.27	PSR Scheme . . . . .	41
2.28	MAX MSP Patch for PSR Implementation . . . . .	42
2.29	Psychoacoustic Curves for Horizontal and Vertical Panning. Image courtesy of [47] . . . . .	44
2.30	Active Triangle Concept on VBAP. Image courtesy of [30] . . . . .	45
2.31	Example of Implementation of VBAP. Image courtesy of [30] . . . . .	46
2.32	MAX MSP Implementation of VBAP . . . . .	49
2.33	Ambisonics X Y Z W Signals. Image courtesy of [55] . . . . .	50
2.34	B-Format Encoding System. Image courtesy of [52] . . . . .	50
2.35	Ambisonics Directivity Patterns in Function of the Order. Image courtesy of [56] . . . . .	51
2.36	Ambisonics Reproduction Configuration. Image courtesy of [43] . . . . .	52
2.37	WFS Loudspeakers Disposition. Image courtesy of [59] . . . . .	53
2.38	WFS Implementation at the University of Technology of Berlin. Image courtesy of [60] . . . . .	54
2.39	SoundField Capsules. Image courtesy of [61] . . . . .	56
2.40	Eigenmike em32 Microphone. Image courtesy of [64] . . . . .	56
2.41	OCT. Image courtesy of [55] . . . . .	57
2.42	INA-5. Image courtesy of [55] . . . . .	58
2.43	Multichannel Arrays Overview. Image courtesy of [27] . . . . .	58
3.1	ORTF 3D. Image courtesy of [65] . . . . .	68
3.2	Capsules Orientation on the Vertical Plane. Image courtesy of [65] . . . . .	68
4.1	TB7 Room at the University of Surrey. Image courtesy of [69] . . . . .	71
4.2	Frontal View of the Loudspeaker Array . . . . .	73
4.3	Lateral View of the Loudspeaker Array . . . . .	73
4.4	Aerial View of the Loudspeaker Array . . . . .	74
4.5	Aerial View of the Sitting Position . . . . .	75
4.6	Frontal Listening Position With and Without the Curtain . . . . .	75
4.7	Experiment Interface . . . . .	76
4.8	Stimuli Position Chart and Figure . . . . .	77
5.1	General Certainty for Different Methods in All the Stimuli Positions. Center Position . . . . .	81

5.2	Pairwise Comparison Between the Certainty for Different Methods in All the Stimuli Positions. Center Position . . . . .	82
5.3	General Certainty for Different Methods in All the Positions. Off Center Position . . . . .	82
5.4	Pairwise Comparison Between the Certainty for Different Methods in All the Stimuli Positions. Off-Center Position . . . . .	83
5.5	General Certainty for Different Methods in the Points on the Right in Respect to the Center of the Array, Further Away from the Listener. Off-Center Position . . . . .	83
5.6	Descriptive Statistics of Certainty for VBAP and Hybrid 15.5 for the Points in the Right in Respect to the Center of the Array, Further Away From the Listener. Off-Center Position . . . . .	84
5.7	Kruskal-Wallis Test for Horizontal and Vertical Angular Error Considering all the Methods. Center Position . . . . .	86
5.8	Comparison Between Vertical Angular Error Considering All the Methods. Center Position . . . . .	86
5.9	Elevation Error in Function of the Different Elevation Values. Center Position . . . . .	87
5.10	Horizontal Angular Error for the Different Methods. Center Position . . . . .	87
5.11	Vertical Error for Different Heights . . . . .	89
5.12	Horizontal Error for Different Heights . . . . .	89
5.13	Azimuth Localization Error. Off-Center Position . . . . .	89
5.14	Pairwise Comparison Between Methods for Horizontal Localisation. Off-Center Position . . . . .	90



# Chapter 1

## Introduction

### 1.1 Aims of the Thesis

Binaural stereophonic system has been the recording and reproduction standard since the Lord Rayleigh[1] theory and Blumlein implementation was adopted back in 1931.

Even if this format is still the most used one for home listening, both with loudspeakers and headphones, in the last few decades some improvements have been made in the Immersive Audio field, arriving at the multichannel system as the new state of the art for listening.

Giving a clear definition of what 3D audio is can result a tough task, since the term is sometimes abused, for example, to describe monodimensional methods realised with a circumference of loudspeakers, but which do not reproduce height sensation.

For this reason, it is possible to include in the immersive methods all the ones which allow the listener to feel surrounded by the sound, making him feel inside the real scene.

Possible examples might be a broadcast from the Royal Opera at home, having the impression of listening to it from the best seats at the opera house or watching a football match in a bar and feeling inside the stadium.

Almost surely, the multichannel system (intended as more than two channels provided) will not completely replace binaural stereophony, but, the research into these methods is significant due to the increase of use, not only for special reproduction purposes like Dolby Atmos or DTS:X in cinema, but also for domestic use. Dolby Atmos, which allows the use of up to 64 loudspeakers, is turning into the



standard for audio mixing, and some of the newest productions with this format has been recently released and distributed by Apple Music.

Not only cinema and music, but also VR experience has changed, due to the immersive experience in terms of audio and visual.

Since the debate about this research field is still open, another important reason to keep looking for alternative solutions concerns the overcoming of limitations of other state-of-the-art technologies by exploiting known features and limitations of how humans perceive sound sources and their relative position for the listener, in terms of localization of the sound source.

For this reason, the goal of the thesis work is to compare different multi-channel panning methods for rendering an audio source in a 3D space, analysing the pros and the cons of the methods in terms of localization and certainty.

The rendering of the auditory scene can be achieved mainly in two ways, with physical motivated methods, using signals prerecorded which are previously captured with particular microphone arrays, or with perceptual reconstruction methods, the purpose of which is to reproduce perceptual laws, feeding the loudspeakers with an original signal amplified and delayed to render a real sound source.

This thesis is concentrated on the second category, but some information about the first is provided for the reader, to give a comparison between alternatives methods.

The project, carried out at the Institute of Sound Recording of the University of Surrey (Guildford, UK) with the supervision of Prof. Enzo De Sena, ultimately proposes an alternative method for 3D panning of sound.

The method was then experimented and the performance is discussed later in this work.

## 1.2 Organization of the Thesis

- The **second chapter** consists of an overview about the background of the research topic. As an introduction, all the phenomena that control auditory perception along horizontal and vertical planes are illustrated, with the following explanation of how panning works in terms of Level and Time Difference between channels and the various techniques for reproducing these effects with microphone arrays.

Then, different panning approaches that use these techniques are shown for the horizontal and vertical dimensions, starting from the classic Stereo Configuration to 3D methods as VBAP (Vector Base Amplitude Panning, by Pulkki) or WFS (Wave-Field Synthesis).

- After this part describing the background, an original panning approach for full-sphere is presented in the **third chapter**. The method is called Hybrid because it combines Time Intensity Difference for the horizontal plane and Intensity Difference only for the vertical plane. The starting point is a previous project from Enzo De Sena (who is also one of the supervisors of the thesis) et al.[2] about panning on the horizontal plane, called PSR (Perceptual Soundfield Reconstruction), a method that has been extended to the vertical plane in this thesis project.
- After the theoretical design of the method, in the **fourth chapter** an experiment with mainly trained listener participants has been run to evaluate the Hybrid approach. The goal of the experiment was to compare the performance of an existing method (VBAP) with different implementations of the Hybrid approach, evaluating the localization accuracy and certainty of these. In fact, during the experiment, the listeners were exposed to different stimuli from different positions, and then they were asked to indicate the position of the sound events and to establish the degree of certainty of each answer.
- The experiment performances are investigated in the **chapter five**.
- At last, the conclusions of the work are presented in **chapter six**.

At the end of every chapter, some conclusions are presented, concerning the highlights of the most important elements presented useful for the reader.

## Chapter 2

# Stereophony Methods for Horizontal and Vertical Plane - Background

### 2.1 Introduction

The listening experience of a sound source is a subjective phenomena called **auditory perception** [3].

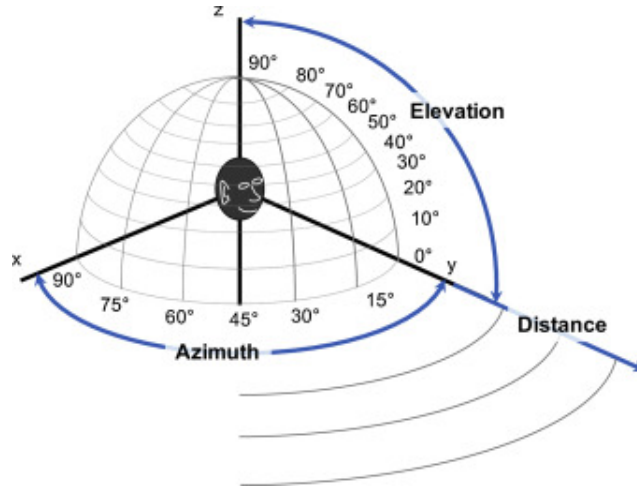
The perception moment has been defined in a more extended way by the German neurologist Hans Lungwitz as *the moment when the perceiver and the perceived encounters each other in a such a way that the perceived became conscious of what is perceived* [4].

The auditory perception is linked to the perception of the sound, defined as the mechanical vibration and waves of an elastic medium, in the frequency range of the human hearing (from 16 Hz to 20 kHz).

Auditory events and auditory perceptions, as defined by Blauert in [3], are distinct phenomena that are often, but not always, linked each other. Indeed, in the majority of the case the sound perception is due to a sound event, but there are sound perceptions not produced by sound events (for example hearing diseases such as tinnitus) or sound events not perceived by a subject (for example if is not loud enough). It is important to specify this difference to underline the subjectivity aspect of auditory perception, and of perception in general.

The problem regarding the determination of the position for a sound event takes the name of **localisation**. With the position of the listener fixed, it is possible to determine the position of a sound source by three main parameters expressed by spherical coordinates: **azimuth**, **elevation** and **radius**, Image 2.1. These are respectively the horizontal and vertical angle and the distance from the listener to the sound event. The azimuth goes from  $0^\circ$  to  $360^\circ$ , which are both the same median point, while the elevation takes values from  $-90^\circ$ , under the listener, to  $90^\circ$ , completely over the head of the listener[5].

The planes which identify spatial hearing are two, one vertical, which divides symmetrically left and right spaces, called **Median Plane**, and a horizontal one, which divides upper and lower planes, called **Saggital Plane**. Referring to the ears, the median plane is equidistant from both ears, while the other one is where the ears are located.



**Figure 2.1:** Representation of the Horizontal and Vertical Angles of Azimuth and Elevation, and of the Distance from Listener to the Sound Source. Image courtesy of [6]

## 2.2 Human Auditory System

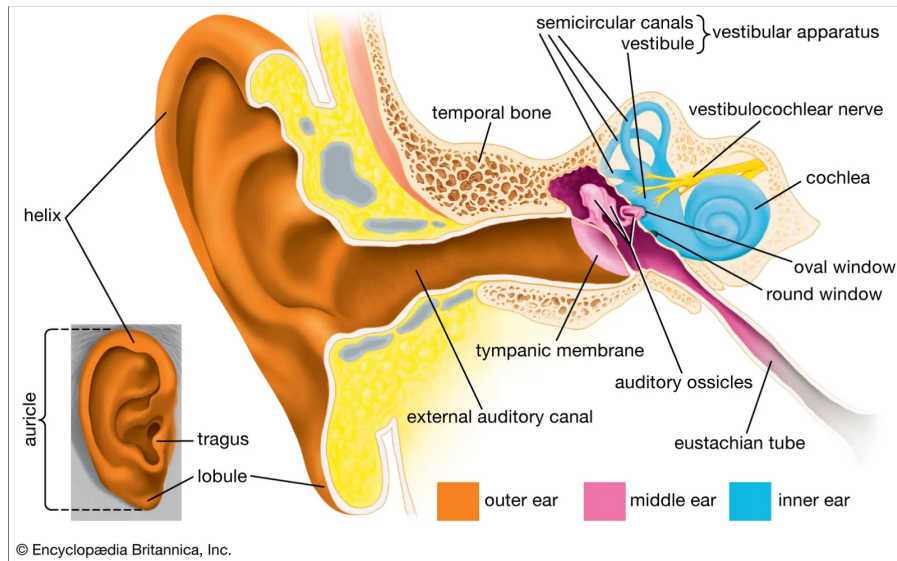
### 2.2.1 Auditory System Anatomy

The auditory system is the apparatus responsible for hearing, thanks to the process of auditory transduction, consisting in the transformation of external sound stimuli (compression and depression of the air that arrives to the ears) into electrochemical stimuli elaborated the brain [5]. The apparatus is also responsible for the equilibrium. This system, illustrated on the figure 2.2, is mainly composed of three parts:

**Outer Ear** (or auris externa, formed by the auricle, or pinna, and the external auditory canal), **Middle Ear** (or auris media, which includes the Eardrum, also called tympanic membrane, the auditory ossicles and the mastoid system) and **Inner Ear** (or auris interna, where the bony labyrinth is located) [7].

The pinna is cartilage in a shell form with the role of collecting auditory signals and sending them to the external auditory canal, positioned at the entry of the temporal bone, and finally to the eardrum. The human auricle can't be oriented, and this is the main difference with respect to some other mammals, that mainly use this part for the sound localisation of predators, and that is the reason why it is more evolved in them.

The external auditory canal is a 25 mm long duct, with an elliptic section of a maximum diameter of 6-8 mm [7]. The end of this canal coincides with the first element of the middle ear, the tympanic membrane. Together with the transmission of the signal, the other role of the auditory canal is to amplify the sound, similar to an organ pipe, before sending this to the tympanic membrane[8].



**Figure 2.2:** Human Auditory System Anatomy. Image courtesy of Encyclopædia Britannica, Inc.

The tympanic membrane, with a circular form and a diameter of 8-9 mm, can vibrate in presence of pressure due to a sound waves, and in this way can transmit the signal to the auditory ossicles (positioned in the tympanic cavity) thanks to the link between this and one of the bones, the annulus. The tympanic cavity needs to move back to its rest position after the vibration, and this is made possible because

the chamber conserves the external pressure thanks to the Eustachian tube, linking between this chamber and the pharynx, that balances the external pressure at every deglutition.

The vibration of the tympanic membrane and, in consequence, of the first ossicle, starts a chain movement that makes possible the movement of the other two linked ossicles, the incus (in the middle) and the stapes (at the other extreme). This last ossicle presses on the first part of the bony labyrinth in the inner ear, the cochlea. In addition to the transmission, the ossicles also amplify the vibration up to 20 times. Even if this chain movement seems really simple, this presents an issue about the transmission between two different fluids, the air (a gas) and the liquid inside the inner ear. In fact, low energy vibrations would not be transmitted in this way, but the presence of a chain system overcomes this problem of the impedance ratio between the two mediums [8]. Also, this system protects the cochlea from the high energy low frequency sound, which can damage it, thanks to the reflex of the stapedius [5].

The inner part of the ear starts with the aforementioned contact point between the stapes and the cochlea, in the inner part of the temporal bone, called the oval window.

The cochlea is only one part of the bony labyrinth, which, not only is responsible for the auditory transduction, but also for the equilibrium.

The bony labyrinth is made by the cochlea and by a central portion, called the vestibule.

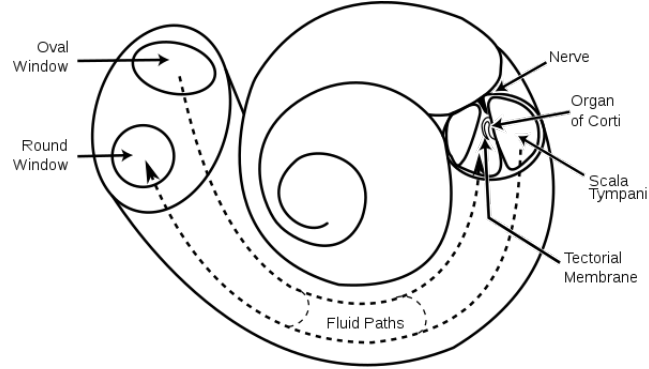
The cochlea is a spiral made up of three chamber canal with a width of 10 mm. The chambers are the **Scala Vestibuli** (where the oval window is located), the **Scala Timpani** (where the round window is located in the base) and the **Scala Media**, between these, which contains the Organ of Corti, a fundamental part of the hearing process.

The first two canals are full of a liquid called perilymph, while the third one contains endolymph, which contains a major concentration of  $K^+$  ions. The first two chambers are connected at a point called helicotrema.

When the stapes presses on the oval window, in a mechanism similar to a piston, the pressure generates a movement of the fluid inside the cochlea. The perilymph is incompressible as all the liquids are, so, the role of the liquid is only to transmit the vibration to the other extreme of the canal, the round window, which is linked to the middle ear.

Besides receiving the vibration, high frequencies are encoded in this region, while low frequencies are encoded near the oval window. In addition, the vibrations are transmitted also to the scala media and to the endolymph, where movement

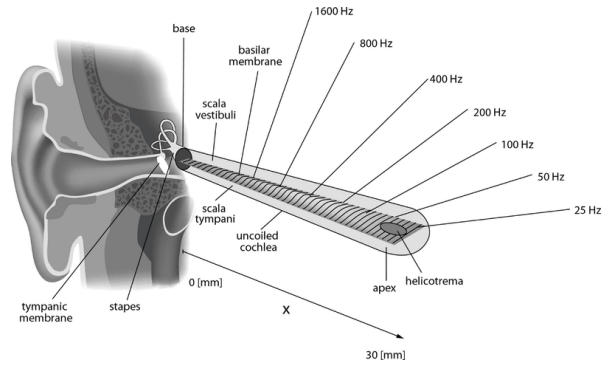
generates the stimulus of the receptors of the sound, consisting in approx 20.000 hair cells positioned in the Corti organ. Two types of hair cells are present, the inner and the outer cells, with the first type making up 95% of the total. The liquid moves the hair of cells and this movement generates an impulse to the Cochlear nerve that is sent to the brain, and specifically to the auditory cortex, for the reception of the sound.



**Figure 2.3:** Structure of the Cochlea. Image courtesy of [9]

### 2.2.2 The Theories of Hearing

Due to the complex mechanism of hearing, which is still a field of study for the researchers, two different theories about how we hear exist: **Place Theory of Hearing** and **Temporal Theory of Hearing** [10].



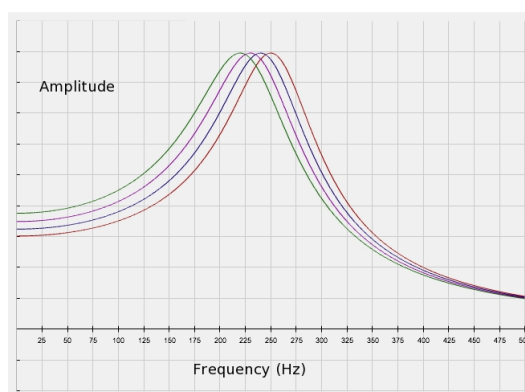
**Figure 2.4:** Resonance Zones of the Cochlea in Place Theory of Hearing. Image courtesy of [10]

## Place Theory of Hearing

As seen in the previous section, different parts of the cochlea resonate at different frequencies, closest to the stapes for high frequencies and furthest away for low frequencies. Since every group of hairs vibrates at a certain frequency, it can be assumed that every part of the cochlea corresponds to a certain frequency band of perception.

The cochlea can be imagined as divided into different resonance frequencies bands, the part furthest from the ossicles has a lower resonance frequency while the other has a higher resonance frequency. Every part sends its frequency-domain impulse to the brain and this gives humans the perception of a certain sound.

The resonance curves of this theory overlap and are very broad, with a high damping, resulting in a difficulty for the ear to trigger correctly frequencies near each others. This results in an overlap of the perceived frequencies. For example, in the image below, four overlapping curves for near frequencies are presented.



**Figure 2.5:** Overlapping Zones of Resonance. Image courtesy of [10]

In addition, a sharp resonance would mean a lower accuracy for the distinction of the sound duration, since it is harder to distinguish the changing of frequency. The human perception is able to recognize frequency changes of a tenth of a second, and this fact invalidates this explanation of the theory. These factors determine the main problem with the Place Theory of Hearing.

A correction to this theory can be the fact that certain nerves are inhibited when the nearest ones are excited in a certain moment. This is also true for other senses, such as touch or sight.



## Temporal Theory of Hearing

In the Temporal Theory of Hearing, also called Periodicity Theory, the firing timing of nerve impulses carries information about the perceived pitch. If a nerve is fired in a period  $T$ , the brain will interpret this as a sound frequency  $f=1/T$ . Considering a sin-wave with a frequency  $f= 500$  Hz, this has a period  $T=0.002$  s, so a vibration of the nerve at every  $T$  determines a perception of the sound of 500Hz. In addition, different sections of a complex sound waveform repeats periodically, and with different nerves stimulated each time.

Also this theory does not seem to be exactly demonstrable in this way, since the nerves do not seem to fire so often.

Instead of firing for every  $T$ , the nerve might fire initially at  $T$  and then at  $2T$ ,  $3T$  and so on, making it still possible to the brain to perceive the sound correctly, understanding the common divider of every firing frequency.

Another proposed possibility is that the nerves in the cochlea filter and combine signals somehow.

As said before, both theories are valid for some ideas and not valid for some others, and are, for this reason, still an open field of study. In a certain way, the combination of both can give the idea about how the human brain perceives sounds.

## 2.3 Perception on the Horizontal and Vertical Plane

### 2.3.1 Perception on the Horizontal Plane

The determining of the position of a sound source in the horizontal plane is a binaural process, that involves both the ears. The coherent single signals that arrives to each ear are perceived as a unique signal, called **Phantom Source**, by a process defined as **Summing localisation** [3].

In general, every soundwave emitted by a source, can arrive to the target in a direct or indirect way, through reflections on the obstacles near the listener.

The one coming from the direct source is the most important information for the sound localisation, while the other one, generated by reflection with the object in the same ambient, gives information about the nature of the ambience itself [5].

Horizontal localisation is controlled mainly by the different arrival time and the different arrival angle of the sound to the two ears [11]. The intensity of these two phenomena is dependent by the frequency of emission of the source.

## ITD - Interaural Time Difference

The different arrival time to the two ears determines a delay  $\tau$ , called **ITD (Interaural Time Difference)** that allows the auditory system to perceive a source as if it is coming from the same direction of the ear where the soundwave arrives first.

For example, for a source located in a more left position in respect to the center, the soundwave will arrive to the left ear at first, and then to the right ear with a certain delay, which depends on how decentred the sound is, and resulting in the perception of a sound coming more from the left. This mechanism is called **Precedence Effect** or **Law of the First Wave-front**, and prevents the perception of other sound source for 40 ms approximately, after the signal arrives at the first ear, if the other signal is not significantly louder than the first one. After this amount of time, the second signal is perceived as an echo [5]. Blauert, in [3], defines the echo threshold as 2 ms for the clicks and 40 ms for the speech.

The ITD, for an incident soundwave of azimuth  $\theta$ , situated at a distance  $a$ , with sound velocity  $c$ , can be expressed by the following equation designed by Woodworth (and which takes his name), which takes account of the curved path of the sound on a shaped head [12]:

$$ITD(\theta) = \frac{a}{c}(\sin(\theta) + \theta), \quad \text{with } 0 \leq \theta \leq \frac{\pi}{2} \quad (2.1)$$

Wanting to quantify the maximum delay, correspondent to an extreme left or right position, it may be considered the easiest case of a sinusoidal plane-wave that arrives with a sound speed of 343 m/s, with a maximum distance between the two ears of 23 cm [13]. The resulting delay, obtained by the division of the distance and the speed, is equal to 0.67 ms.

The Interaural Time Differences are perceived only for frequencies below 1500 Hz [6], because these frequencies are characterized by a wavelength smaller than the dimension of the head. In the same way, for very low frequencies, wavelengths which are too big determine phase differences correspondent to an undersized delay to be perceived by the listener. This affirmation is demonstrated below:

$$f_{max} = \frac{c}{\lambda_{min}} = \frac{343}{2.57 \cdot 0.087} = 1525 Hz \quad (2.2)$$

The divider  $\lambda_{min}$  is equal to the maximum distance which a wave has to cover to reach the opposite ear, if located at an extreme point, at  $90^\circ$ .

For really low frequencies and ones above 1500 Hz, phase ambiguity phenomena can occur only with ITD, with no directional distinction of the sound event [14], even if, for pure tones ITD cue remains relevant beyond 1500 Hz [6].

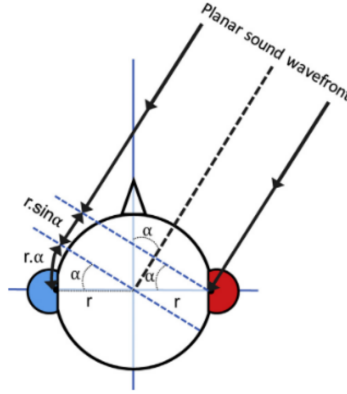
### ILD - Interaural Level Difference

The sound localisation for frequencies beyond 1500 Hz is obtained with the **Interaural Level Differences (ILD)**. In particular, this is true for frequencies function of a wavelength compared to the head dimension, equivalent to a  $f_{min}$  of 1960 Hz.

ILD localisation is possible thanks to the **Shadowing Effect**, where, the sound-waves with these wavelengths are blocked by the dimension of the head, arriving with a greater amplitude to the nearest ear and with a lower one to the furthest. This effect increases from a sound source positioned between  $15^\circ$  and  $60^\circ$ , while a sound source positioned between  $60^\circ$  and  $90^\circ$  is characterised by an acoustic diffraction from different paths that coincide in phase with the head-shadowed ear, causing a summing of the effects and resulting in an increased perceived sound level [15].

A possible mathematical expression of ILD, presented in [16], is made by the following equation, in function of the frequency of the sound source and of the angle of incidence  $\alpha$ :

$$ICLD(\alpha, f) = 0.18 * \sqrt{f \cdot \sin(\alpha)} \quad (2.3)$$



**Figure 2.6:** Sound-Wave Path for the Human Head. Image courtesy of [17]

### Cone of Confusion

ITD and ILD cues are part of the classic **duplex theory** proposed by Lord Rayleigh in 1907 [1]. Even if these are important parameters, they are not enough for audio localisation, since it is possible to find infinite position with the same ITD or ILD.

For this reason, another important concept has been introduced, called **Cone of confusion**. This is defined as the set of the point which shares the same ITD and ILD cues, resulting in an ambiguity of localisation. An extreme cone, as explained in [12], is the median plane, where both delays and intensity difference are 0. The cone of confusion can be visualized as a cone with symmetry axis along a line passing through the listener's ears and upper vertex in the center point between the listener's ears [18]. It is possible to say, now, that ITD and ILD cues determine alone not the precise position, but the cone of confusion where the sound source is located.

The ambiguity on localisation inside the cone can be solved with small head movements, as proposed by Hans Wallack in his work [19] from 1940, or if the sound is repeated [5].

The head rotation helps to localise sounds thanks to what is called "**the cocktail party effect**", reflex that make the subject turn their head in direction of the perceived sound source, introducing ITD, and helping to separate the source from the other sounds and resulting in a better localisation [20].

## HRTF- Head Related Transfer Function

A similar shadowing effect from the external ear, and especially from the auricle, allows humans to establish if sounds arrive from the front or from behind the head of the listener, with a band-pass filter behaviour. This monaural cue, which takes the name of **Head Related Transfer Function (HRTF)** is true also for the vertical localisation [12].

This has been explained by Batteau in [17], as following: both direct and reflected sounds arrives from the pinna to the ear canal, but their relative delay is direction dependent on the different point of reflection on the pinna. The notch and the peaks on the interference between these two sounds spectra are, for this reason, direction dependent, giving spatial information about the position of the sound source. This has been explained mathematically as the combination of two reflections  $A_1$  and  $A_2$ , with different time delays  $\tau_1$  and  $\tau_2$ . The transfer function of the pinna is expressed as:

$$H(f) = 1 + A_1 \cdot e^{-j2\pi f\tau_1} + A_2 \cdot e^{-j2\pi f\tau_2} \quad (2.4)$$

This extreme simplification make the theory suitable only for frequencies between 2-3 kHz, comparable to the pinna dimension of 65 mm. In addition, the complex and subjective shape does not allow to find such a model for the entire frequency domain.

For this reason, further studied have been made, such as the narrow-band noise theory proposed by Blauert in [3], with the conclusion of the pinna effect similar to a filter effect on the arrival sound. Another important conclusion is the importance of the frequency range around 5-6 kHz range, for the vertical localisation and the front-back ambiguity.

The reception of a signal to both ears from a point source can be modelled as an LTI system. The anatomy of the auricle, and in general of the head, influence the reception of the sound source, characterising a filtering process defined as HRTF (Head Related Transfer Function), and are mathematically defined, for both ears, as:

$$\begin{aligned} H_L &= H_L(r, \theta, \phi, f, a) = \frac{P_L(r, \theta, \phi, f, a)}{P_0(r, f)} \\ H_R &= H_R(r, \theta, \phi, f, a) = \frac{P_R(r, \theta, \phi, f, a)}{P_0(r, f)} \end{aligned} \quad (2.5)$$

with  $P_L$  and  $P_R$  defined as the complex sound pressure in a frequency domain for both left and right ear and  $P_0$  representing the complex pressure in the free-field

which there would be at the center of the head, without the presence of the head. This parameter can be calculated, as proposed by Morse and Ingrad [21] as:

$$P_0(r, f) = j \frac{k \rho_0 c Q_0}{4\pi r} e^{-jkr} \quad (2.6)$$

where  $\rho_0$  is the density of the air,  $c$  is the speed of the sound,  $Q_0$  is the intensity of the point sound source and  $k$  is the wavenumber.

The parameter  $a$  of the HRTF couples corresponds to a subjective parameter due to the subjective anatomic configuration of each human ear, and this, with also the distance  $r$  between the source (since the diffusion is made in a free-field) and the angle  $\phi$  are parameters which are less considered in the simplest model of the HRTF.

### **Accuracy of the Horizontal Location**

In general, the accuracy of perception depends on the quality of the sound material arriving to the listener. In [3], a test has been made with a white noise stimulus of 70 phne for 100 ms, for calculating the accuracy in function of the angle of incidence of the planewave for the listener.

From this, an uncertainty of 3-4° has been obtained for the frontal position, azimuth of 0°, of 5-6° for the back position, with an azimuth of -180°, and of 10° for a lateral position, with an azimuth of 270°.

Together with the incidence angle, another parameter which influences the accuracy of perception on the horizontal plane is the frequency. In fact, as demonstrated in the work of Yost and Zhong in [22], humans better perceive the frequencies below 1000 Hz, worst the frequencies between 1000Hz and 3000 Hz and in an intermediate way the frequencies over 3000 Hz. The poorest perception in this range of frequencies is due to the poor influence of the ITD and ILD cues, which suffers from too high frequencies and too low frequencies respectively.

Lastly, considering the broadband, it has been demonstrated in [22] that the wider is the band, the better is the perception of the sound, with the comparison between one or more than one octave.

Accuracy of azimuthal sound source localization by interaural time difference (ITD) and interaural level difference (ILD) according to frequency.

Binaural localization cue	Localization accuracy		
	< 1000 Hz	1000–3000 Hz	> 3000 Hz
ITD	Good	Mediocre	Impossible
ILD	Impossible	Mediocre	Good

**Figure 2.7:** Accuracy of Horizontal Localisation for Frequencies Ranges. Image courtesy of [6]

### 2.3.2 Perception on the Vertical Plane

The perception on the vertical plane is more complex than the one on the horizontal, due to the particular disposition of the ears, equidistant from the median axis and located in the sagittal plane.

#### Pinna Effect and HRTF

First of all, the perception along the vertical plane is not regulated by bin-aural cues, since the two ears are equidistant to the median plane and located at the same height. Instead, this type of localisation of a sound source is possible thanks to monaural principles such as the one from the pinna folds, and in particular thanks to the reflections of the sound in these areas [3]. This is called **pinna effect**. In addition, the pinna reflections generate the HRTF which are involved in the vertical localisation, but this concept has been explained in the previous section.

#### Frequency Dependence of the Vertical Localisation

In his work of 1930 [23], Pratt made an experiment about the localisation of pure tones along the median plane, determining a difficulty for the listener to perceive the height of pure tone, but noting a correlation between the frequency of the tone and the perceived height. In fact, even if the tones were played from the same height, the listener perceived those higher for higher frequencies and lower for lower frequencies.

## Accuracy of the Vertical Location

The vertical localisation, due the dimension of the pinna, is related to the frequency of the sound source. It has been demonstrated in [24] that short soundwaves are better localised by the ears, due to the correspondence to higher frequencies, especially for complex sounds with components higher than 7 kHz.

**Accuracy of sound source localization in the vertical plane by head-related transfer function (HRTF) according to frequency.**

Monaural localization cue	Localization accuracy	
	< 7000 Hz	> 7000 Hz
HRTF	Moderate	Good

**Figure 2.8:** Accuracy of Vertical Localisation for Frequencies Groups. Image courtesy of [6]

### 2.3.3 Perception of Distance

#### Real Distance from the Source

In general, the perception of distance is a hard goal to archive correctly. As defined in [12], humans tend to underestimate distance of about 1.6 m if the source is further and overestimate closer ones of 1.6 m. A mathematical formulation for this law is expressed by Zahorik in his work [25] of 2002, where the perceived distance  $r'$  is expressed in function of the real distance  $r$  as:

$$r' = kr^\alpha \quad (2.7)$$

In the equation,  $k$  is a constant equal to 1.23 and  $\alpha$  is a parameter dependent on various factors as the environment and the subjects, but it can be considered approximately equal to 0.4.

#### Loudness

Loudness is another parameter which influences the perception of the distance. In fact, in free field it decreases by 6dB for doubling the distance from the listener to the source (by the  $1/r$  law). This parameter is most effective if the listener is used to the loudness of the source which emits the sound.



## **Air Absorption and Head Diffraction**

The last parameters that influence distance perception is the air absorption for high frequencies, as a low pass filter effect, and acoustic diffraction of the head. These two, as the other parameters, are still subject of study, and remain an open discussion field.

### **2.3.4 The Parameters of Perception- Position**

Some parameters which characterise the position of a sound source are defined in [3] and reported in this section, with the aim of being useful for the understanding of the next sections.

The **localisation Blur** is defined as the smallest change in a specific attribute (or specific attributes) of a sound event, or to another auditory event which is sufficient to produce a change in the localisation of the auditory event. For example, in terms of direction, it can be defined as the small variation of direction which is translated in a change of perception.

The **Locatedness** is defined as the spatial perception of the sound event, in terms of its extent, and evaluated together with position and the extent of other auditory events.

The last two parameters, defined in [26] and [27], are the **Minimum Audible Angle (MAA)**, defined as the minimum change in the direction of a static source to define this as changing from left or right from the original direction, and the **Minimum Audible Movement Angle (MAMA)**, corresponding to the smallest arc which a moving sound source can move for being discriminated from a stationary source.

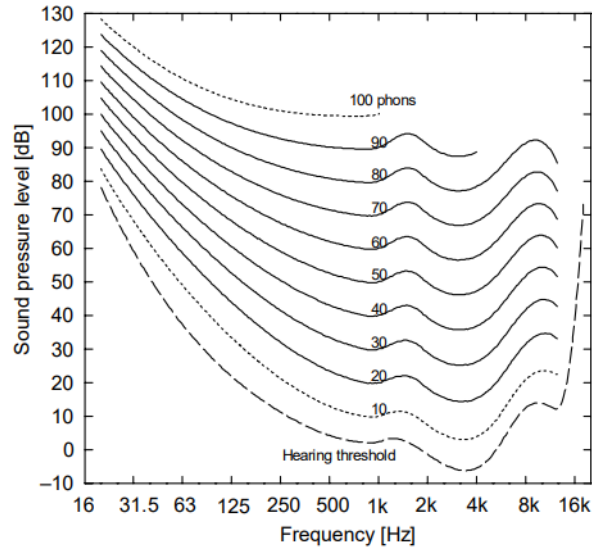
### **2.3.5 Other Parameters of Perception**

The perception of the sound is not only characterized by its position in space, but also by some other subjective parameters, as the loudness, the pitch and the timbre [5]. All of them are linked, thanks to psychoacoustic studies, to objective parameters of frequency and intensity.

The **loudness** is linked to the intensity of the sound but also to the frequency. In fact, Fletcher and Munson, in their study called "Loudness, its definition, measurement and calculation" [28], elaborate the **Equal Loudness Contours** to

indicate the trend of same perceived loudness at different pure tone frequencies. [5]. The value of every curve is expressed in phon and it is equal to the perceived loudness in dB for every curve, with the reference of 1000 Hz. For example, a curve of 20 phons is equal to a curve which Intensity at 1000 Hz is perceived as 20 dB. A particularly interesting curve is the one at 0 phone, that is called absolute audibility threshold, and represents the minimum amplitude for a human to hear a pure tone of that frequency. This curve is used, for example, for the digital audio compression, to filtrate all the non audible contribution of the spectrum. Another interesting curve is the highest one, at 100 phons, which represents the pain threshold [8].

During history, some other different curves have been elaborated, and currently the ISO 226:2003 [29] standard curves are used (in Figure 2.9).



**Figure 2.9:** ISO 226:2003 Standard Curves. Image courtesy of [29]

The **pitch** is linked to the frequency, or, in the case of complex sounds, to the fundamental frequency, if this is not masked by the harmonic frequencies, for a phenomena called virtual pith. In general, the pitch helps to establish, between two sounds, which one is higher or lower.

The **timbre** is linked to the waveform of a sound, so, in general, to the frequency content, the amplitude and the envelope, evolution of the first two characteristics among time. This parameter allows to distinguish the same frequency with the same amplitude played by two different instruments, like a Fourier transform made by the ear [5].

## 2.4 Stereophony

The Stereophonic listening has been defined by Bernfeld in [20] as *"the listening of signals emitted by two or more loudspeakers, each creating crossed signals at both listener ears"*.

These type of methods implement the localisation of the sound source using ICLD (Inter-Channel Level Differences) and ICTD (Inter-Channel Time Differences) principles, corresponding to similar concepts of the previously viewed ITD and ILD, but which can not be considerate as the same, due the complex relationship between them and the fact that every signal not only goes to a single ear, but affects also what the other one perceives.

Stereophony methods for the horizontal plane can be classified by the type of technology used and by the number of channels of reproduction.

For the first classification, it is possible to divide the methods into **physically motivated systems**, which reproduce a physical approximation of the desired sound field with particular microphone arrays recording to capture a stereophonic image. Belonging to the second class the **perceptual motivated systems**, which reconstruct the sound only rendering the perceptually characteristics, with less computational needs [27]. In this second case, it is possible to initially feed the loudspeakers with the same signal, and then apply delays and intensity differences in order to obtain panning.

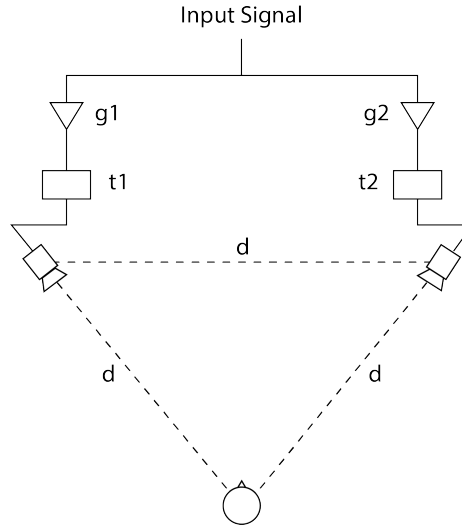
### 2.4.1 Two-Channels Stereophonic Setup

The two-channels stereophonic setup is a binaural stereo configuration made by an imaginary equilateral triangle where the loudspeakers are positioned on two of the vertex, with an angle of  $60^\circ$ , and the head of the listener at the other opposite point. This setup is the most common example of an audio reproduction system.

Considering feeding the loudspeaker with the same signal, the easiest case of a sound located in the center is obtained, equidistant to both the loudspeakers and in the same axes, perfectly in front of the listener, as a monophonic signal.

However if the goal is to locate the sound in an arbitrary position in the Stereophonic arc between the loudspeakers, it is possible to feed the two loudspeakers with different signals, that will be elaborated as a single one by the listener's brain for the Summing localisation Principle. This process is called **Panning** of the sound

source and consists in positioning the sound source (called, in this case, **phantom source** in a point between the loudspeakers) working on delays of emission and Amplitude Ratio between the two loudspeakers.



**Figure 2.10:** Standard 2-Channel Stereophonic Setup

A con of this binaural method is the **cross-talk effect**, which consists in the signal of one channel that arrives to the opposite ear, affecting the stability of the stereo image. This problem, which is present only with binaural stereophony realised with loudspeakers, can be easily solved with a cross-talk cancellation method, called **transaural audio**, which eliminates the sound from the wrong channel if the listener is situated in the sweet spot, equidistant from the sources. Out of this zone, in general for all the multichannel system, the image can be right perceived with a head tracker system. In addition, this system can simulate also the presence of the sound source behind the listener, even if there are no loudspeakers in that point [27].

In general cross-talk effect is one of the reasons why the couples ITD/ICTD and ILD/ICLD can not be considered as the same.

Different panning methods are possible, with the use of the only Time Delays or Intensity Difference, obtaining Time Difference Methods (TD) and Intensity Difference Methods (ID) respectively. It is also possible to use both, obtaining methods defined as Time Intensity Difference ones (TID), which use psychoacoustics curves to better merge both the principles.

The panning obtained by the **Sine Panning Law** [20] is an example of Intensity

Difference method. This law express the gains of the loudspeakers ( $G1$  and  $G2$ ) in function of the ratio between the azimuth angle  $\theta$  of the two loudspeakers and the desired angular position of the sound source  $\Phi$ . The law takes the following expression:

$$\frac{\sin(\theta)}{\sin(\Phi)} = \frac{G1 - G2}{G1 + G2} \quad (2.8)$$

This formula is valid if the listener's head is pointing directly forward. If the head of the listener can move, it is useful to consider the alternative approach of the **Tangent Panning Law** [30], which expresses the position of the sound source in terms of a tangent ratio, taking the following expression:

$$\frac{tg(\theta)}{tg(\Phi)} = \frac{G1 - G2}{G1 + G2} \quad (2.9)$$

In these equations the assumption is made for a sound source with only amplitude changes between the channels, and this is demonstrated in [20], with the first equation, which is true for frequencies below 500-600 Hz, where the perception is only regulated by ITD, while, for higher frequencies, where the perception is regulated by both ITD and ILD, the second law is true.

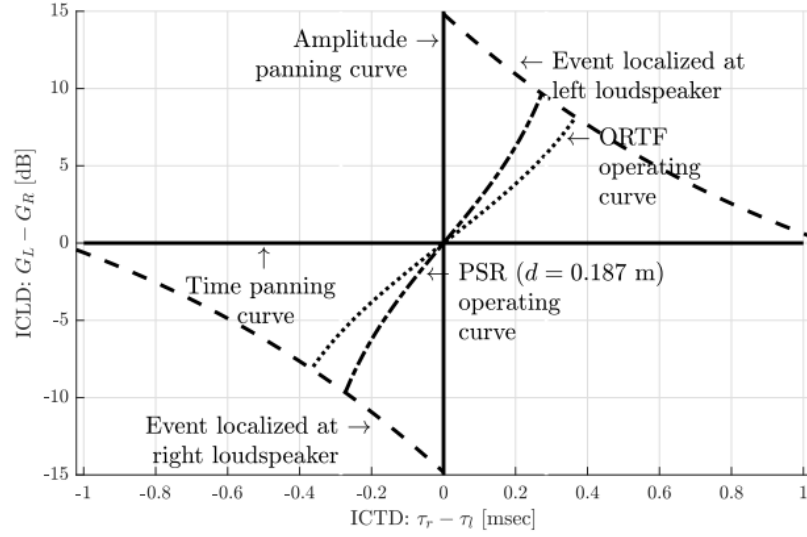
The introduction of delays in the Panning methods requires the use of Time-intensity psychoacoustic curves, as the ones of Frannsen [31] and Williams [32] (Image 2.11), that express three lines of points with coordinates (time, amplitude) for which the auditory event is localised in the left, center and right position. On the X-axis the time delay between the left and the right channel in milliseconds is represented, while on the Y-axis the gain difference between the right and the left channel, in dB, is represented.

Drawing a line (operative curve) from one curve to the other, a panning curve between the two channel is obtained. A vertical line passing for the 0 ms x-coordinate point is equal to an amplitude only panning, while a horizontal line passing for the 0 dB y-coordinate is equal to a time difference panning. At last, the use of a straight line for combining the two curves results in a Time Intensity Linear Difference (TILD) method. [2].

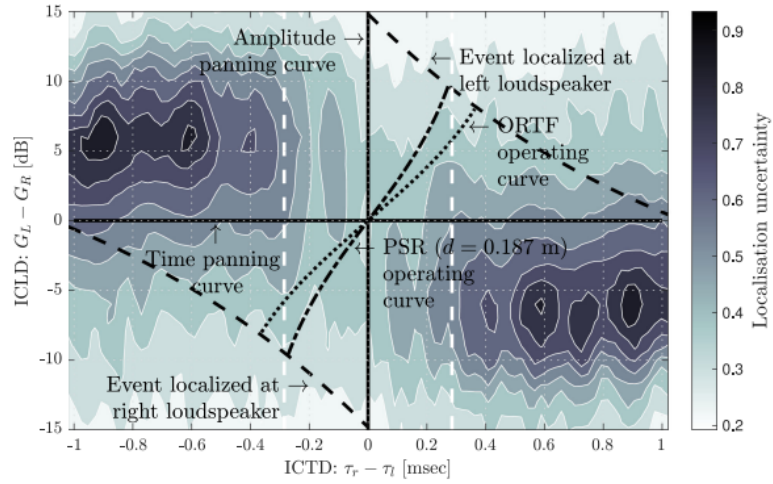
An ICLD difference of  $\pm 12dB$  is enough to pan a source completely on a direction without TID, while, in the same way, a delay of  $\pm 1$  ms is the minimum delay for panning a source in an extreme direction without the use of level differences.

The analogy between physical methods and perceptual ones is clear if a linear operating curve passing for the central point (0,0) is considered. Considering the

perceptual reconstruction method, changing the microphones radius of the array (distance between the capsules) on the physical one, correspond to vary the angular coefficient of the panning line.



**Figure 2.11:** Williams Time-intensity Psychoacoustic Curves. Image courtesy of [33]



**Figure 2.12:** Contour Plot of Localisation Uncertainty for a Centered Position. Picture courtesy of [33]

As investigated in [33], the choice of the curve can influence the degree of certainty of the perceived position of the sound source. Indeed, in the cited work an experiment has been run to determine the curve which provides the best certainty in localisation, in function of different delays and level differences. The figure 2.12 shows the localisation Uncertainty for the centered position, with analogue plots obtained for off-centered positions.

Form the figure it is possible to see that the best panning curve for localisation certainty is the one used on the Perceptual Sound Field Reconstruction method (PSR), explained in section 2.6.

Considering the physical reconstruction method, the PSR curve is really close to the one which has been obtained for the ORTF method (see section 2.4.3).

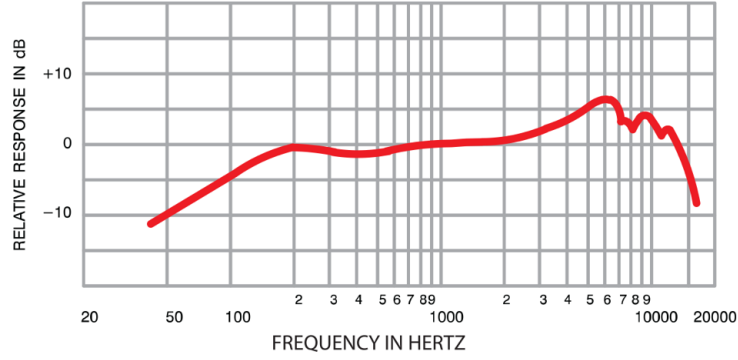
## 2.4.2 Microphones and Characteristics

As said in the introduction, this thesis work is concentrated on perceptual reconstruction method for panning. In order to the most complete overview of the other possibilities, also physical methods for rendering stereophony are presented, realised with particular microphone arrays which record sound material, each of which will be send to a different loudspeaker on a reproduction system.

The basic element of the recording is the **microphone**, transducer of the air pressure variations into an electrical signal.

Every microphone is different from another for the **frequency response**, which gives an information about how the sensibility changes (in terms of amplitude) along the frequencies, but especially for its **polar diagram**, which expresses the sensibility of the microphone in terms of the angle of incidence of the planewave. These parameters, which are considered for a soundwave incident on the capsule, are dependent on the construction features.

For the frequency response, it is possible to say, on first approximation, that a flat trend of the graph is an index of quality of the microphone, because it keeps the original sound, but, for some transducer, non-flat frequency responses are an index of particular and desired characteristics, for example, for recording a voice a particular coloration is preferred.



**Figure 2.13:** Frequency Response of a Shure Microphone. Image courtesy of [34]

Talking about the polar diagram of a microphone, this parameter can classify microphones into 7 main categories: Omnidirectional, Subcardioid, Cardioid, Supercardioid, hypercardioid, Shotgun and Bidirectional (or figure-8).

The Omnidirectional microphone is characterised by a spherical directionality, which means that every incidence-dependent soundwave is received in the same way by the microphone, unlike the Subcardioid that possesses an attenuation on the back.

The difference between the other microphones is the null angle, considered at the angle where the amplitude of the recorded soundwave is zero. Cardioid microphone has a zero at  $180^\circ$ , Supercardioid has two zeros as the hypercardioid (but changing the back answer), while the 8-figure, as the name said, has two zeros as well, but located at  $90^\circ$  and  $270^\circ$ , resulting in a polar response like an 8. The shotgun is the only microphone with four zeros.

Every directivity pattern is characterised by an equation which expresses the response of the microphone  $\Phi$  in function of the angle of incidence  $\theta$ .

The general equation has the form:

$$\Phi(\theta) = a_0 + a_1 \cos(\theta) + a_2 \cos^2(\theta) + \dots + a_N \cos^N(\theta)$$

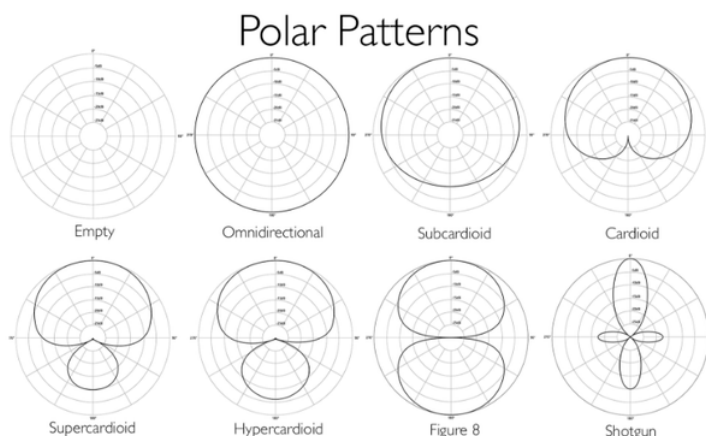
The number of parameters of the previous equation defines the order of the microphone. For example, if the terms  $a_2 \neq 0$ , the microphone will be of order 2. The commercial microphones are mainly of 0 or 1st order, with some examples of high order microphones presented in the following sections.

Examples of common directivity patterns are the following:

$$\begin{aligned} \text{Omnidirectional: } \Phi(\theta) &= 1 \\ \text{Cardioid: } \Phi(\theta) &= 0.5 + 0.5 \cdot \cos(\theta) \\ \text{Supercardioid: } \Phi(\theta) &= 0.375 + 0.625 \cdot \cos(\theta) \\ \text{Hypercardioid: } \Phi(\theta) &= 0.25 + 0.75 \cdot \cos(\theta) \\ \text{Figure-8: } \Phi(\theta) &= \cos(\theta) \end{aligned}$$



The polar diagram equation is plotted and represented inside particular circumferences. This representation gives information also about the dependency on the frequency, with every circumference in the graph which represents an octave. A particular aspect related to this is the fact that, for high frequencies, all the microphones became directive. These basic polar patterns have to be intended as an ideal configuration, for categorization purposes. Every microphone is affected by a gap, even if small, to this behaviour.



**Figure 2.14:** Polar Patterns of a Microphone. Image courtesy of [35]

### 2.4.3 Microphone Arrays for 2-Channel Stereophony

Basic microphones directivity patterns can be merged to obtain custom directivity. This is possible thanks to the microphone array, formed by a group of two or more microphones in a particular configuration, depending on what is required to be recorded.

The first distinction between the microphone arrays can be made in three categories: Coincident, Near-Coincident and Spaced arrays, and this categorization distinguishes the distance of the capsules of the microphones.

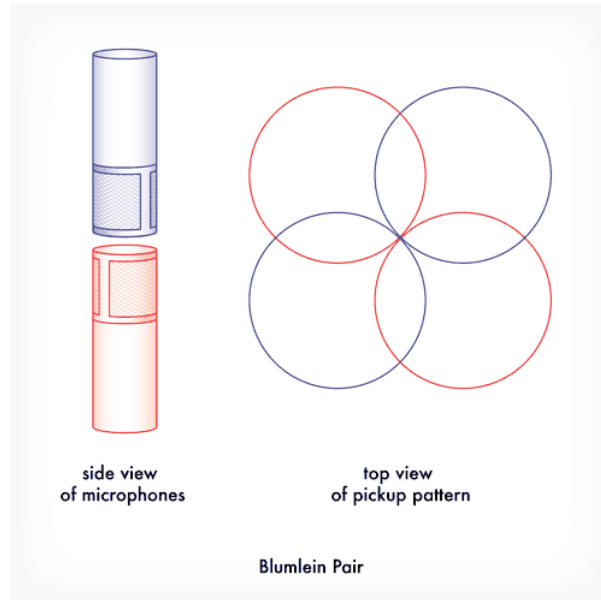
In addition, every group of microphones works with a different concept between ICTD, ICLD, or both of them.

#### Coincident Microphone Arrays

The coincident Microphones are characterized by a distance between the capsule

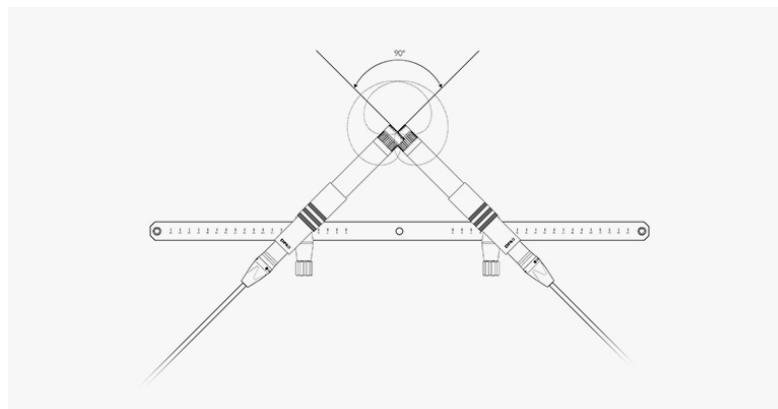
$r = 0$  and a variable angle between the capsules [20]. Actually, the distance equal to zero is ideal, since the microphones have not to touch each other to avoid mechanical noise. This configuration works only by capturing differences of intensity, so with the only concept of ICLD, with a delay  $\tau=0$ . Examples of this configuration are the Blumlein Array, the X-Y and the Mid-Side.

The **Blumlein array** configuration is probably one of the first to have been invented. The array, dating back to the '30s, is formed by two 8-figure microphones with the coincident capsules angles at  $90^\circ$ . This allows to have four different areas, left and right for both front and back direction, but, as a con, every channel can be affected by the reverb of the opposite one, so, for this reason, this configuration has been substituted by particular microphones as SoundField, presented in 2.10.1.



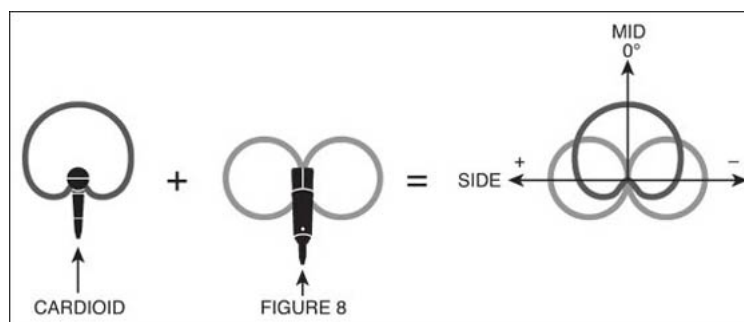
**Figure 2.15:** Blumlein Stereophonic Configuration. Image courtesy of [36]

The **X-Y** is a very common configuration realised with two cardioids placed at  $90^\circ$ . In respect to Blumlein, the frontal area has a major coverage, with a more stable stereophonic pattern. This configuration determines a weak spatialization due to the low reverb recorded. Also, another important parameter to take into account is the distance, because a large distance from the source determines a loss in the low frequencies.



**Figure 2.16:** X-Y Configuration. Image courtesy of [37]

Lastly, the **Mid-Side** configuration is made by the merging of two signals, as suggested by the name, one for the Mid and one for the Side, realised with a frontal Cardioid and with a lateral 8-figure. This configuration is not a proper stereophonic configuration, because the result Mid+Side is calculated by a matrix for both the Left and Right channels. The best recording angle is obtained from 0 to 90°, out of which the monophony is more rendered.



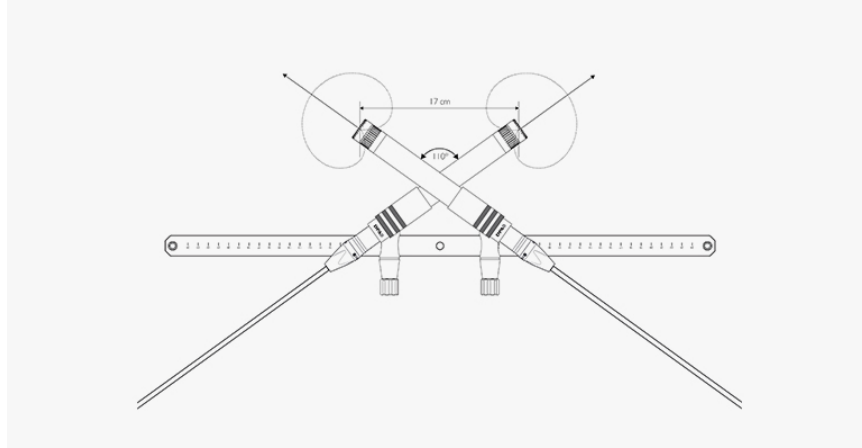
**Figure 2.17:** Mid-Side Configuration. Image courtesy of [38]

### Near-coincident Microphone Arrays

This configuration is characterised by a limited distance between the capsules of the microphones (the order is the human head dimension, for capturing localisation cues similar to the human auditory ones), and a variable angle between the axis. This group of microphones works with both the concepts of ICLD and ICTD. The use of cardioid determines, also in this case, a loss at low frequencies if the distance between listener and source is too large, resulting in a lack of energy and richness of the sound.

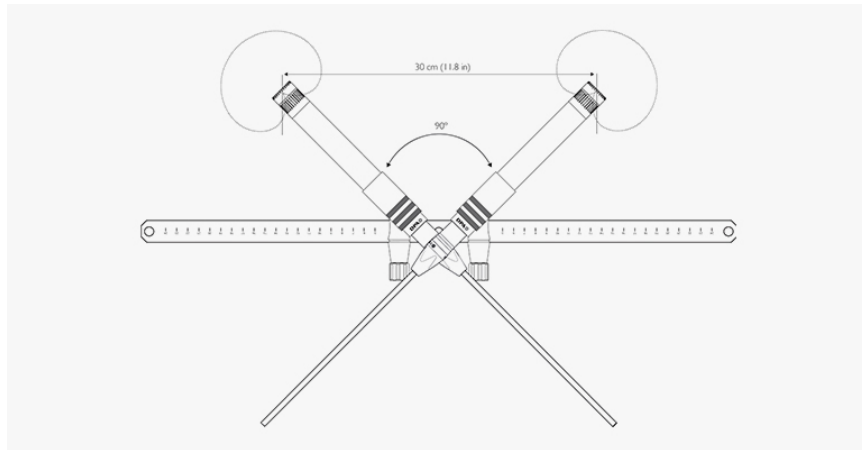
NOS and ORTF are part of these configuration arrays.

The configuration **ORTF**, Office de Radiodiffusion Télévision Française, is made by two cardioid microphones with a distance of the capsules of  $r=170$  mm and an angle between the two axis of  $110^\circ$ . This configuration gives an optimal stereophonic image with a wide angle of stereophony.



**Figure 2.18:** ORTF Configuration. Image courtesy of [37]

The **NOS** configuration, from the Nederlandse Omroep Stichting, is made by two cardioid with an angle of  $90^\circ$  and a distance of 30 cm. This bigger distance introduces a major delay and a wider stereophony.

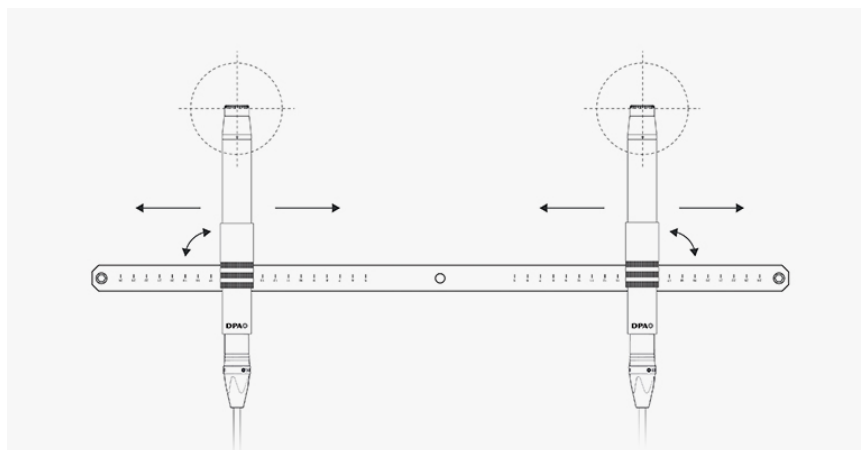


**Figure 2.19:** NOS Configuration. Image courtesy of [37]

## Spaced Microphone Arrays

The configurations with two (or more) microphones, with a distance of 60 cm or more, are classified as Spaced microphone arrays. This type of technique uses mainly ICTD, since the distance between the microphones is lower in respect of the distance from the mic axis to the source [20]. For the wider distance, omnidirectional microphones are necessary in the majority of cases, to better capture the sound in the center of the array. The use of ICTD implies the need of an optimal ambience of recording, since the reflections of the sound in the room become important. In general these arrays are used for recording a big set of instruments. If only one instrument has to be recorded, using near-coincident methods is the best solution.

The easiest case of this type of array is the **A-B** technique, made by two parallel omnidirectional, which gives a great response at low frequencies, but a low mono compatibility, due to some comb filtering effects.



**Figure 2.20:** A-B Configuration. Image courtesy of [37]

In an A-B configuration, the distance is variable, for this reason it is more correct to consider the A-B as a techniques family [37]. The right distance between the capsules and the resulting delay is determined in function of the outer position of the sound source, with a diagram as the one in the image below. The right configuration, in terms of angle, has also to take into account the distance from the source, to determines how much direct sound and reflected sound has to be included in the recording. Similar curves are obtained also for the previous configurations.

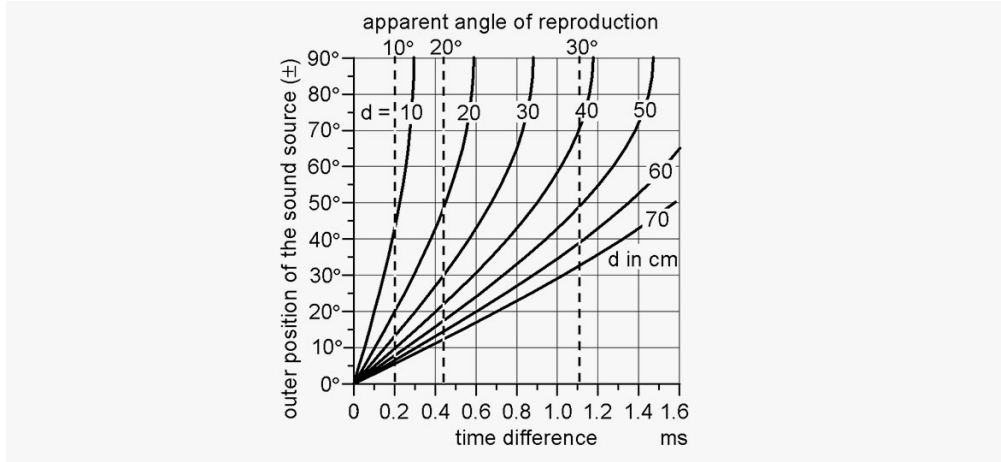


Figure 2.21: A-B Configuration Curves. Image courtesy of [37]

## Conclusions on the Microphones Techniques

The right microphone array is determined by the type of single or group of instruments that is desired to be rendered. The parameters which influence this choice are:

1. **SRA**, defined as the sector of original sound field, in front of the microphone array, where the stereophonic image is perceived as stable in the loudspeakers system. This parameter is obtained from particular curves (called ISO-SRA), and are a function of the distance between the capsules and the angle between these. For example, the X-Y has an SRA of  $180^\circ$ , because the angle of reproduction is  $(\pm 90^\circ)$ . Out of this region, sounds can be captured, but their images are not rendered in a stable way.
2. **Angular Compression/Expansion** is the ratio between the SRA and the angle of reproduction. For example, X-Y has an SRA of  $180^\circ$  and the Angular Compression/Expansion value, for a system of  $60^\circ$ , is equal to 3.
3. **Angular Distortion** is an effect which causes a right reproduction of the source in the center and at the extreme points of the reproduction system, while the other points are moved of  $5^\circ$  or  $6^\circ$  toward the outside. Every microphone array has its own Angular Distortion.

## Dummy Head

A separate paragraph is dedicated to the **Dummy Head system** [39] [40]. Even if the characteristics can allow it to be classified as a near-coincident microphone array, this system, of which first model dates back to 1952 (Neumann KU-80) is mainly used for 3D audio, or for binaural recording. Since it is possible to define the previous arrays as **space-related**, because the goal is to encode spatial information of the sound, this method is called **head-related** because it is based on human head information.

Dummy Head, or Head and Torso Simulator (HATS), share the same shape of a human head, with two pressure omnidirectional microphones located on the eardrums, and this configuration allows the listener to hear as if his head is centered on the sound scene. This system has many artistic applications, but it can also be used for studies on the effect of certain sound sources in other fields, as for the monitoring of ambience and industrial noise, or for another important application, the HRTF acquisition.

The recorded sound arrives at the microphones and is then equalized with a diffuse field system, which allows for the recordings to be suitable also for the loudspeakers reproduction on two or more channels.



**Figure 2.22:** Dummy Head System Neumann KU100. Image courtesy of Georg Neumann GmbH, Berlin

## 2.5 From 2-Channels to Multichannel Stereophony

As seen in the previous section, stereophony system can be grouped by the number of channels. The methods can be divided into binaural stereophony, seen in 2.4, multichannel with a low number of loudspeakers (like 5.1 to 9.1), and systems with more than 9 loudspeakers (as 10.2 or 22.2). The goal of the last two is, in general, to try to reconstruct a complex image of the sound field around the listener, which represents the center of the auditory scene.

The methods based on the number of channels are defined in the ITU-R BS.2159-8 [41] report, with a description of the position of every channel in the space and the address of use for every method.

Method	#Channels	Characteristics
2-Channels	2	Easy to implements but small sweet spot
Multichannel	from 5 to 9	Large enough sweet spot but needs of managing psychoacoustics effects
Sound Field Reconstruction (SFR)	>9	Elegant Mathematically but hard to render

Considering all these parameters, the multichannel solution seems to be the best one, because it is the method that allows to obtain a low localisation error with a source easy to localise in the largest possible area (sweetspot), needing a lower computational cost.

Another distinction between multichannel techniques can be made considering the audio format, as discrete or matrix. In the first case there is a one to-one correspondence between channels and speakers, while in the second type, the original channels are encoded in a smaller number of channels[27].

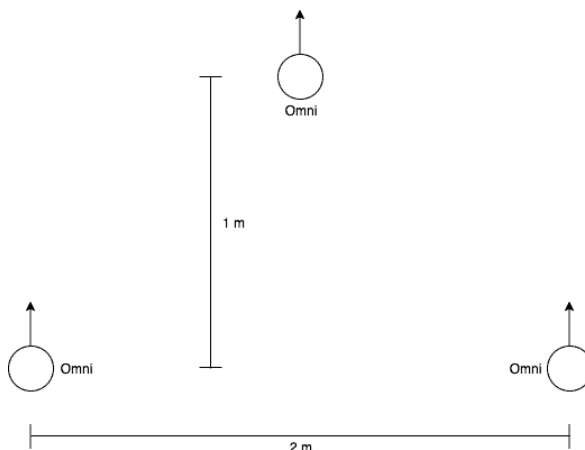
The first question that a person can ask their selves is: why are more than two channels necessary to render spatial information?

The first reason can be found by considering a particular Spaced Microphone Array made mainly for orchestral recordings by Decca Records, called Decca Tree. This microphone array solves the **hole in the center problem** of the A-B system, due to the too large distance between the microphones which causes a poor stereophonic image at the center of the array. In AB, in fact, the signal at the extremes is correctly recorded, but the instruments in the center are poorly rendered.

This array is made by three omnidirectional microphones with spherical capsules, with two lateral microphones as the A-B, positioned at a 2m distance and a central one, 1 m ahead the other two, for a better mono image than the A-B, and this is



due to being more affected by the precedence effect, which provides lower mixing level needs, and this results in a lower comb-filtering effect.



**Figure 2.23:** Decca Tree Configuration. Image courtesy of [42]

In general, multichannel recording systems improve binaural stereophony with loudspeakers in terms of accuracy, localisation, and naturalness of the source, in addition to a wider sweetspot.

But a high number of microphones is not always the best solution for reaching this result. Imagine recording an orchestra with a high number of elements and using single microphones, this is really expensive in terms of equipment and also, this is probably not the best solution to obtain a proper acoustic of the room.

In this case, the solution is in between, with one or a couple of microphones for the front scene, called **main microphones** and **room microphones**, to capture the acoustic of the room.

In this section some examples of this configuration are presented.

### 2.5.1 5.1 Multichannel Configuration

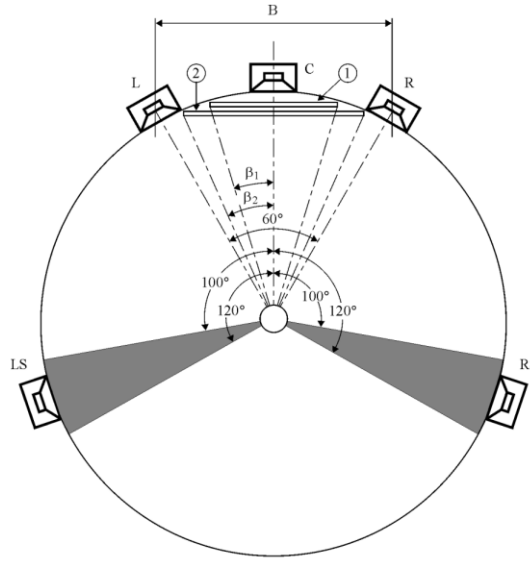
The 5.1 configuration is the easiest the and most common commercial multichannel setup, which is represented by commercial standards as Dolby Digital 5.1 and DTS 5.1.

The technology was invented by Dolby in the 70's for a cinema purpose, and was immediately appreciated for the better audio spatialization in respect to the classic stereophony, with particular attention to the central channel useful for the dialogues.

This system is based on five reproduction channels positioned in a circular disposition, with the listener at the center of the circumference.

Considering the listener view azimuth as  $0^\circ$  and a positive angle in a clockwise sense, two channels (L and R) are positioned at  $\pm 30^\circ$ , as the normal stereophony. What has been added are two surround channels (LS and LR) at  $\pm 110^\circ$ , which provide a better spatial perception on the back position, and a center channel, located at a  $0^\circ$ . This last channel, as said before, is really useful for the dialogues in the field of cinema, since the visual is a stronger perception cue in respect to the audio, with the sound perceived as if it is coming from the center, the screen, regardless of the actual L/R position[43].

All the 5 channels are aimed to be positioned at 1.20 m height, at the average elevation of the human ears in a seated position. A low frequency channel (LFE) is also added on the floor, realised with a subwoofer which received the original sound and then provided a filtering of this with a low pass filter. The best position for this channel is under the seating position. The correct disposition and the guidelines for the use of this system are shown in the ITU-R BS.775-3 standard [44].



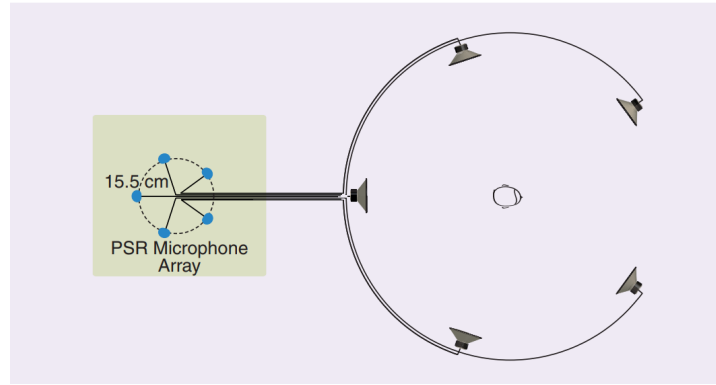
**Figure 2.24:** 5.1 Surround Configuration. Picture courtesy of [44]

A particular extension of the 5.1 configuration is the 7.1, which adds another two surround channels at  $\pm 140^\circ$  fed with the same signal of the other surround channel of the same side, but both with a lower gain [44].

## 2.6 Perceptual Sound Field Reconstruction- PSR

Perceptual Soundfield Reconstruction is a psychoacoustic panning method which aims to render a convincing auditory image on the horizontal plane using time intensity panning. The method was initially proposed by Johnston et al. in [45] and in this section further considerations from the work of De Sena et al. [2] are presented.

The array configuration for PSR is a circular disposition with the listener at the center. Every loudspeaker, equidistant from the others and from the listener, is linked to his microphone, positioned in the same configuration, with a one-to-one correspondence. The microphone array radius is variable, but it has been demonstrated that the optimal one is 15.5 cm.



**Figure 2.25:** PSR Configuration for 5.1. Image courtesy of [27]

### 2.6.1 How Many Loudspeakers Render a Single Sound Source?

Let's now consider we have access to an arbitrary number of loudspeakers. The question is, for an arbitrary configuration, are all the loudspeakers necessary to render a single sound source?

The starting point can be the case of near field, with a single source positioned near the listener at a certain distance  $r$ . Every loudspeaker emits a certain time-dependent sound pressure  $x$  equal to:  $x(t) = s \cdot e^{j2\pi ft}$ .

The sound  $y(t)$  received by the listener is mitigated by the delay of arrival with the following expression:  $y(t) = \frac{1}{r} \cdot x(t - \frac{r}{c})$ , where the ratio between  $r$  and  $c$  comes from the velocity law  $x(t) = v \cdot t \rightarrow t = \frac{x}{v}$ . In addition, the term  $\frac{1}{r}$  takes into account the spherical shape of energy.

Rewriting the previous equation, we obtain:

$$y(t) = \frac{1}{r} \cdot x\left(t - \frac{r}{c}\right) = \frac{s}{r} \cdot e^{j2\pi f \cdot (t - \frac{r}{c})} = \frac{s}{r} \cdot e^{j2\pi f t} \cdot e^{(-j2\pi f \frac{r}{c})} \quad (2.10)$$

Considering  $f=c$  (speed of sound, 343 m/s),  $\frac{2\pi f}{c} = k$  (wave-number) and with the assumption of far-field (source and listener distant enough), which implies  $\frac{1}{r} = 1$ , the final expression for the pressure  $p$  produced by an  $l$ -th loudspeakers in a certain listening position  $P[x,y]$  became:

$$p_l(P, t) = s_l \cdot e^{jkct} \cdot e^{jk \cdot [xcos\phi_l + ysin\phi_l]} \quad (2.11)$$

The previous result can be extended to a speakers array made by  $L$  loudspeakers, located in a circular position with angles  $0 \leq \phi_0 \leq \phi_1 \leq \dots \leq \phi_{L-1} \leq \phi_L \leq 2\pi$ . The assumption made on this system is that the array is centered in the origin and the radius is large enough that, in the listening point, called  $P$ , the wave can be approximated by plane-wave source.

The total pressure and velocity of the soundfield is made by the sum of the individual loudspeakers. The expressions of these variables is equal to:

$$p(P, t) = \sum_{l=0}^{L-1} p_l(P, t) \quad (2.12)$$

$$v(P, t) = \frac{1}{\rho c} \sum_{l=0}^{L-1} p_l(P, t) \cdot n_l \quad (2.13)$$

where  $\rho$  is the density of the air and  $n_l$  correspondent to the versor co-directional with the acoustic axis of the  $l$ -th loudspeaker.

The product between complex pressure and complex velocity produce the **complex intensity**, with a real part called **active intensity** which is co-directional with the wave propagation. The expression of the complex intensity is the following:

$$I(x) = \frac{1}{2} \cdot p(P, t) \cdot v^*(P, t) = \frac{1}{2\rho c} \cdot \sum_{l=0}^{L-1} \sum_{m=0}^{L-1} I_{lm}(x) \quad (2.14)$$

with  $I_{lm} = p_l(P, t) \cdot p_m^*(P, t)$ .

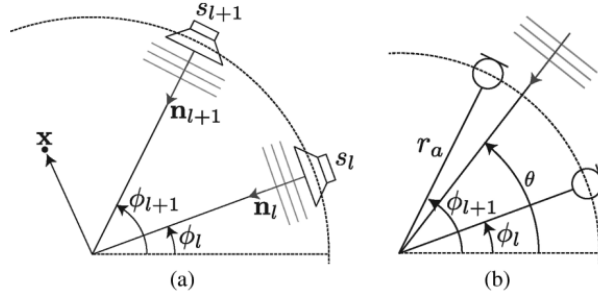
Every component  $I_{lm}$ , with  $l \neq m$ , contributes to a complex field which fluctuate around the space with a frequency  $|u_{lm}|$  and with direction of propagation equal to

$(\phi_m + \frac{\phi_l}{2} - \frac{\pi}{2})$ , orthogonal to the median plane between  $l$  and  $m$ . On the other side, every component  $I_{mm}$  contributes to a uniformly spatialized field in the direction of the  $m$ -th loudspeaker.

The **active intensity field** is expressed as:

$$I_a(x) = \frac{1}{2\rho c} \cdot \left[ \sum_{m=0}^{L-1} |s_l|^2 n_l + 2 \cdot \sum_{l=0}^{L-1} \sum_{m=l+1}^{L-1} |s_{lm}| \cdot x \cos(\varphi_{lm} + \langle u_{lm}, x \rangle) n_{lm} \right] \quad (2.15)$$

The goal is to have an active intensity field without fluctuations, because the fluctuations are a cause of interference, and to reach this the second element of the previous equation has to be minimized or suppressed. It is not possible to completely suppress the second term, because this would mean having only one active channel at a time, but it can be minimised only using two channels at a time, in particular two adjacent ones,  $m$  and  $m+1$ , as the interference intensity depends also on the angle between the distance of the speakers, so that's the way to minimize this.



**Figure 2.26:** Recording (a) and Reproduction (b) System for a Multichannel Array. Image courtesy of [2]

Let us now consider the same system of loudspeakers whose exits are fed by a microphone array in the same position. This constraint is important as every loudspeaker emits the same gain of the microphone, and the gain can be considered as equal to the polar diagram of the microphone in function of the incidence angle of the planewave ( $\Gamma(\theta)$ ), with values dependent on the following expression:

$$s_l = A\Gamma_i(\theta) \cdot e^{jkr_a \cos(\theta - \phi_l)} \quad (2.16)$$

where  $A$  is the amplitude of the soundwave and  $\Gamma$  is the directivity pattern of the  $l$ -th microphone.

For a planewave that arrives between two adjacent channels  $l$  and  $m$ , the cross-term gain contribute can be considered, whose expression is:

$$s_{lm} = A\gamma_{lm}(\theta) \cdot e^{j\varphi_{lm}(\theta)} \quad (2.17)$$

$$\text{with } \gamma_{lm}(\theta) = \Gamma_l(\theta) \cdot \Gamma_m(\theta) \text{ and } \varphi_{lm}(\theta) = 2kr_a \sin(\theta - \frac{\phi_l + \phi_m}{2}) \sin(\frac{\phi_l + \phi_m}{2})$$

In other words, this result demonstrates that, for rendering a single soundwave in a circular array of N loudspeakers (with  $N \geq 5$  for considering the system as a multichannel one), only two adjacent loudspeakers at a time are necessary, the ones in between the incidence angle  $\theta$ , positioned in  $\phi_m$  and  $\phi_{m+1}$ , with  $\theta \in [\phi_m, \phi_{m+1}]$ . In this way, the complex problem of the positioning of a sound source in a multichannel array is solved as a normal stereophonic configuration. This is not always true for multichannel configuration, for example with HOA (High Order Ambisonics), all the loudspeakers play at the same time for the rendering of a single sound source. On the other hand, this result implies the research for a better selectivity of the microphones of the array, with the gain  $\Phi(\theta)$  which has to be 0 for  $\theta \notin [\phi_m, \phi_{m+1}]$ .

### 2.6.2 Design of the Microphone Directivity

Having two adjacent active microphones at a time implies the first constraint on the equal loudness of the two directivity pattern, as:

$$\Gamma_m^2(\theta) + \Gamma_{m+1}^2(\theta) = 1 \quad (2.18)$$

This means that the sound at the center is constant in every direction. Under this constraint, valid for all the  $\theta \in [\phi_m, \phi_{m+1}]$  and  $m=0 \dots L+1$ , it is possible to express the terms ICLD and ICTD as:

$$ICLD = \Phi = \frac{A \cdot \Gamma_2(\theta)}{A \cdot \Gamma_1(\theta)} = \frac{\Gamma_2(\theta)}{\Gamma_1(\theta)} \longrightarrow \Phi[dB] = 20 \log_{10} \left( \frac{\Gamma_2(\theta)}{\Gamma_1(\theta)} \right) \quad (2.19)$$

$$ICTD = \tau_m(\theta) = \frac{2r_a}{c} \cdot \sin\left(\frac{\phi_0}{2}\right) \cdot \sin\left(\frac{\phi_0}{2} - \theta\right) \quad (2.20)$$

with  $\phi_0 = \phi_{m+1} - \phi_m$ . These two parameters define the angular position of a phantom source in between two loudspeakers. It is easy to demonstrate from the second equation that an array radius  $r_a = 0$  (the case of coincident array) implies an ICTD=0. The delay is the function of the angles between the two adjacent microphones, the angle of incidence of the sound wave and the radius array. The maximum delay  $\tau_{max}$  is only dependent on the angle between the two loudspeakers, and is obtained by the formula:

$$\tau_{max} = \frac{2r_a}{c} \cdot \sin^2\left(\frac{\phi_0}{2}\right) \quad (2.21)$$

Depending on the different Panning method used, the ratio between the two gains, or directivity patterns, changes. In the case of ICLD, this is equal to the following expression, and it is possible to define the gains as polar patterns since a one-to-one correspondence is considered between the loudspeakers and the microphones of the array.

$$ICLD = \Phi = \frac{\Gamma_{m+1}(\theta)}{\Gamma_m(\theta)} = \frac{\sin(\theta - \phi_m)}{\sin(\phi_m - \theta)} \quad (2.22)$$

The aim of this method, which can be easily brought back to the tangent panning law, renders a more defined phantom source, but the introduction of TD permits to have a more natural and realistic sound, together with the need for less frequency selective microphones, which are easier to build.

As said in section 2.4, the use of TID methods implies the use of near-coincident arrays and psychoacoustic curves as the ones of Franssen or Williams. In this case, since the ID is mitigated by the use of TD, the ratio between the two gains is corrected by a factor  $\beta \geq 0$ , which takes account of this. The formula 2.22 becomes:

$$ICLD = \Phi = \frac{\Gamma_{m+1}(\theta)}{\Gamma_m(\theta)} = \frac{\sin(\theta - (\phi_m - \beta))}{\sin((\phi_{m+1} + \beta) - \theta)} \quad (2.23)$$

$$\beta = \arctg\left(\frac{\eta \sin(\phi_{m+1} - \phi_m)}{1 - \eta \cos(\phi_{m+1} - \phi_m)}\right) \quad (2.24)$$

with  $\eta = 10^{\frac{\eta_{dB}}{20}}$ , corresponding to the converted value of the ICLD from the one read on the psychoacoustic curve, which has a value in dB. The maximum delay  $\tau$  considered in the curves is 1 ms, because over 3 ms the pattern of the microphones became omnidirectional.

A value  $\beta = 0$  is a particular case of ID method, resulting in an  $\eta = 10.40$ .

The value of  $\eta$  is given by the following equation, obtained as a generalised logistic approximation of Williams curve in dB :

$$\eta[dB] = ICLD_{max} = 221.5913 - \frac{230.1794}{1.0 + e^{-(1000 \cdot \tau_{max} + 2.1786)}} \quad (2.25)$$

Starting from the angles of the loudspeakers and incidence of the planewave, the delay was found. This value can be used to obtain the ICLD, ratio between

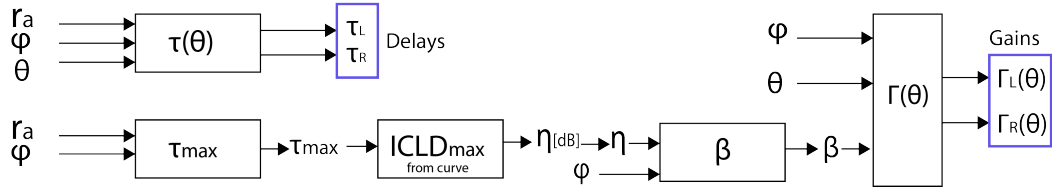
the microphones polar pattern.

The ratio gives infinite solutions, but, at the start, the constraint 2.18 was used, that, with the other constraint of selectivity  $\Gamma(\theta) = 0$  if  $\theta \notin [\phi_m, \phi_{m+1}]$  results in the following final equation for the polar pattern:

$$\Gamma(\theta) = \begin{cases} [1 + \frac{\sin^2(\theta+\beta)}{\sin^2((\phi_0+\beta)-\theta)}]^{-\frac{1}{2}} & \text{if } 0 \leq \theta \leq \phi_0 \\ [1 + \frac{\sin^2(\theta+\beta)}{\sin^2((\phi_0+\beta)-\theta)}]^{-\frac{1}{2}} & \text{if } -\phi_0 \leq \theta \leq 0 \\ 0 & \text{elsewhere} \end{cases} \quad (2.26)$$

This polar pattern can not be implemented with commercial microphones, but can be well approximated by microphones with order  $N \geq 2$ . The orders of the microphones are investigated in section 2.9.1 dedicated to the theory of Ambisonics.

To conclude, a new approach for the multichannel reproduction system has been presented, with only two free parameters, the radius of the array, proposed at 15.5 cm by Johnston for a more natural sound, and the angle between the loudspeakers. This system will be the starting point for the extension of the study to the vertical plane, presented in the next chapter.



**Figure 2.27:** PSR Scheme



### 2.6.3 MAX-MSP Implementation

A MAX-MSP implementation of the PSR method has been realized for a system of two loudspeakers. The system receives the azimuth angle for both the array and the sound source distance from the center (considering the median point as  $0^\circ$ , the maximum angle  $\theta$  in the extreme left point  $\theta = \frac{\phi}{2}$  and the minimum angle in the extreme right point  $\theta = -\frac{\phi}{2}$ ). The patch receives also the radius of the array, useful to calculate the delays, and the sound sample to reproduce.

The system calculates, starting from these parameters, the correspondents delays and gains (as polar pattern of the microphones) for the two loudspeakers, giving the illusion of a phantom source in the indicated position.

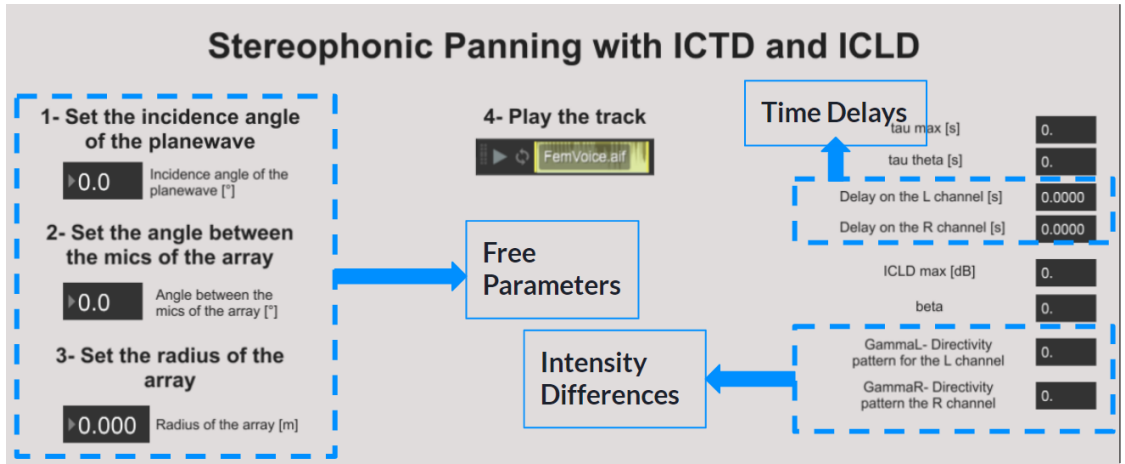


Figure 2.28: MAX MSP Patch for PSR Implementation

## 2.7 Panning Techniques for Vertical Plane

The panning methods illustrated in the previous section are all **pantophonic techniques**, with loudspeakers positioned only on the horizontal axis.

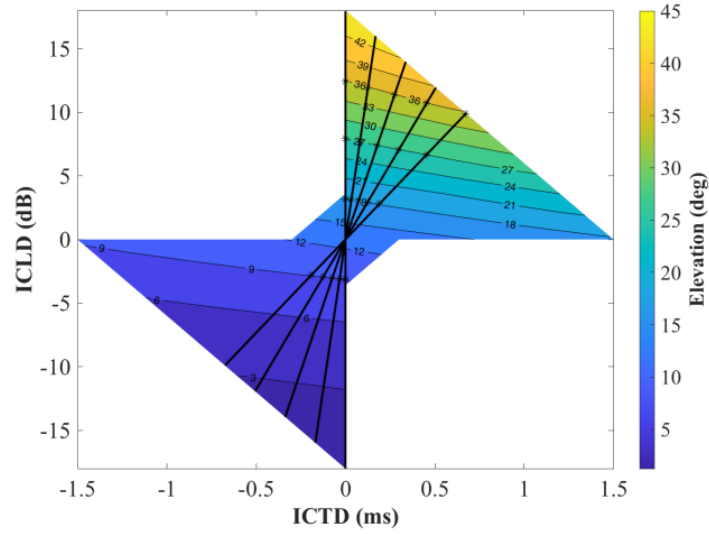
In order to render a source in a 3D space, it is still possible to use this type of approach, filtering the sound material with the HRTF of the listener, but an alternative is to use a **perophonic model**, with loudspeakers positioned on a full sphere array[30]. The presence of height loudspeakers is not only motivated by the vertical spatialization, but it has been demonstrated that this took also to a more natural perception of the sound [46]. In this section, some perophonic techniques are presented.

A path to follow for the vertical panning can be the one of recycling the already discussed ICLD and ICTD methods between loudspeakers positioned at a different height. Regarding this, an interesting ICLD method is VBAP, proposed by Ville Pulkki in [18] and one in which both ICLD and ICTD are used, that has been proposed in [47].

It is possible to pan sources with time differences, thanks to the **vertical precedence effect** proposed by Tregonning and Martin in [48], but the results of this perceptual experiment show a maximal delay around 5-10 ms, which is really higher compared to the 1 ms max for the horizontal plane, even if these high delays increase vertical localisation spread.

Another effect shown in [47], which can occur also in vertical panning, is the comb filtering. This can happen if the ears and the vertically arranged loudspeakers are not equidistant or some head movements are present, but also if TDs are introduced in the panning method, as in the horizontal plane. Since it is impossible to think that a listener can keep his head in a stationary position, the comb-filtering effect is not avoidable, so the introduction of TD can be a good path to follow if this introduces a wider sweet spot and a more natural sound as demonstrated with PSR method.

An example of this method is the one proposed in [47] which expresses TD and ID on a psychoacoustic curve similar to the ones of Williams and Fransen, but for two dimensions.



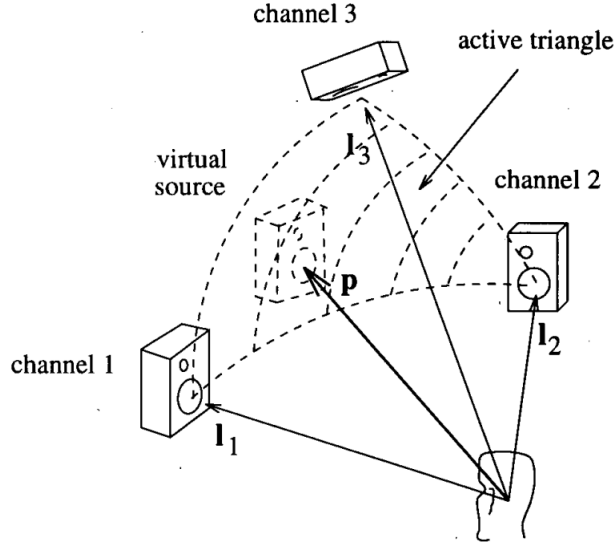
**Figure 2.29:** Psychoacoustic Curves for Horizontal and Vertical Panning. Image courtesy of [47]

## 2.8 VBAP- Vector Based Amplitude Panning

Vector Base Amplitude Panning (VBAP) method is a perophonic technique proposed by Ville Pulkki in [30] in 1997 for the positioning of a virtual sound source in a 3D space, and is based on a vector reformulation of the Tangent Panning Law for 3D. This method, as for the PSR, is useful because allows us to build a multichannel sound system with an arbitrary number of channels, in order to create a two or three dimensions sound field.

VBAP method is an amplitude panning method in which the formula for the sound positioning is obtained by vector and vector bases. In the work both 2D and 3D rendering is considered, but, for the aims of this work, only the second one is considered, since it is the one used in the experiment presented in the Chapter 4.

Starting from the classic stereophonic configuration, the method is extended with the addition of a third loudspeaker located in another arbitrary position with the only boundaries to be equidistant to the listener, as the others, and not to be placed at the same height as the other two. This configuration results in a triangle, with the virtual source positioned inside this triangle, or rather, inside a sphere whose radius is correspondent, for the three loudspeakers, to their own distance from the listener's head. This region is called **active triangle**.



**Figure 2.30:** Active Triangle Concept on VBAP. Image courtesy of [30]

The first condition is similar to the 2.18, for the equal loudness of the loudspeakers in a central point of the triangle. This results in:

$$g_1^2 + g_2^2 + g_3^2 = C \quad (2.27)$$

with the sum of the square gains equal to an arbitrary constant  $C$ .

Let us consider now the three dimensional vectors that goes from the listener to each loudspeaker, with the left, right and central channel numbered, respectively, as 1,2 and 3. It is also defined a matrix  $L_{123}$ , made by the three vectors:

$$\begin{aligned} l_1 &= [l_{11}, l_{12}, l_{13}]^T \\ l_2 &= [l_{21}, l_{22}, l_{23}]^T \\ l_3 &= [l_{31}, l_{32}, l_{33}]^T \\ L_{123} &= [l_1, l_2, l_3] \end{aligned} \quad (2.28)$$

The direction components of the virtual source are expressed by the vector  $p$ , as:

$$p = [p_1, p_2, p_3]^T \quad (2.29)$$

The vector  $p$  can be expressed as a linear combination of the directional vector  $l_n$ , as:

$$\begin{aligned} p &= g_1 l_1 + g_2 l_2 + g_3 l_3 \\ p^T &= g L_{123} \end{aligned} \quad (2.30)$$

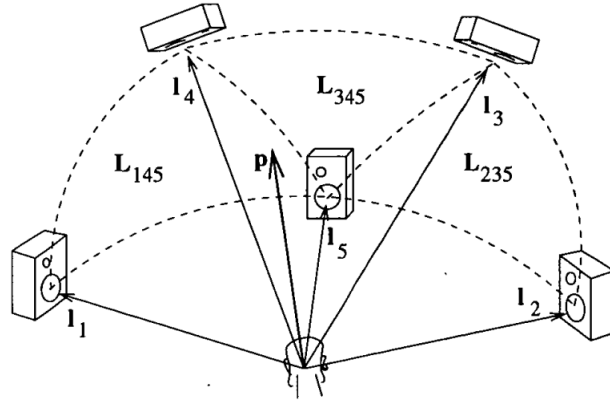
with  $g$  correspondent to the vector of the gains. If the vector  $L_{123}^{-1}$  exists, which is true if  $L_{123}$  is a base for a three-dimensional space, the vector  $g$  can be calculated as:

$$g = p^T L_{123}^{-1} = [p_1, p_2, p_3] \begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \quad (2.31)$$

It is verified in the work that, in the case of a sound source located on the line between the two loudspeakers, the same gains are equivalent if calculated with the tangent panning law. Finally, the components of  $g$  can be used as gains after being scaled, for keeping the total power constant, by a factor:

$$g_{scaled} = \frac{\sqrt{C}g}{\sqrt{g_1^2 + g_2^2 + g_3^2}} \quad (2.32)$$

The addition of more loudspeakers to this 3 element configuration make the number of active triangles in the space grow, and, as a consequence, the space in which the phantom source can be placed. Every triangle forms a base  $L_{n,m,k}$  (with  $m, n$  and  $k$  the label of each loudspeaker) and every loudspeaker can be part of multiple bases. Even if the position of the loudspeakers in VBAP is free enough, the only strict conditions are not to intersect the active triangles and to place the loudspeakers in the best possible configuration to optimize the space.



**Figure 2.31:** Example of Implementation of VBAP. Image courtesy of [30]

In summary, VBAP allows the positioning of a sound source inside a sphere detected by loudspeakers triangles with the same distance to the listener head. The maximum number of active loudspeakers for rendering a source is three, so this

allows to simplify the panning compared to methods such as HOA. In addition, the loudspeakers triplet is uniquely determined by the position of the sound source, thanks to the vector expression.

The source can be placed only inside one of the triangles, and smaller the triangle is the smaller the localisation error, with the con of a higher number of loudspeakers to keep the same dimension of the array.

The main properties of VBAP are three and are listed below:

1. If the source is located at the same azimuth and elevation of a loudspeaker, only that loudspeaker will reproduce the sound. This provides maximum sharpness of the virtual source, but, as con, if the source is placed in between more than one source, it will have a larger spread [27].
2. If the source is positioned in a line between two loudspeakers, the panning follows the tangent panning law between these two, with the gain of the third loudspeaker equals to zero.

To demonstrate this, we can consider the case of two loudspeakers, 1 and 2, writing the formula 2.24 for two channels as:

$$g = p^T L_{12}^{-1} = [p_1 p_2] \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{bmatrix}^{-1} \quad (2.33)$$

The term  $L_{12}^{-1}$  can be rewritten as:

$$L_{12}^{-1} = [p_1 p_2] \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{bmatrix}^{-1} = \frac{1}{l_{11}l_{22} - l_{21}l_{12}} \cdot \begin{bmatrix} l_{22} & -l_{12} \\ -l_{21} & l_{11} \end{bmatrix} \quad (2.34)$$

The values of the terms of the matrix are:  $l_{11} = l_{21} = \cos(\phi_0)$ ,  $l_{12} = -l_{22} = \sin(\phi_0)$  and the direction of the virtual source are:  $p_1 = \cos(\theta)$  and  $p_2 = \sin(\theta)$ .

The formula (2.27) can be rewritten using these consideration, as:

$$g = \frac{1}{l_{11}l_{22} - l_{21}l_{12}} \cdot [p_1 l_{22} - p_2 l_{21}, p_2 l_{11} - p_1 l_{12}] \quad (2.35)$$

Separating the component  $g_1$  and  $g_2$  of  $g$ , is obtained:

$$g_1 = \frac{1}{l_{11}l_{22} - l_{21}l_{12}} \cdot [p_1 l_{22} - p_2 l_{21}] = \frac{\cos(\theta)\sin(\phi) + \sin(\theta)\cos(\phi)}{2\cos(\theta)\sin(\theta)} \quad (2.36)$$

$$g_2 = \frac{1}{l_{11}l_{22} - l_{21}l_{12}} \cdot [p_2 l_{11} - p_1 l_{12}] = \frac{\cos(\theta)\sin(\phi) - \sin(\theta)\cos(\phi)}{2\cos(\theta)\sin(\theta)} \quad (2.37)$$

The relation  $\frac{g_1 - g_2}{g_1 + g_2}$  can be rewritten using the values of the gains from 2.29 and 2.30, as:

$$\frac{g_1 - g_2}{g_1 + g_2} = \frac{2\sin(\theta)\cos(\phi)}{2\cos(\theta)\sin(\phi)} = \frac{\tan(\theta)}{\tan(\phi)} \quad (2.38)$$

This demonstrates that the panning law used on VBAP for two loudspeakers is equivalent to the tangent panning law.

3. If the source is located on the center of the triangle, the gains are the same for all the loudspeakers of the base. This is possible thanks to the assumption made in the equation 2.27.

Further experiments have overcome the idea of the fixed position for the loudspeakers, a really important constraint for practical applications. This is possible with methods as DBAP [49] (Distance Based Amplitude Panning) which remove this constraint without losing performances [50].

### 2.8.1 MAX MSP Implementation of VBAP

A possible implementation of this method is realized for the software MAX MSP (6.1 minimum version required) by Nathan Wolek and is available for free in the related GitHub Repository [51]. This patch allows both the positioning on a 2D or 3D array of loudspeakers, by indicating the azimuth and elevation of every loudspeaker (paying attention to indicate at first the lower plane ones and then the higher plane ones) and of the sound source.

It is also possible to define the other two parameters: gain and spread. The first one indicates a value equal to 0 or 1, for the gain of the system, while the second indicates how the sound is well-sharped in a loudspeaker.

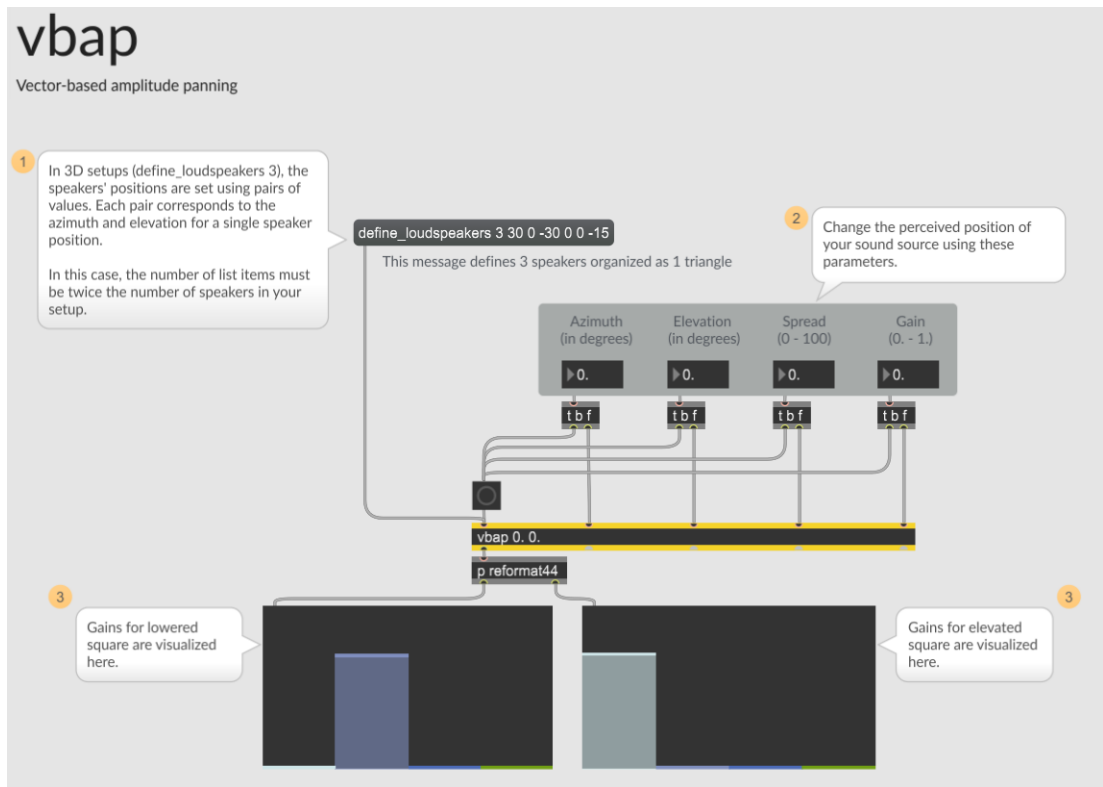


Figure 2.32: MAX MSP Implementation of VBAP

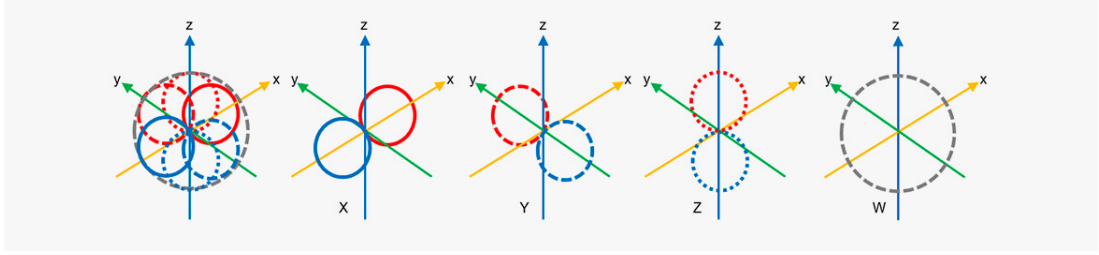
## 2.9 Other Options

### 2.9.1 Ambisonics

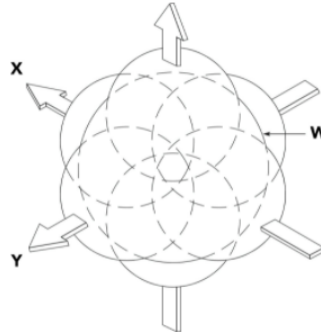
Ambisonics is a 3D surround technique introduced in the 70s by the British National Research Development Corporation, with the goal of rendering a better perception of the auditory scene with respect to classic stereophony and to surround formats (like 5.1 and 7.1), simulating the sound field in a certain position inside a room.

In the ambisonics theory [52] [53] [54], the sound source is encoded into 4 components of width, depth, height and omnidirectional pressure. X, Y and Z represents the 3D dimension of width, depth and height, and are recorded in a similar way to three 8-figure microphones. The fourth parameter, W, represents the omnidirectional pressure component, with a spherical pattern, similar to an omnidirectional microphone.





**Figure 2.33:** Ambisonics X Y Z W Signals. Image courtesy of [55]



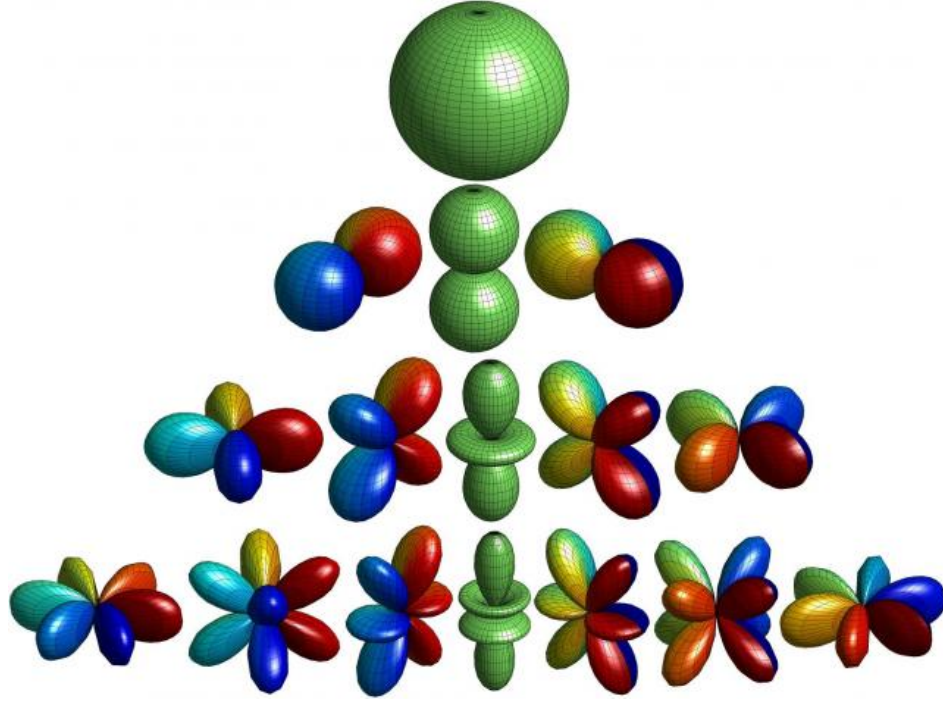
**Figure 2.34:** B-Format Encoding System. Image courtesy of [52]

The order of the microphone defines also the type of information decoded. This information divides the microphones into:

- **Order 0:** The decoded signals consist only of the pressure of the recorded soundfield. This parameter corresponds to the recorded parameter W.
- **Order 1 (FOA, First Order Ambisonics):** The decoded signals contain information about the pressure gradient. These parameters correspond to the X, Y, Z parameter, velocity of the air particles moved by a sound event in the three dimensions [43].
- **Order  $\geq 2$  (HOA, High Order Ambisonics):** The decoded signals contains information about the pressure derived from the recorded soundfield, corresponding to the acceleration of the sound.

For an order N,  $(N + 1)^2$  signals are generated, resulting in a better directivity of the microphone with the increasing of this parameter. On the other side, more

signals implies a higher computational power needed, a particular design if a real microphone is used, and an increasing number of loudspeakers used for the reproduction [54].



**Figure 2.35:** Ambisonics Directivity Patterns in Function of the Order. Image courtesy of [56]

Different coding techniques leads to different encoding formats:

- **A-Format:** The A-format encoding is realised with four sub-cardioid capsules located on the surface of a tetrahedron. The name of the four channels are LF, LR, RF and RB (Left and Right, Forward and Back). All the capsules record with the same gain level.
- **B-Format:** This type of encoding is realized starting with the A-Format, applying the following mathematical equations:

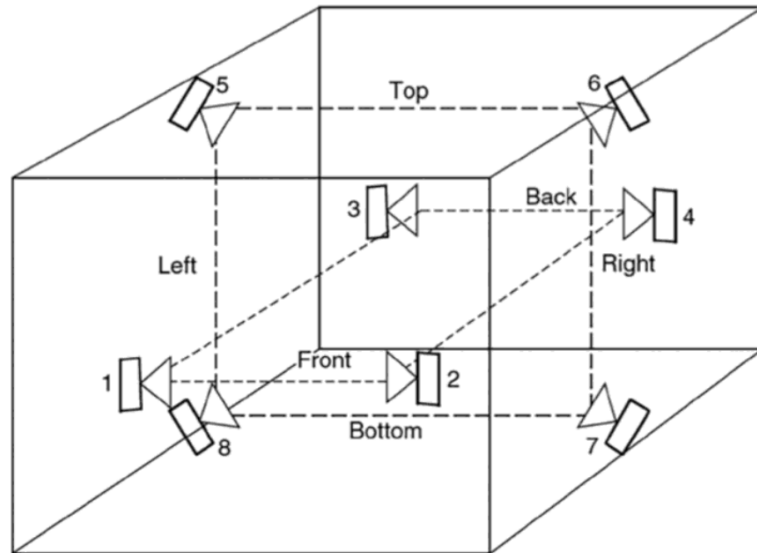
$$\begin{aligned} X &= 0.5((LF - LB) + (RF - RB)) \\ Y &= 0.5((LF - RB) - (RF - LB)) \\ Z &= 0.5((LF - LB) + (RB - RF)) \\ W &= 0.5(LF + LB + RF + RB) \end{aligned}$$

- **C-Format:** Format used to reduce the number of channels of the B-Format, usually from four to two. Since the transmission on four channels is now technologically executable in an easy way, this format is rarely used anymore.
- **D-Format:** B-Format signal ready for the reproduction on different system as 5.1, 7.1 or binaural.

The most common format is the B-format, used in the First Order Ambisonics (FOA) technique. The best characteristic is the independence from the reproduction system used. This technique uses order 0 microphones to capture the directional cues of the sound and order 1 microphones to capture the information from the propagation space.

Ambisonics signals can be recorded via particular microphones, presented in the section 2.10.1, which use the physical reconstruction method to render the soundfield in a certain point of the room, or can be encoded in terms of spherical harmonics of far and near field.

The reproduction system is realised with the use of 8 channel minimum, positioned in a configuration as the one below:

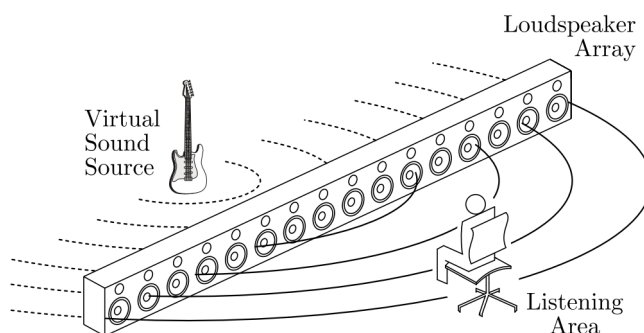


**Figure 2.36:** Ambisonics Reproduction Configuration. Image courtesy of [43]

## 2.9.2 Wave-Field Synthesis - WFS

The Wave-Field Synthesis (WFS) [57] [58] is a full-sphere panning method based on the Kirchhoff-Helmholtz integral, which explains the Huygens principle. In the Huygens principle, every soundwave can be described as the contribute of secondary sources.

The aim of the technique is to render one or more virtual sound sources on a high number of loudspeakers positioned in a volume around the listener. The idea is to render the wavefronts from the volume of the loudspeakers, with every loudspeaker which acts as a secondary source for the production of the wavelet as the sources are created behind the array.



**Figure 2.37:** WFS Loudspeakers Disposition. Image courtesy of [59]

This configuration presents an interesting theoretical base, with the main advantage of the wider sweetspot, but it comes with some practical limitations if implemented.

- The disposition, as seen in the figure ??, is made by a high number of channels which have to be fed, resulting in a higher computational cost. In addition, the great number of different loudspeakers has to be fed with the same number of different signals, which have to be recorded or generated and then stored.
- The perfect reconstruction of the wavefront is possible only if the size of the loudspeakers is less than half the wavelength of the maximum frequency reproduced. In this case, for 20 kHz, this would require a loudspeaker the size of 8.6 mm. If this rule is not applied, distortion can degrade the spatial accuracy of the reconstruction, and this is one of the main topics of research for WFS improvement.

- Another con is the fact that these loudspeakers, to respect the Huygens-Fresnel principle, have to emit continuously, and this is not possible practically. In addition the loudspeakers have to change direction in the quietest way possible.
- The theory of WFS implies that the only sound present is the one reproduced by the loudspeakers. This results in a listening room with no reflection, because each reflection slopes the accuracy of the reconstruction. The only suitable reproduction room is, then, the anechoic one.

One of the implementations of this technology is installed at the University of Technology of Berlin and includes 500 independent Loudspeakers.



**Figure 2.38:** WFS Implementation at the University of Technology of Berlin. Image courtesy of [60]

## 2.10 Microphone Arrays for Multichannel

In this section some microphone arrays are presented, those for multichannel recording, with a mix between physical reconstruction methods, made by pre-built microphones with more than one capsule, and particular systems based on cardioids, supercardioids and hypercardioids, arranged in particular tree structures.

### 2.10.1 Physical Reconstruction Methods: Soundfield and Eigenmike

#### SoundField

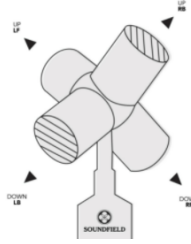
**SoundField microphone**[61] can be considered the relative of the Blumlein theory. It is based on the Ambisonics theory [52], and for this reason has 4 capsules positioned in a tetrahedron surface. The recording made by the microphone is in A-Format, and then this is converted into B-format, to obtain X, Y, Z, W.

The channels are named, based on the position, LF, LR, RF, RB (Left and Right Forward and Back). The capsules are placed as close as possible to avoid phase effect, typical of the multi-microphones configuration, and the really close distance is also compensated for via software.

Particular processing, of which the mathematical formulation has been shown in the previous sections, transforms every A-Format signal into a B-Format [62].

The elaboration of this signal is made by a particular system available in analog or as a plugin (as the case of SPS200). The software, called SPS200 Surround Zone allows to define the output format (Stereo, 5.1, etc.), the polar pattern and other useful parameters.

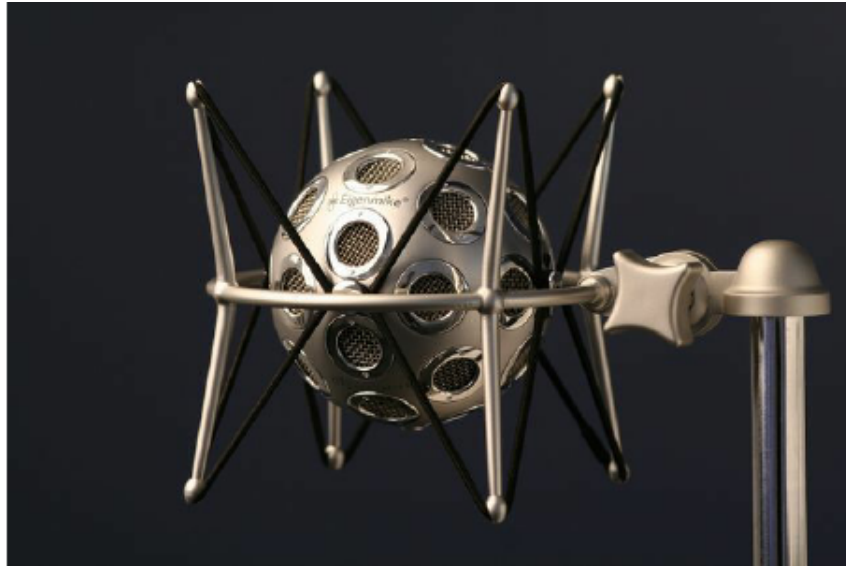
This system allows a great 3D multichannel, but it can be used also for stereo recordings. The microphone has a frequency response from 40Hz to 20KHz, a Maximum SPL of 130 dB and is powered by a 48 V phantom supply.



**Figure 2.39:** SoundField Capsules. Image courtesy of [61]

### **Eigenmike em32**

The Eigenmike em32 system[63], by mh acoustic, is realized with 32 condenser capsules positioned on a sphere of 8 cm. Inside the body of the microphone, 32 A/D/A 24-bit converters are located, which communicate with a CAT5 protocol. The signal of every capsule is combined to create a set of Eigenbeams, called also High Order Ambisonics (HOA) signals, with the number of elements of the set corresponding to the user-determine beam-form, up to 4. The Eigenbeams are then combined to steer multiple simultaneous beam-patterns, that can be focused to specific directions in the acoustic field. The process of eigenbeamforming allows to position the soundfield in the desired direction, with the use of a particular plugin called EigenUnits.



**Figure 2.40:** Eigenmike em32 Microphone. Image courtesy of [64]



### 2.10.2 Tree-Structure Arrays

As proposed in [27], these arrays can be grouped into two main families: **five-channel main microphone** techniques and **front-rear separation** techniques. The first family is characterized by five channels which record the signals of the previously seen 5.1, the second uses separate arrays for the direct and the ambience sound field. Not always the one-to-one correspondence between loudspeakers and microphones is respected, so the signal has to be mixed separately after the recording.

#### INA-5

This five-channel main microphones system is based on five cardioids, a central one, two located at  $\pm 90^\circ$  in the Left and Right position and two LS and LR channel at  $\pm 150^\circ$ .

It is possible to recognize a Decca Tree configuration between C,L and R, while the LS and RS are two channel surround for the back and the ambience.

#### OCT

This five-channel main microphones system is made with a central cardioid, two L and R supercardioid (sometimes replaced with Omnidirectionals filtered with a low pass filter, to improve the answer at low frequencies) at  $\pm 90^\circ$  and two cardioid LS and LR at  $\pm 150^\circ$ .

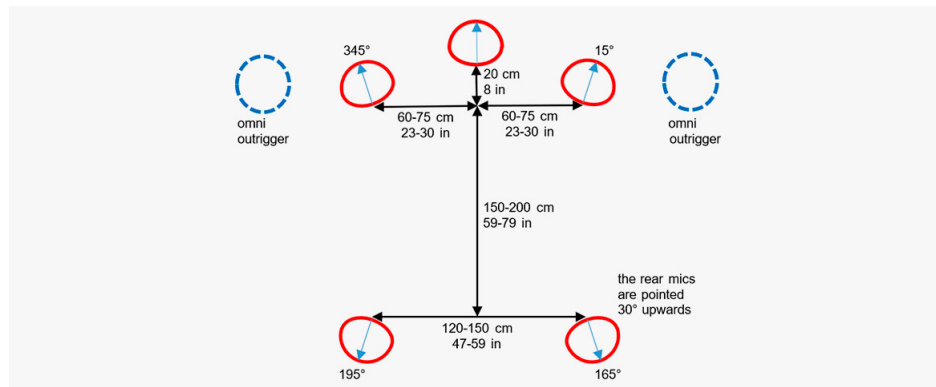


Figure 2.41: OCT. Image courtesy of [55]



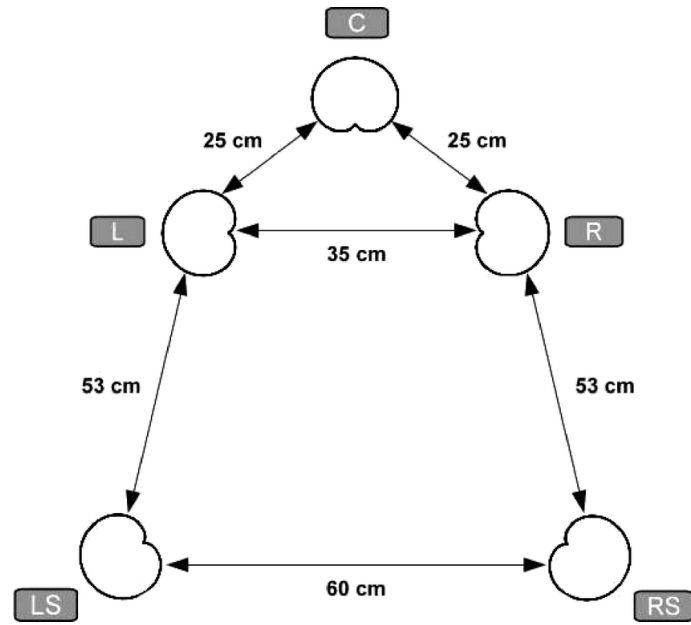


Figure 2.42: INA-5. Image courtesy of [55]

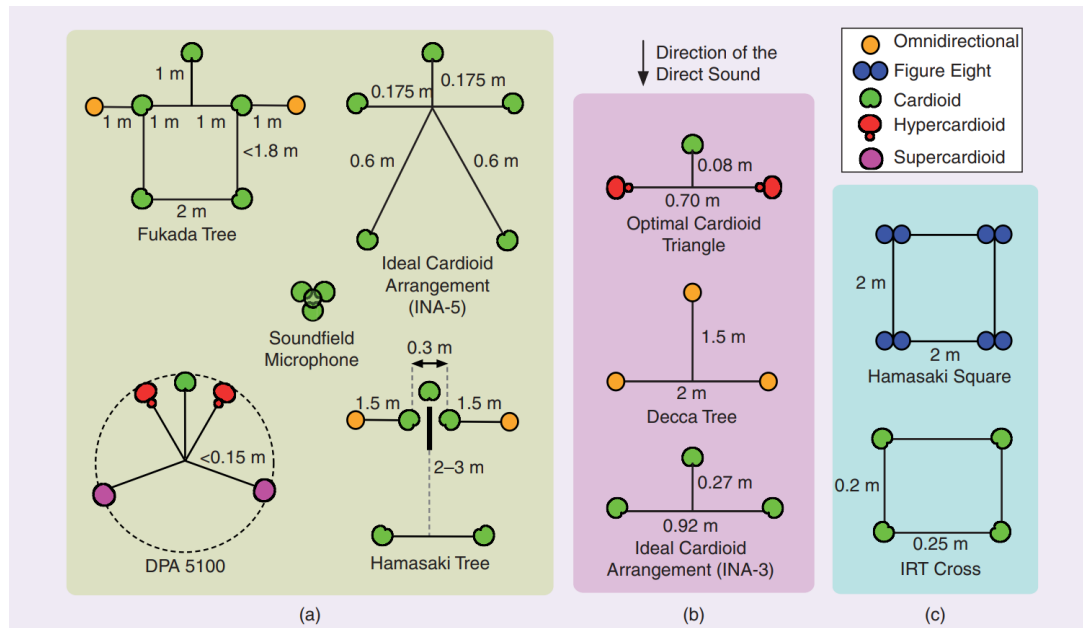


Figure 2.43: Multichannel Arrays Overview. Image courtesy of [27]

## **Other Multichannel Arrays**

The complete panoramic of the multichannel microphone arrays is provided by the image 2.43, from [27] with the already considered main microphone systems and some examples of frontal scene (b), including the Decca Tree seen in the section 2.5, and rear-only configuration (c).

### **2.10.3 Arrays for a Higher Number of Channels**

microphone arrays for a higher number of channels (as 10.2 or 22.2) are yet experimental and not standardised, even if they are described on the in the ITU-R BS.2159-4 report [41].

## **2.11 Chapter Conclusions**

In this background chapter some human hearing psychoacoustics theories and applications are presented.

The human auditory system distinguishes the location of sound events thanks to binaural mechanisms (which involves TID and TIL) and monaural mechanism (as HRTF). These cues allows the listener to localise the sound in terms of azimuth, elevation and distance, which are the main parameters to determine the position.

The application of these concepts allows to position a sound (phantom source) on the auditory scene, thanks to what is called Summing localisation principle, which uses ICTD and ICLD, sometimes together.

Examples of these methods are presented, with 2 or more loudspeakers, positioned in a single plane or with different elevations, to render a vertical spatialization. Other models which use HRTF are available for rendering height illusion, even without the need for a second layer of loudspeakers on a different elevation. These methods are called ear-signal based.

The signal received by the loudspeakers can be prerecorded with particular microphone arrays, or rendered with psychoacoustic methods. In the chapter some examples of both are presented, with particular attention to the second route with PSR (Perceptual Sound Field Reconstruction), used as base for the Hybrid method obtained in chapter 3, and VBAP, used as comparison for the Experimental part in chapter 4.

## Chapter 3

# Hybrid Time Amplitude Approach

### 3.1 Why Use Time Differences?

The synthetic stereophony methods, realised with perceptual assumptions, are mainly based on ID panning, since the literature suggests the most stable stereophonic image, while discrediting the use of TD for the introduction of artifacts as tonal coloration for comb-filtering effects. In addition, the direction of the sound source is not always easily controlled in the presence of time delays [27]. All the previous cons are true, but the introduction of TD could make the subject perceive a more natural sound, thanks to the casual effect of comb-filtering.

In the reference work [2] this idea has been confirmed, even if the cons remains true for delays which are not carefully chosen. The experiments demonstrated that:

1. There is no perceptual difference between ID and TID methods on the center in terms of localization.
2. For a back and lateral position, ID method work better in a general way, but degrades more moving away from the center in respect to TID.
3. In the same way, certainty of the position (also called locatedness) is better in the center for ID methods, but degrades faster in an off-center position. This, together with consideration 2, results in a smaller sweetspot for ID compared to TID method.

Even if the debate is still open, this demonstrates that, if an accurate attention is paid to the considerations of psychoacoustics, it is possible to obtain a better

and stable auditory image with the use of time intensity methods.

In any case, the delays are useful only on the horizontal plane thanks to the binaural cues, with which human auditory system perceives sound, and especially thanks to the position of the ears, which does not allow the same for the vertical plane.

## 3.2 Explanation of the Hybrid Approach

### 3.2.1 Introduction

The aim of this section is to extend the panning method from a system of two loudspeakers, located on the horizontal plane, to a system of three loudspeakers, with the third one located in between the previous two, but on a higher point on the vertical plane, forming a triangle. All the three loudspeakers are equidistant from the listener.

A Time Amplitude Panning method (PSR presented in the section 2.6) has been used for the horizontal plane, while an Amplitude Panning method is employed for the vertical plane, based on the tangent panning law.

### 3.2.2 Horizontal Panning

As shown in the section 2.6, the model has three free parameters: the radius of the array  $r_a$ , the angle between the microphones (or loudspeakers) of the array  $\phi$  and the angle of incidence of the planewave  $\theta$ , with  $-\frac{\phi}{2} \leq \theta \leq +\frac{\phi}{2}$ . Let us assume, as central point,  $\theta = 0$ , the most left point  $\theta = \frac{\phi_0}{2}$  and the most right point  $\theta = -\frac{\phi_0}{2}$ .

The delays and the intensity difference of the two channels are calculated from these three terms, for the final goal of obtaining the polar diagram of the microphones, correspondent to the gains of the channels. The formulas are as following, reported in the table to be easily remembered by the reader:

Delays	$ICTD = \tau_m(\theta) = \frac{2r_a}{c} \cdot \sin(\frac{\phi_0}{2}) \cdot \sin(\frac{\phi_0}{2} - \theta)$
Gains	$\Gamma(\theta) = \begin{cases} \left[1 + \frac{\sin^2(\theta+\beta)}{\sin^2((\phi_0+\beta)-\theta)}\right]^{-\frac{1}{2}} & \text{if } 0 \leq \theta \leq \phi_0 \\ \left[1 + \frac{\sin^2(\theta+\beta)}{\sin^2((\phi_0+\beta)-\theta)}\right]^{-\frac{1}{2}} & \text{if } -\phi_0 \leq \theta \leq 0 \\ 0 & elsewhere \end{cases}$

From now on, let us define the gains of the two loudspeakers as  $\Gamma(\theta_L) = g_1(\theta)$  and  $\Gamma(\theta_R) = g_2(\theta)$ .

### 3.3 Extension to the Vertical Panning

#### 3.3.1 First Faulty Approach: the $\alpha$ Parameter

A first faulty approach has been followed for the research of the vertical extension of the hybrid method. This is reported as following.

Considering the gains which come from PSR named as  $\hat{g}_1$  and  $\hat{g}_2$ , and using the assumption that, for every change of the elevation, these two gains will change by the same amount  $\alpha$ , the following factorization of the new gains,  $g_1$  and  $g_2$ , can be defined as:

$$\begin{cases} g_1 = \alpha \hat{g}_1 \\ g_2 = \alpha \hat{g}_2 \end{cases} \quad (3.1)$$

The second condition is the equal loudness of the loudspeakers in the center of the array, which, for monodimensional and bidimensional loudspeaker, is respectively equal to:

$$\begin{aligned} \hat{g}_1^2 + \hat{g}_2^2 &= 1 \\ g_1^2 + g_2^2 + g_3^2 &= 1 \end{aligned} \quad (3.2)$$

Combining 3.1 with the second equation of 3.2, is obtained:

$$\begin{aligned} \alpha^2 \hat{g}_1^2 + \alpha^2 \hat{g}_2^2 + g_3^2 &= 1 \\ \alpha^2 \cdot (\hat{g}_1^2 + \hat{g}_2^2) + g_3^2 &= 1 \end{aligned} \quad (3.3)$$

Considering the first equation of 3.1,  $\hat{g}_1^2 + \hat{g}_2^2 = 1$ , and replacing in 3.3:

$$\begin{aligned} \alpha^2 + g_3^2 &= 1 \\ \alpha^2 &= 1 - g_3^2 \end{aligned} \quad (3.4)$$

And finally obtaining the final parameter  $\alpha$

$$\alpha = \sqrt{1 - g_3^2} \quad (3.5)$$

The parameter  $\alpha$  defines the panning between horizontal and vertical directions, with the gain of  $g_3$  which varies between zero and one. A zero value corresponds to a source completely on the lower plane, while a one value corresponds to a source completely on the central upper loudspeaker.

The value of  $g_3$  is function, for the tangent panning law, of the angle  $\gamma$ , elevation of the upper loudspeaker in respect to the lower plane, of  $\omega$ , elevation of the sound-source, and of a certain gain  $g_k$  function of the azimuth angle of the soundsource.

**This configuration resulted faulty** for the assumption made on the lower loudspeakers, first equation of 3.2, which produced correct gains for the extreme vertical points correspondent to the ones where the loudspeakers were located, but not for points inside the triangle.

### 3.3.2 Panning for 3 Loudspeakers

#### Tangent Panning Law for the Vertical Plane

Let's define  $\gamma$  as the vertical angle formed by the head of the listener and the upper loudspeaker, and  $\varphi$  as the vertical angle of the phantom source. The zero is located in the central point between the loudspeakers, at  $\frac{\gamma}{2}$ . The goal is to find  $g_3$ , gain of the higher central loudspeakers in function of the angles and of  $g_k$ , gain of the virtual central loudspeaker, equidistant in the horizontal plane to the other two.

From the **Tangent Panning Law formula**:

$$\begin{aligned} \frac{\tan(\varphi)}{\tan(\gamma)} &= \frac{g_3 - g_k}{g_3 + g_k} \rightarrow \tan(\varphi) \cdot (g_3 + g_k) = \tan(\gamma) \cdot (g_3 - g_k) \rightarrow \\ g_3 \cdot \tan(\varphi) + g_k \cdot \tan(\varphi) &= g_3 \cdot \tan(\gamma) - g_k \cdot \tan(\gamma) \rightarrow \\ g_3 \cdot (\tan(\varphi) - \tan(\gamma)) &= -g_k \cdot (\tan(\gamma) + \tan(\varphi)) \rightarrow \end{aligned} \quad (3.6)$$

$$g_3 = \frac{g_k \cdot (\tan(\gamma) + \tan(\varphi))}{\tan(\gamma) - \tan(\varphi)}$$

#### Preliminary Assumptions

For a system of three loudspeakers, in an analog way as considered for two, the sum of the square of the gains of the loudspeakers is defined, in function of the horizontal and vertical incidence angles of the planewave, called  $\theta$  and  $\varphi$ , as constant. We can arbitrarily define the constant equal to 1, obtaining:

$$g_1^2(\theta, \varphi) + g_2^2(\theta, \varphi) + g_3^2(\theta, \varphi) = 1 \quad (3.7)$$

Considering factorization between the terms of the sum, every gain can be rewritten in the form:

$$\begin{aligned}
 g_1(\theta, \varphi) &= \overline{g_1}(\theta) \cdot \hat{g}_1(\varphi) \\
 g_2(\theta, \varphi) &= \overline{g_2}(\theta) \cdot \hat{g}_2(\varphi) \\
 g_3(\theta, \varphi) &= \overline{g_3}(\theta) \cdot \hat{g}_3(\varphi) = A \cdot \hat{g}_3(\varphi)
 \end{aligned} \tag{3.8}$$

In the last equation, it has been assumed  $\overline{g_3}(\theta)=A$ , since the gain of the central higher loudspeaker does not depends on the horizontal angle of incidence of the planewave  $\theta$ , so it is equal to a constant  $A$ . Another assumption is that  $\hat{g}_1(\varphi) = \hat{g}_2(\varphi)$ , since both gains of the loudspeakers in the lower plane change in the same way along the vertical direction.

Furthermore, in a certain point  $\theta_m, \varphi_m$ , for example  $\theta_m = 0, \varphi_m = 0$ , holds the equation:

$$g_1(\theta_m, \varphi_m) = g_2(\theta_m, \varphi_m) = g_3(\theta_m, \varphi_m) = \frac{1}{\sqrt{3}} \tag{3.9}$$

Let's define the function **Horizontal ICLD** as:

$$f(\theta) = \frac{\overline{g_2}(\theta)}{\overline{g_1}(\theta)} \rightarrow \overline{g_2}(\theta) = f(\theta) \cdot \overline{g_1}(\theta) \tag{3.10}$$

### How To Obtain the Gains

Starting from 3.7, the expression 3.8 can be substituted in this, obtaining:

$$\begin{aligned}
 g_1^2(\theta, \varphi) + g_2^2(\theta, \varphi) + g_3^2(\theta, \varphi) &= 1 \rightarrow \\
 (\overline{g_1}(\theta) \cdot \hat{g}_1(\varphi))^2 + (\overline{g_2}(\theta) \cdot \hat{g}_2(\varphi))^2 + A^2 \cdot \hat{g}_3^2(\varphi) &\rightarrow \\
 \overline{g_1}^2(\theta) \cdot \hat{g}_1^2(\varphi) + \overline{g_2}^2(\theta) \cdot \hat{g}_2^2(\varphi) + A^2 \cdot \hat{g}_3^2(\varphi) &
 \end{aligned} \tag{3.11}$$

Using [5] we can express  $\overline{g_2}(\theta)$  in function of  $\overline{g_1}(\theta)$  as:

$$\overline{g_1}^2(\theta) \cdot \hat{g}_1^2(\varphi) + f^2(\theta) \cdot \overline{g_1}^2(\theta) \cdot \hat{g}_2^2(\varphi) + A^2 \cdot \hat{g}_3^2(\varphi) = 1 \tag{3.12}$$

Now, let's substitute  $\hat{g}_3^2(\varphi)$  from 3.6, and the following result is obtained:

$$\begin{aligned}
 \overline{g_1}^2(\theta) \cdot \hat{g}_1^2(\varphi) + f^2(\theta) \cdot \overline{g_1}^2(\theta) \cdot \hat{g}_2^2(\varphi) + A^2 \cdot \left[ \frac{\hat{g}_1(\varphi) \cdot (\tan(\gamma) + \tan(\varphi))}{\tan(\gamma) - \tan(\varphi)} \right]^2 &= 1 \rightarrow \\
 \overline{g_1}^2(\theta) \cdot \hat{g}_1^2(\varphi) + f^2(\theta) \cdot \overline{g_1}^2(\theta) \cdot \hat{g}_2^2(\varphi) + A^2 \cdot \frac{\hat{g}_1^2(\varphi) \cdot [(\tan(\gamma) + \tan(\varphi))]^2}{(\tan(\gamma) - \tan(\varphi))^2} &= 1
 \end{aligned} \tag{3.13}$$

Let's define the function **Vertical ICLD** as  $\frac{\tan(\gamma)+\tan(\varphi)}{\tan(\gamma)-\tan(\varphi)} = h(\varphi)$  and substitute this expression in the equation:

$$\overline{g_1}^2(\theta) \cdot \hat{g}_1^2(\varphi) + f^2(\theta) \cdot \overline{g_1}^2(\theta) \cdot \hat{g}_2^2(\varphi) + A^2 \cdot \hat{g}_1^2(\varphi) \cdot h^2(\varphi) = 1 \quad (3.14)$$

Let's group together the terms in function of  $\hat{g}_1^2(\varphi)$ :

$$\begin{aligned} \hat{g}_1^2(\varphi) \cdot [\overline{g_1}^2(\theta) + f^2(\theta) \cdot \overline{g_1}^2(\theta) + A^2 \cdot h^2(\varphi)] &= 1 \\ \hat{g}_1^2(\varphi) \cdot \{\overline{g_1}^2(\theta) \cdot [1 + f^2(\theta)] + A^2 \cdot h^2(\varphi)\} &= 1 \end{aligned} \quad (3.15)$$

The goal now is to separate the terms function of  $\theta$  from the ones function of  $\varphi$ :

$$\overline{g_1}^2(\theta) \cdot [1 + f^2(\theta)] = \frac{1}{\hat{g}_1^2(\varphi)} - A^2 \cdot h^2(\varphi) = C \quad (3.16)$$

The left and right terms are dependent separately on  $\theta$  and  $\varphi$ , so, it can be imposed that each part of the equation is equal to a constant C, and rewriting the equation in function of the two gains,  $\overline{g_1}^2(\theta)$  and  $\hat{g}_1^2(\varphi)$ , is obtained:

$$\begin{aligned} \overline{g_1}^2(\theta) &= \frac{C}{1 + f^2(\theta)} \\ \hat{g}_1^2(\varphi) = \hat{g}_2^2(\varphi) &= \frac{1}{C + A^2 \cdot h^2(\varphi)} \end{aligned} \quad (3.17)$$

### Expressions of the Gains $\hat{g}_n(\varphi)$

Each gain  $\hat{g}_n(\varphi)$  can be expressed in the form:

$$\begin{aligned} \hat{g}_1^2(\varphi) &= \frac{1}{C + A^2 \cdot h^2(\varphi)} \\ \hat{g}_2^2(\varphi) &= \frac{1}{C + A^2 \cdot h^2(\varphi)} \\ \hat{g}_3^2(\varphi) &= \hat{g}_1^2(\varphi) \cdot h^2(\varphi) \end{aligned} \quad (3.18)$$



### Values of the Constants A and C

In order to find the values of the constants, the equation 3.9 for the central point can be considered, since this is the only one which refers to the point, with coordinates  $\theta_m, \varphi_m$ , where the value of the gains is determined.

Starting from the first equation of 3.8, the result can be extended also for the second one, because of  $\hat{g}_1(\varphi) = \hat{g}_2(\varphi)$ :

$$\begin{aligned} g_1(\theta_m, \varphi_m) &= \overline{g_1}(\theta_m) \cdot \hat{g}_1(\varphi_m) \\ g_1^2(\theta_m, \varphi_m) &= \overline{g_1}^2(\theta_m) \cdot \hat{g}_1^2(\varphi_m) = \frac{1}{3} \\ \frac{C}{1 + f^2(\theta_m)} \cdot \frac{1}{C + A^2 \cdot h^2(\varphi_m)} &= \frac{1}{3} \end{aligned} \quad (3.19)$$

In this point, both  $f(\theta_m)$  and  $h(\varphi_m)$  are equals to 1, so:

$$\begin{aligned} \frac{C}{2 \cdot (C + A^2)} &= \frac{1}{3} \\ C &= \frac{2 \cdot (C + A^2)}{3} \\ C &= 2A^2 \end{aligned} \quad (3.20)$$

From the third equation of 3.8, for the point with coordinates  $(\theta_m, \varphi_m)$ :

$$g_3^2(\theta, \varphi) = A^2 \cdot \hat{g}_3^2(\varphi) = \frac{1}{3} \quad (3.21)$$

Using the expression  $\hat{g}_3^2(\varphi) = \hat{g}_1^2(\varphi) \cdot h^2(\varphi)$ :

$$A^2 \cdot \hat{g}_1^2(\varphi) \cdot h^2(\varphi) = \frac{1}{3} \frac{A^2}{C + A^2 \cdot h^2(\varphi)} = \frac{1}{3} \quad (3.22)$$

Using 2.20:

$$\frac{3A^2}{3A^2} = 1 \quad (3.23)$$

This equation has infinite solutions, so this demonstrates that an arbitrary value of A can be chosen. Let's define A=1, and, consequently C=2.

### Values of the Gains $\overline{g}_n(\theta)$

The values of  $\overline{g}_1(\theta)$  and  $\overline{g}_2(\theta)$  can be calculated in an analog way from the first equation of 3.17, as:

$$\begin{aligned}\overline{g}_1(\theta) &= \sqrt{\frac{C}{1+f^2(\theta)}} \\ \overline{g}_2(\theta) &= \sqrt{\frac{f(\theta) \cdot C}{1+f^2(\theta)}} \\ \overline{g}_3(\theta) &= A\end{aligned}\tag{3.24}$$

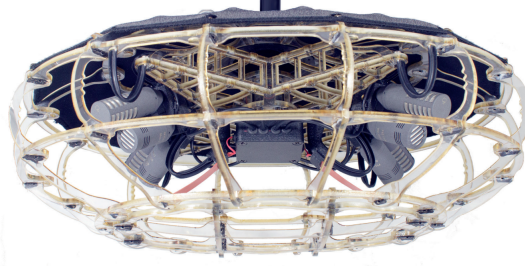
### 3.3.3 Final Values of the Gains $g_n(\theta, \varphi)$

After all these considerations, each gain  $g_n(\theta, \varphi) = \overline{g}_n(\theta) \cdot \hat{g}_n(\varphi)$  can be calculated for a two-dimensional array of three loudspeakers as:

$$\begin{aligned}g_1(\theta, \varphi) &= \overline{g}_1(\theta) \cdot \hat{g}_1(\varphi) = \frac{\sqrt{2}}{\sqrt{1+f^2(\theta)}} \cdot \frac{1}{\sqrt{2+h^2(\varphi)}} \\ g_2(\theta, \varphi) &= \overline{g}_2(\theta) \cdot \hat{g}_2(\varphi) = \overline{g}_2(\theta) \cdot \hat{g}_1(\varphi) = \frac{f(\theta) \cdot \sqrt{2}}{\sqrt{1+f^2(\theta)}} \cdot \frac{1}{\sqrt{2+h^2(\varphi)}} \\ g_3(\theta, \varphi) &= \overline{g}_3(\theta) \cdot \hat{g}_3(\varphi) = A \cdot \hat{g}_1(\varphi) \cdot h(\varphi) = \frac{h(\varphi)}{\sqrt{2+h^2(\varphi)}}\end{aligned}\tag{3.25}$$

### 3.4 ORTF3D Microphone Array

The Hybrid Time Amplitude approach is a perceptual method, but an analog implementation of the system, realised by Shoeps with a microphone array called ORTF 3D [65], exists.

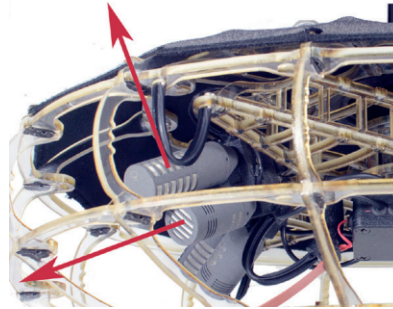


**Figure 3.1:** ORTF 3D. Image courtesy of [65]

As the name suggests, this array is an evolution of the 2-channel ORTF and of the 5.1 version called ORTF surround, and is based on eight supercardioids placed on two planes (upper and lower), which form rectangles with 10 and 20 cm length sides, and angles of  $80^\circ$  and  $100^\circ$ . The method used is ORTF because, in the two channel version, it is the one which provides a wide angle of stereophony ( $100^\circ$ ), and a good decorrelation between channels.

Since the microphones planes are placed one on top of the other, there is no distance between the capsules on the vertical plane, but only an angle of  $90^\circ$ , determining a X-Y configuration and a ID method on the vertical.

For the horizontal plane, on the other hand, the small distance between the capsules introduces delays with the intensity difference, in the same way as the PSR System, used on the Hybrid approach.



**Figure 3.2:** Capsules Orientation on the Vertical Plane. Image courtesy of [65]

Further studies in [65] demonstrated the suitability of the method for 3D Audio and VR application, with a good fit for format as Dolby Atmos and Auro3D.

### 3.5 Chapter Conclusions

In this chapter, the Hybrid method, the core of the thesis, has been illustrated, and its performances will be discussed and compared with VBAP in the next chapter.

Firstly, the advantage of Time Differences has been illustrated, remembering that these improve horizontal perception but not vertical one.

Then, starting from the PSR method, the formulation for the gains of the 3-loudspeakers Hybrid method are illustrated, based on a factorization (equation 3.8) and three main constraints:

1. The sum of the squares of the gains of the loudspeakers is always constant, and in particular, there is a point, with coordinates  $(\theta_m, \varphi_m)$  whose sum is equal to  $\frac{1}{\sqrt{3}}$ . This is also called equal loudness constraint.
2. Considering the central loudspeakers 3, if the source moves only in a horizontal direction, varying the azimuth, the gains of this loudspeakers resulted not affected. In other words, the azimuth contribute of the gain of this loudspeaker is equal to a constant A ( $\overline{g_3}(\theta)=A$ ).
3. Moving only on a vertical dimension, the gains of the lower loudspeakers change in the same way. So, the part of these two gains dependent on the vertical angle  $\varphi$  is the same for both the loudspeakers,  $\hat{g}_1(\varphi) = \hat{g}_2(\varphi)$ .

Solving the calculus, a closed formula for the three gains has been derived, function of the horizontal and vertical ICLDs,  $f(\theta)$  and  $h(\varphi)$ , respectively. Previously, the first faulty approach has been exposed, to show an alternative route followed in the research.

Hybrid method is a perceptual method realised thanks to psychoacoustics principles. Some similar physical implementation of the same principles, like the ORTF3D method presented in the section 3.4, are available on the market.

## Chapter 4

# Experiment Methodology and Setup

### 4.1 Introduction

The goal of the experiment has been to compare the performances of the Hybrid approach with the ones of the Vector Base Amplitude Panning (VBAP) method, in terms of Localization and certainty of the answer, also defined as Locatedness.

During the experiment, subjects answered the question “Where do you perceive the sound event?”, localizing the sound source generated from an array of three loudspeakers, in a triangle disposition, with a particular interface that captured the information of azimuth and elevation angle. In addition, the participants were asked to rate the degree of certainty of each answer, on a scale of continuous values from 0 (fully uncertain) to 100 (fully certain). For this second part, the subjects answer the question “How well you can assign a particular direction to the perceived source?” [66], “How certain are you of the direction of the perceived source?” [67]. An acoustic curtain with a numbered grid was located in front of the listener, to have a reference of the position, with the grid reproduced on the PC screen.

### 4.2 Experiment Setup

#### 4.2.1 Room

The experiment has been run in an acoustically isolated room (TB7 in the Teaching Block Building of the University of Surrey, image 4.1). The design of the room follows the ITU-R BS 1116 standard [68].



**Figure 4.1:** TB7 Room at the University of Surrey. Image courtesy of [69]

### 4.2.2 Loudspeakers

To playback the stimuli, an array formed by three active Genelec 8330A loudspeakers has been used, in an isosceles triangle disposition, with two external lower loudspeakers and a central one in a higher plane, 4.2. The loudspeakers were calibrated to a nominal level of 85 dBA, with the possibility of raising or lowering the volume given to the subjects.

Due to the equalization of the existing spherical system and the particular desired configuration of the loudspeakers, a custom loudspeakers triangle has been created.

The two lower loudspeakers were positioned at the same height of 1.21 m in respect of the distance from the floor to the emission centre, with a base angle of  $60^\circ$  and with a mutual distance of 2 m. The distance from the wall for both was the same, equal 87 cm, image 4.3.

The vertical angle  $\alpha$  between the ears of the listener and the upper loudspeakers was  $30^\circ$ , and the horizontal distance between the listener and the lower line between the two loudspeakers was chosen as 2 m, image 4.4.

From this information it is possible to obtain the position of the third loudspeaker in terms of height and horizontal distance from the listener. Considering a circumference with  $r=c=2$  m (distance between the listener and the vertical loudspeaker) the equation of the sphere can be written as:

$$x^2 + y^2 + z^2 = c^2 \quad (4.1)$$

with  $z$  equals to the height difference between the lower and upper loudspeakers and  $x$  equals to the projection of the distance between the listener and the upper loudspeaker in the lower plane. Considering  $y=0$ , results:

$$x^2 + z^2 = c^2 \quad (4.2)$$

The value of  $x$  can be obtained from  $z = x \cdot \sin(\alpha)$ , so, rewriting the previous equation, results in:

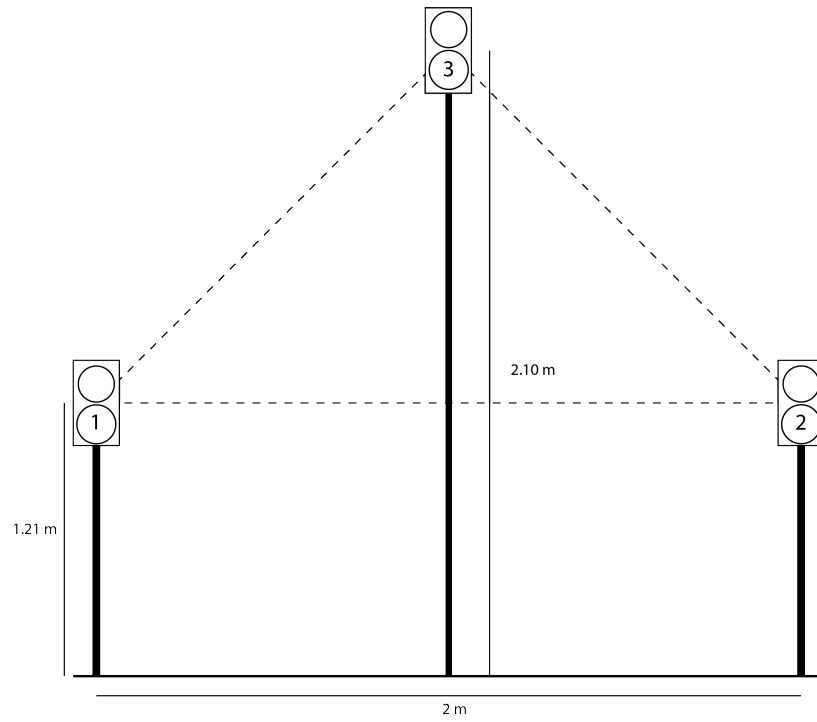
$$x = \frac{z}{\sin(\alpha)}. \quad (4.3)$$

Solving the equation with the conclusions about  $x$  and  $c$ , the following result is obtained:

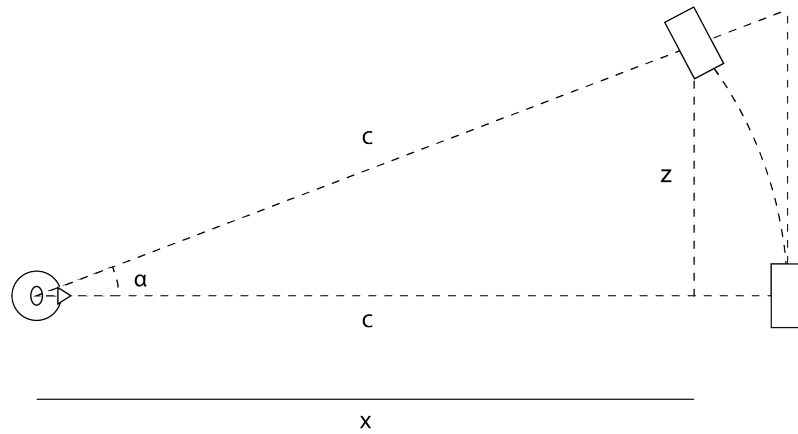
$$z^2 = \frac{(c^2 \cdot \sin^2(\alpha))}{(1 + \sin^2(\alpha))} \quad (4.4)$$

$$x^2 = \frac{c^2}{(1 + \sin^2(\alpha))} \quad (4.5)$$

For  $\alpha=30^\circ$  and  $c=2$  m, the central loudspeaker is located at a height  **$z=0.89$  m** from the lower plane (2.10 m from the floor) and  **$x=1.78$  m** of horizontal distance from the listener, so 22 cm nearer to the listener with respect to the line between the two lower loudspeakers.

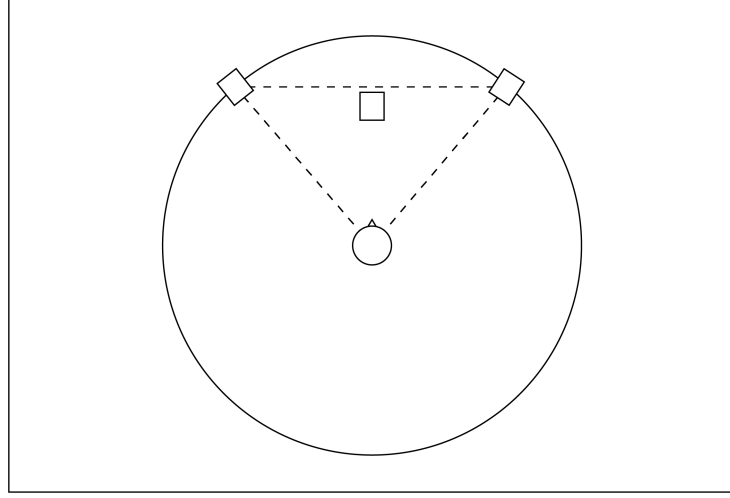


**Figure 4.2:** Frontal View of the Loudspeaker Array



**Figure 4.3:** Lateral View of the Loudspeaker Array





**Figure 4.4:** Aerial View of the Loudspeaker Array

### 4.2.3 Curtain

The system was hidden from the listener with an acoustic curtain (Image 4.6), useful, first of all, for hiding the placement of the loudspeakers, to avoid influencing the subjects' answers (the array was smaller than the curtain). The dimensions of the curtain were 238 x 169 cm, respectively for width and height.

In addition, a grid with 3 lines and 4 columns, resulting in 12 squares, was drawn on the curtain, useful to the subjects to have a reference with the interface to indicate the answers, .

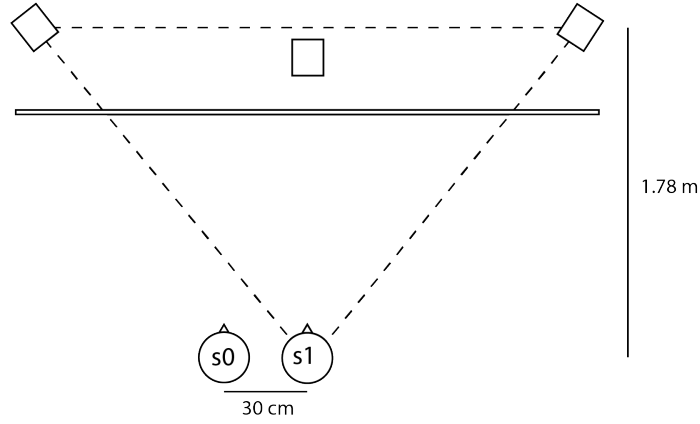
The screen has been horizontally centred in respect to the loudspeakers and it was 19 cm wider than the array for each part, to avoid giving information about the placement of the array to the subject. For the same reason, the curtain covered a surface of 76,5 cm under the lower plane of the loudspeakers and 1 cm over the higher loudspeaker. The smaller coverage over the higher loudspeaker was due to the position of the ceiling of the room, which was too close, and also due to the fact that the highest point of the experiment was 34,7 cm lower than the highest point of the curtain.

### 4.2.4 Listening Positions

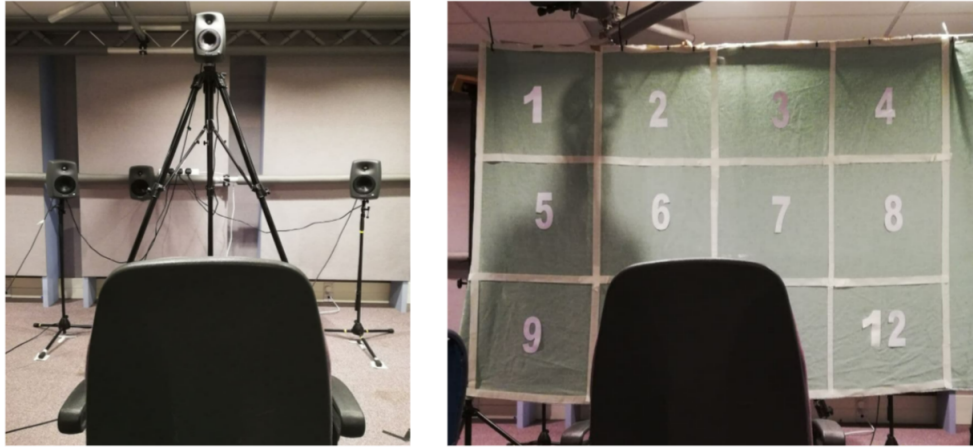
Two listening positions have been used in the experiment, figure 4.5. The first centred position S1 consisted of the listener seated facing the loudspeakers array, at a distance of 1.78 m from the axis formed by the lower loudspeakers, and with the head, in particular the ears, at the same elevation of the lower plane of

the loudspeakers, at 1.21 m. The measurement of the distance from the lower loudspeakers has been obtained from the stereo disposition (the side of the stereo triangle was 2 m long).

In addition to this one, a second position S0 has been used, with the listener facing the loudspeakers, but in a left off-center position of 30 cm. To provide the same position, with the head of all the listeners at the same level as the lower loudspeakers, a chair with an adjustable height has been used. Before the start of each session, the height of the chair has been adjusted for each participant.



**Figure 4.5:** Aerial View of the Sitting Position



**Figure 4.6:** Frontal Listening Position With and Without the Curtain

### 4.2.5 Interface

For the experiment, the answers of the subjects has been recorded with a specially designed graphical user interface, realized with MAX MSP, which was displayed on a monitor placed in front of them.

For the first experiment, the user interface has been used to proceed through the samples and to register the perceived positions and the locatedness information.

The subject, for each stimulus, had to indicate, on the screen, with a cursor, the position of the perceived source in terms of X and Y coordinates. In addition, the participants were asked to rate the degree of certainty of each answer, on a scale of continuous integer values from 0 (fully uncertain) to 100 (fully certain), with three other intermediate values at the points: 25 (“I am really not sure”), 50 (“I have a doubt”) and 75 (“I have a slight doubt”), as suggested in [67]. The listener can reproduce every sample as many times as he wants.

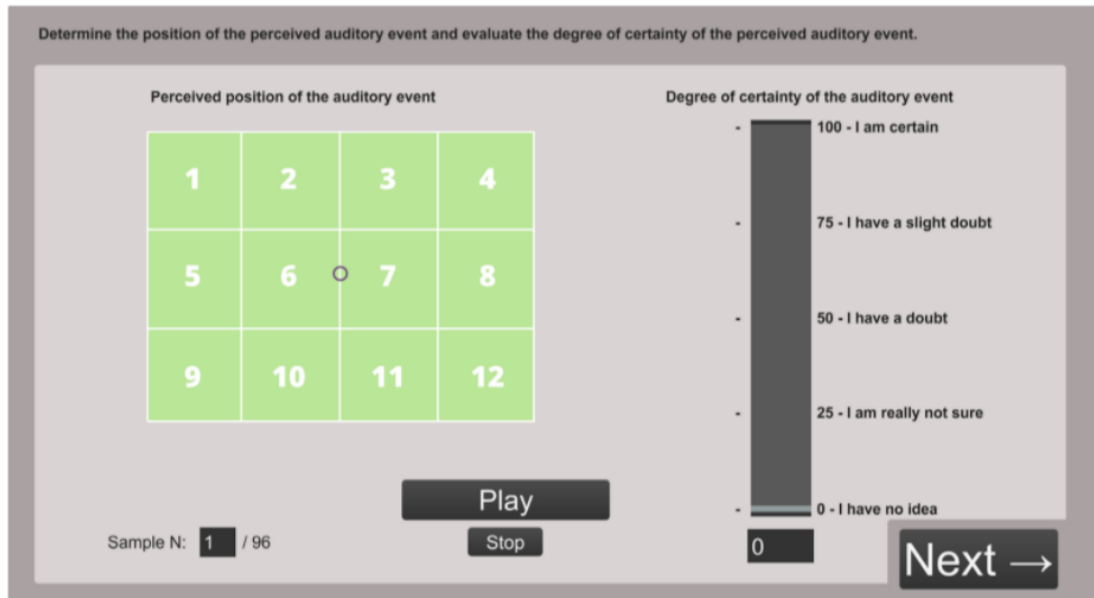


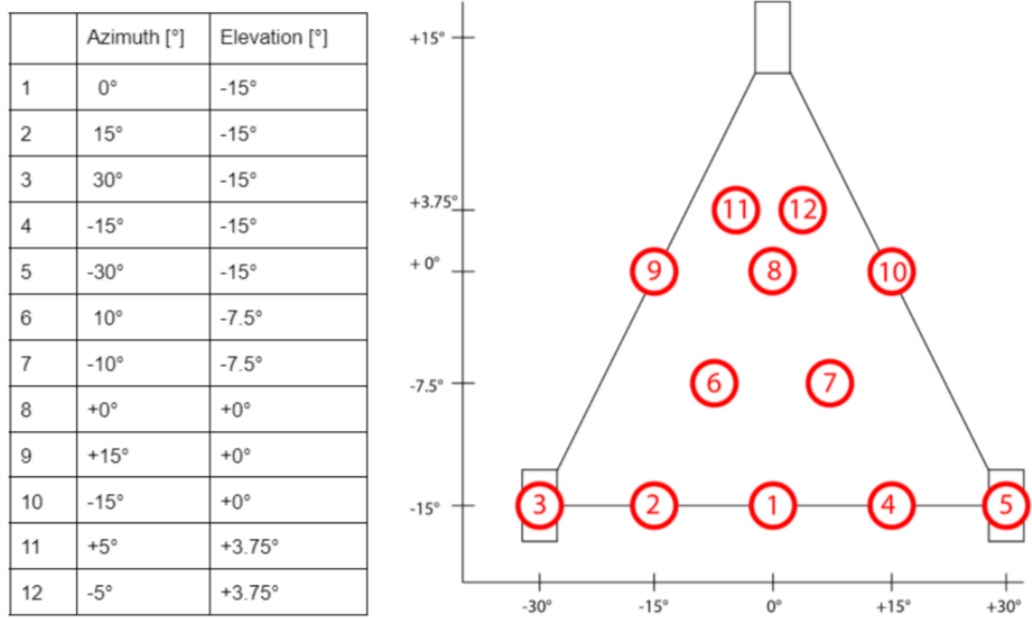
Figure 4.7: Experiment Interface

## 4.3 Methodology and Stimuli

### 4.3.1 Methodology

For the experiment, 12 virtual sound positions inside the triangle of the loudspeakers have been considered. The horizontal angle of the loudspeakers  $\Phi$  is  $60^\circ$ , with the soundwave angle positioned at  $-30^\circ < \theta < +30^\circ$ , with the centre point in zero. The

most negative angle refers to a source positioned at the most extreme right point, while the biggest value of the angle refers to the extreme left point. The vertical angle between the head of the listener and the upper loudspeakers  $\gamma$  is  $30^\circ$ , with the same zero references for the plane wave position, lowest point at  $-15^\circ$  (lower plane) and the highest point at  $15^\circ$  (corresponding to the position of the higher loudspeaker). The azimuth and elevation angle of each source proposed in the experiment is represented in the figure and chart 4.8.



**Figure 4.8:** Stimuli Position Chart and Figure

The experiment was made up of 96 samples to analyse, consisting of the 12 positions for the four methods (VBAP and Hybrid with array radius of 0 cm, 15,5 cm and 50 cm), presented in a random order for every participant and every seating position.

The measures of the radius were chosen like that to compare the result with the VBAP Intensity Difference method ( $r_a = 0cm$ ), with the same measure of the experiment presented in [2] and correspondent to the proposed best one for PSR ( $r_a = 15.5cm$ ) [45], and with an ( $r_a = 50cm$ ), which is a measure proposed by the Tonmeister Student of the University of Surrey Tom Thorpe and represents the minimum distance which can be used in a loudspeaker array of  $60^\circ$  with a first order microphone.

The subjects task was to listen to each of the 96 samples and indicate, on the interface positioned on the screen under the curtain, on the floor, the perceived

position of the sound source and the degree of certainty. Every participant had to point their head straight to the curtain to listen to the stimuli and then when he/she was sure about the answer, look down to watch the screen and indicate the answer. The subject gave the answer of localization placing the cursor on a point on the grid, and then indicated the degree of certainty of the location of the perceived source. The subjects were not aware of the characteristics of the array.

The experiment has been run in two parts, one in a centred position and one in an off-center position, in a random order for every participant. Between the two parts, a pause of at least 5 minutes was made, with the participants free to start the second part whenever they wanted after this amount of time. The average time, considering both the parts and the pause was of 1.30 hours.

### **4.3.2 Stimuli**

As stimuli, two anechoic samples from Bang & Olufsen “Music for Archimedes” CD has been used, consisting of a sample of female speech, and an African bongo, as representative common program material [67]. Every sample was faded off after 7 s, or to the closest point to avoid clicks. Each subject ran a training test before the actual session, to familiarize themselves with the sample and the system of the experiment.

### **4.3.3 Experiment Execution**

Before the beginning of the experiment, the conductor explained to each participant the method of work and the level of exposition. In addition, the person had to read, check and sign a consent form with information about data storage and usage and potential risk of exposition to the audio levels (see Appendix A). After this part, the elevation of the chair was set, according to the height of the participant in a sitting position. The reference was the position of the ears at 1.21 m of height, the same as the loudspeakers. The next step consisted in the calibration of the system, with the participant who had the possibility to slightly adjust the relative volume of the loudspeakers.

Before each part of the experiment started, a familiarization session was run, consisting of twelve samples in three different positions for the four methods, considering the easiest point, the hardest point and one in the average difficulty of localization. The familiarization has been replicated also for the off-center position, with a randomized order between the two seating positions.

The familiarization part does not count towards the final results but it is useful only for the participant to understand and familiarize themselves with the experiment and with the interface. At the end, the conductor entered the room again to check if the subject had any questions and started the proper experiment, before leaving the room again.

At the end of each part, every participant was asked for their opinion about the test.

#### **4.3.4 Subjects**

Twelve participant without any hearing impairment took part in the test. Of this number, eleven were trained listeners, students or researchers from the Institute of Sound Recording of the University of Surrey. The participants were all men apart from one woman (the only non trained listener). The participants were voluntary recruited via email.

### **4.4 Chapter Conclusions**

In the previous chapter, the main setup of the experiment has been presented. The goal of the experiment has been to compare the performance of VBAP with the Hybrid method, this last one implemented with different array radius (0, 15,5 and 50 cm).

The listener was located in the center of a stereophonic triangle, with the position of the third central loudspeaker, positioned in a higher position, founded through trigonometry rules.

In addition to this center position, also an off-center position has been used, with the listener located 30 cm horizontally leftward de-centered.

The subject had to define the position of different sound stimulus in terms of X and Y position. This task is called Localisation. The second question was about the degree of certainty of the previous answer, defining the "Locatedness" of the sound event.

The answers were collected with a particular interface realised with MAX MSP.

The results of the twelve experiments will be discussed in the next chapter.

## Chapter 5

# Experimental Results and Discussion

In this section the results of the experiment presented in the previous chapter will be discussed. The data will be analyzed in terms of:

- **Locatedness of the phantom source:** an important perceptual attribute of multichannel reproduction, defined as the degree of certainty of the location of the auditory events [67].
- **Angular Error:** The difference between the given answer (for each x and y dimension) and the actual value of the auditory event. The horizontal answers on the lower plane will be particularly interesting for the comparison with a similar experiment included in [2], while for the vertical ones, a comparison between VBAP and Hybrid 15.5 will be investigated.

The analysis of the data has been made with the software IBM SPSS Statistics.

The performance of each method has been compared with a non parametric Kruskal-Wallis test, since the data was not normally distributed. The parameter used to check the differences between methods is the statistical significance  $\alpha$ , with a chosen value of  $\alpha=0.05$ ; The test to prove the hypothesis gave back a value p (called **p-value**), which represents how random the result of the observation is. If  $p \leq \alpha$ , the null hypothesis is rejected, and the results take on statistical significance, resulting in some differences between the methods compared.

## 5.1 Locatedness

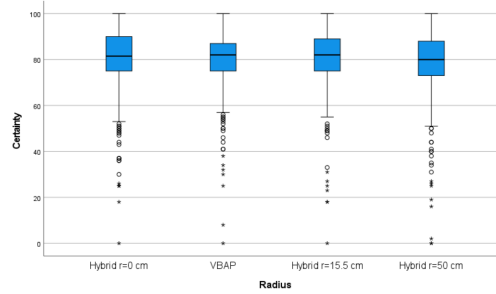
The data of the twelve listeners has been grouped by method, and then the total average certainty has been calculated.

### 5.1.1 Center + Off-Center

As expected, the perception in a center position reached better values of certainty in respect to the off-center position (AVG Answer 79,00 vs 75,76). It is important to remember that the listeners were unaware of which one was the center one and which one was not. In addition, the positions were presented in a different order among the participants, due to the blind characterisation of the test.

### 5.1.2 Center Position

From the analysis of the Certainty in the center position, no differences between the methods has been found, since the significance of the test was 0.323. This results in an equal certainty for the methods in a center position, see Figure 5.1



**Figure 5.1:** General Certainty for Different Methods in All the Stimuli Positions. Center Position

Also going through the pairwise comparison, all the methods are equal to each other, see Figure 5.2, with the Hybrid 50cm method which performs worse than the others, as expected. Also individually considering the certainty results in the Azimuth extreme points ( $\theta = 30^\circ$ ,  $\theta = -30^\circ$ ,  $\theta = 15^\circ$ ,  $\theta = -15^\circ$ ) and in the center (Azimuth=  $0^\circ$ ), no particular differences between methods has been noticed, with a significance of 0,277 and 0,411 respectively.



Confronti pairwise di Radius					
Sample 1-Sample 2	Statistica del test	Errore std.	Statistica test standard	Sig.	Mod. Sign. *
Hybrid r=50 cm-VBAP	33,155	27,707	1,197	,231	1,000
Hybrid r=50 cm-15.00	43,585	27,707	1,573	,116	,694
Hybrid r=50 cm-Hybrid r=0 cm	45,601	27,707	1,646	,100	,599
VBAP-15.00	-10,431	27,707	-,376	,707	1,000
VBAP-Hybrid r=0 cm	12,446	27,707	,449	,653	1,000
15.00-Hybrid r=0 cm	2,016	27,707	,073	,942	1,000

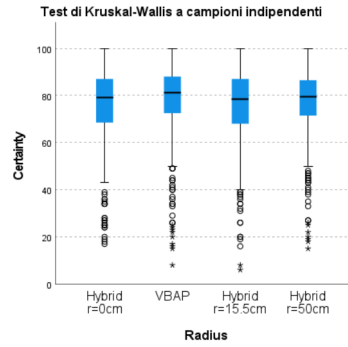
**Figure 5.2:** Pairwise Comparison Between the Certainty for Different Methods in All the Stimuli Positions. Center Position

Considering now a comparison between VBAP and Hybrid 15.5 on the lower plane only for the center position, no differences between the methods has been found ( $p=0.348$ ), as found for the PSR method in [2].

Lastly, considering the single heights, no differences of locatedness has been found.

### 5.1.3 Off-Center Position

The same Kruskal-Wallis test has been made for Certainty on the Off-Center position, see Figure 5.3. Also in this case, no difference has been founded between the methods, with a significance of 0.173.



**Figure 5.3:** General Certainty for Different Methods in All the Positions. Off Center Position

Considering the pairwise comparison, in Figure 5.4, the only couple which is near to a significance  $\leq 0.5$  are (Hybrid 15.5 cm - VBAP) and (Hybrid 0 cm - VBAP). For the second case, there is not an explanation for the result, since similar results were expected.

Confronti pairwise di Radius					
Sample 1-Sample 2	Statistica del test	Errore std.	Statistica test standard	Sig.	Mod. Sign. <sup>a</sup>
Hybrid r=0cm-15.00	-.974	27,708	-.035	,972	1,000
Hybrid r=0cm-Hybrid r=50cm	-18,575	27,708	-.670	,503	1,000
Hybrid r=0cm-VBAP	-54,035	27,708	-1,950	,051	,307
15.00-Hybrid r=50cm	-17,601	27,708	-.635	,525	1,000
15.00-VBAP	53,061	27,708	1,915	,055	,333
Hybrid r=50cm-VBAP	35,460	27,708	1,280	,201	1,000

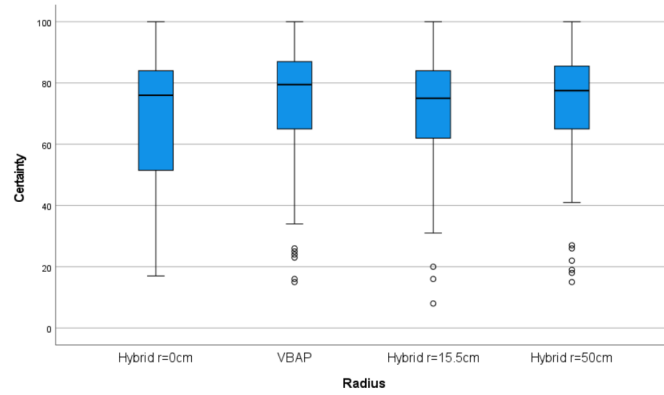
**Figure 5.4:** Pairwise Comparison Between the Certainty for Different Methods in All the Stimuli Positions. Off-Center Position

Considering the point with external azimuth in (a) as ( $\theta = 30^\circ$  and  $\theta = -30^\circ$ ) and (b) ( $\theta = 30^\circ$ ,  $\theta = -30^\circ$ ,  $\theta = 15^\circ$  and  $\theta = -15^\circ$ ), no differences are perceived in terms of certainty, with a significance of 0,145 and 0,210 respectively.

Afterwards, points with negative and positive azimuth were investigated. In the first case, so for points further away from the listener, the significance is equal to 0.087, Image 5.6.

The same comparison has been made also for the points closer to the listener, with positive azimuth, resulting in no difference between the methods,  $p = 0.794$ .

Lastly, answers on the extreme left point have been checked, with a significance of 0.984, resulting in no difference between methods.



**Figure 5.5:** General Certainty for Different Methods in the Points on the Right in Respect to the Center of the Array, Further Away from the Listener. Off-Center Position

Also in this case, VBAP and Hybrid 15.5 have been compared for the values on the lower plane. From the analysis resulted no difference between the methods in terms of certainty (significance = 0.262). Also in the PSR work [2], the certainty between methods was more or less the same comparing VBAP and Hybrid 15.5,

with the only exception of an off-center point where Hybrid resulted better than VBAP.

VBAP	Medio		74,20	1,824
	95% di intervallo di confidenza per la media	Limite inferiore	70,59	
		Limite superiore	77,81	
	Media ritagliata al 5%		75,69	
	Mediana		79,50	
	Varianza		399,321	
	Deviazione std.		19,983	
	Minimo		15	
	Massimo		100	
	Intervallo		85	
	Intervallo interquartile		22	
	Asimmetria		-1,150	,221
	Curtosi		,776	,438
Hybrid r=15.5cm	Medio		71,18	1,693
	95% di intervallo di confidenza per la media	Limite inferiore	67,83	
		Limite superiore	74,53	
	Media ritagliata al 5%		72,19	
	Mediana		75,00	
	Varianza		343,748	
	Deviazione std.		18,540	
	Minimo		8	
	Massimo		100	
	Intervallo		92	
	Intervallo interquartile		22	
	Asimmetria		-,928	,221
	Curtosi		,987	,438

**Figure 5.6:** Descriptive Statistics of Certainty for VBAP and Hybrid 15.5 for the Points in the Right in Respect to the Center of the Array, Further Away From the Listener. Off-Center Position

## 5.2 Localization

### 5.2.1 Conversion of the Localization Data

As seen in the section 4.2.5, every subject gave the answer of localization in a scale, for both horizontal and vertical axis, from a 0 to 127. From now on, let's consider the horizontal axis as **y**, the vertical axis as **z** and the horizontal distance from the listener to the curtain, as **x** (this measure is constant and is equal to 173 cm).

Every answer of the subject has been converted first in  $y_{cm}$  and  $z_{cm}$  measure, and then, from a linear measure ( $y_{cm}, z_{cm}$ ), to spherical coordinates  $(\theta, \gamma)$ , useful to determinate the angle error.

From the given answers  $\bar{ANS}$ , these are converted into the  $ANS_{cm}$  measures with the formulas:

$$y_{cm} = L_y \cdot \left( \frac{\bar{y}}{128} - \frac{1}{2} \right) \quad (5.1)$$

$$z_{cm} = L_z \cdot \left( \frac{\bar{z}}{128} \right) - 76.578 \quad (5.2)$$

where  $L_y$  and  $L_z$  are the length and the height of the curtain, equal to 238 cm and 169 cm respectively. In both the measures are considered the offset due to the centering of the measurement in the point of azimuth and elevation  $(\theta, \gamma) = (0, -15)$ , corresponding to the position of the ears. This is necessary for considering the listening point as the center of the array.

These measures are then converted into angles with a linear (x,y,z) to spherical  $(\rho, \theta, \gamma)$  coordinate conversion. The formulas for the conversion are the following:

$$\begin{aligned} \rho^2 &= x^2 + y^2 + z^2 \\ \tan \theta &= \frac{y}{x} \longrightarrow \theta = \arctan\left(\frac{y}{x}\right) \\ \gamma &= \arcsin\left(\frac{z}{\sqrt{x^2 + y^2 + z^2}}\right) \end{aligned} \quad (5.3)$$

The angles obtained represent the perceived azimuth and elevation of the sound event. The measures, in radians, are then converted into degrees, and compared with the actual measures.

Since in the experiment reference the elevation point correspondent to the head is considered as  $\gamma = -15^\circ$  and from the precious formula we would obtain this point as  $\gamma = 0^\circ$ , and extra offset is added, subtracting every quantity of  $15^\circ$ . The resulting formula for the conversion is:

$$\gamma = \left[ \arcsin\left(\frac{z}{\sqrt{x^2 + y^2 + z^2}}\right) \right]^\circ - 15^\circ \quad (5.4)$$

### 5.2.2 Angular Error

Starting from the angles of the given position of the stimuli  $(\theta, \gamma)$ , for every given answer  $(\theta_0, \gamma_0)$  the absolute angular error for azimuth ( $\theta_{err}$ ) and elevation ( $\gamma_{err}$ ) is calculated, with the following formulas:

$$\begin{aligned} \theta_{err} &= |\theta - \theta_0| \\ \gamma_{err} &= |\gamma - \gamma_0| \end{aligned} \quad (5.5)$$

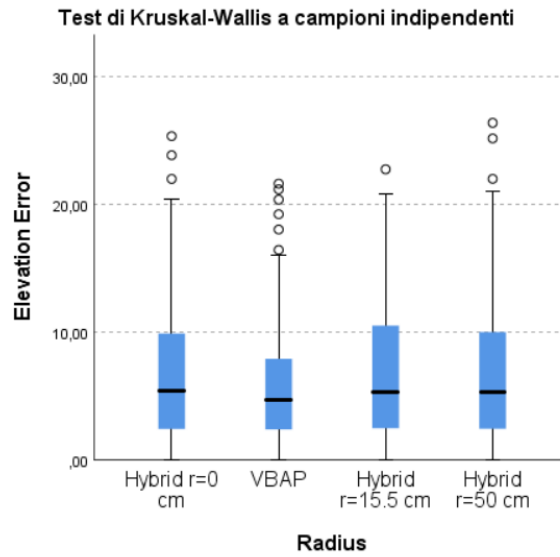
### Center Position

First of all, a comparison between the methods has been run. Considering the horizontal and vertical error individually, reporting no difference between methods, Figure 5.7, with significances of  $p=0.368$  and  $p=0.069$ .

Riepilogo test sull'ipotesi			
	Ipotesi nulla	Test	Sign.
1	La distribuzione di ABSAzim è la stessa sulle categorie di Radius.	Test di Kruskal-Wallis a campioni indipendenti	,368
2	La distribuzione di ABSElev è la stessa sulle categorie di Radius.	Test di Kruskal-Wallis a campioni indipendenti	,069

Le significatività asintotiche sono visualizzate. Il livello di significatività è ,050.

**Figure 5.7:** Kruskal-Wallis Test for Horizontal and Vertical Angular Error Considering all the Methods. Center Position

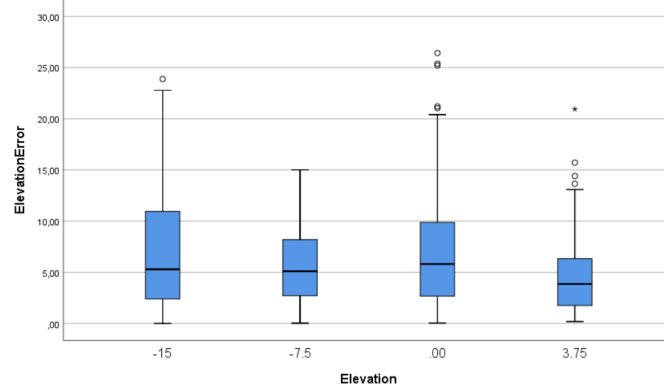


**Figure 5.8:** Comparison Between Vertical Angular Error Considering All the Methods. Center Position

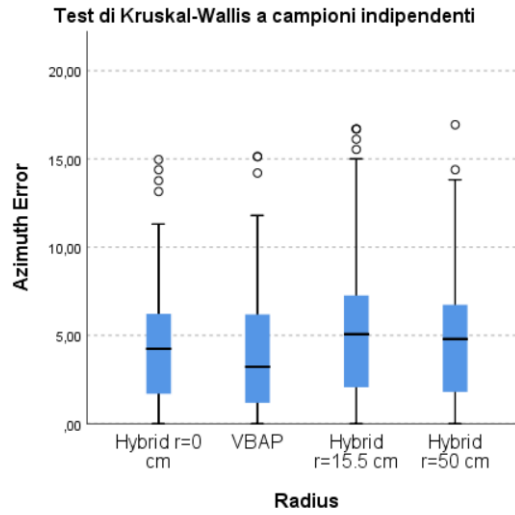
Considering the elevation error in function of the different elevations, the point located at  $3.75^\circ$  is the one with the lowest value of error, with an average error of  $4.58^\circ$ . The vertical point with the worst elevation error is the one located at  $-15^\circ$ , with an average gap of  $6.97^\circ$  between the given answer and the real one. Comparing the different methods in every vertical point, no differences have been noticed singularly in the methods for every vertical point considered.

Evaluating now the performance of the horizontal error for the lower plane only, Image 5.10, the methods are equivalent ( $p=0.075$ ). The errors for the central position are slightly greater than the range from  $1^\circ$  to  $4^\circ$  founded by Blauert in [3]. Considering the same horizontal error for the other height points, at  $-7.5$  Hybrid 15.5 performs slightly better than VBAP ( $p=0.05$ ) and no differences are found

for the central point  $\gamma=0^\circ$  ( $p=0.951$ ) and for the highest point at  $\gamma=3.75^\circ$  ( $p=0.999$ ).



**Figure 5.9:** Elevation Error in Function of the Different Elevation Values. Center Position



**Figure 5.10:** Horizontal Angular Error for the Different Methods. Center Position

Considering the different azimuth for the center position, at the extreme points (Azimuth=30, -30, 15, -15), no differences between the methods have been found for the horizontal ( $p=0.1$ ) and vertical ( $p=0.406$ ) errors. Neither in the central point (Azimuth=0°), any differences between the methods has been found, both for the horizontal ( $p=0.095$ ) and vertical error ( $p=0.596$ ).

Considering the other points singularly, an interesting result has been found for an azimuth of  $\theta = \pm 30^\circ$  (the most left and right points respectively), with VBAP that performs better than Hybrid 15.5 in terms of horizontal error ( $p=0.000$ ) for the Left and ( $p=0.002$ ) for the Right. This is also a known result for VBAP, that better localises sound located in the position of a loudspeaker, but suffers from a higher localisation spread where the loudspeakers are not located [30]. For this reason, the result is not particularly relevant, but has been reported to confirm this aspect.

The differences are not relevant for an azimuth of  $\theta = \pm 15^\circ$ , and this justifies the first result found in this part of the analysis, which compensate the differences in the extreme points with the correlates in the other two points.

Also for azimuth of  $\pm 10^\circ$  and  $\pm 5^\circ$ , no differences between Hybrid 15.5 and VBAP has been found, but the error significantly increases for the other methods, confirming the hypothesis of the best array radius of 15.5 in case of time differences.

As expected, the vertical angular error is always higher than the horizontal one, due to the more complex mechanism for vertical localisation. In addition, the vertical points are often perceived as higher, confirming that seen in the work [70]

## **Off-Center Position**

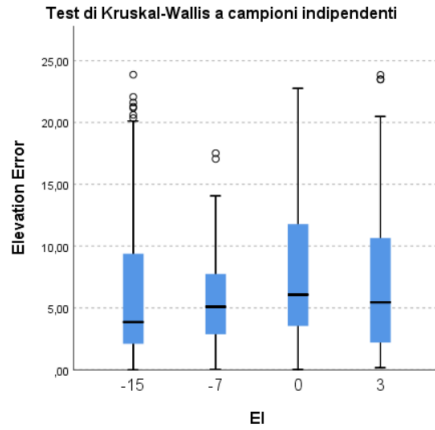
For the off-center position the same method of conversion has been used. It would have been more correct to consider the projection of the position on the sphere from a coordinates system center on the off-center sitting position, to avoid the error in the localization, but, since the introduced error is the same for all the methods, it can be neglected.

As expected, due to the off-center position, both the errors increased with respect to the central position, figures 5.11 and 5.12. Between the two, the one that got worse considerably is the horizontal one, probably due to the horizontal off-center, which implies the lack of ITD and ILD cues.

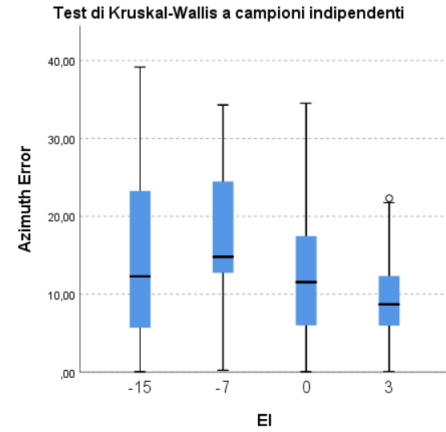
The comparison between the methods shows no differences between the vertical error ( $p=0.164$ ) and some differences resulting from the horizontal one ( $p=0.000$ ). Considering this second case, a big difference resulted from the comparison between VBAP and all the Hybrid methods, Figures 5.13 and 5.14, with a statistical significance of 0.000.

The Hybrid method behaved approximately equal to each other, with the only relevant difference between Hybrid 50 and Hybrid 0 ( $p=0.034$ ), resulting in a lower

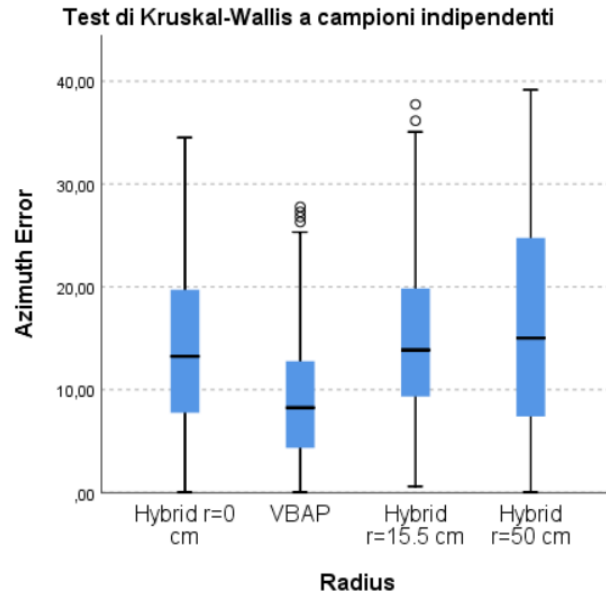
error for the second one. Also the comparison between VBAP and Hybrid 0 was statistically significant, and this is not an expected result.



**Figure 5.11:** Vertical Error for Different Heights



**Figure 5.12:** Horizontal Error for Different Heights



**Figure 5.13:** Azimuth Localization Error. Off-Center Position



Confronti pairwise di Radius					
Sample 1-Sample 2	Statistica del test	Errore standard	Statistica test standard	Sign.	Sign. adattata <sup>a</sup>
VBAP-Hybrid r=0 cm	194,174	27,724	7,004	,000	,000
VBAP-15.00	-234,941	27,724	-8,474	,000	,000
VBAP-Hybrid r=50 cm	-252,892	27,724	-9,122	,000	,000
Hybrid r=0 cm-15.00	-40,767	27,724	-1,470	,141	,849
Hybrid r=0 cm-Hybrid r=50 cm	-58,719	27,724	-2,118	,034	,205
15.00-Hybrid r=50 cm	-17,951	27,724	-,647	,517	1,000

**Figure 5.14:** Pairwise Comparison Between Methods for Horizontal Localisation. Off-Center Position

As in the center position, the lower general error has been found for the highest point, while for the horizontal and vertical error, the best heights are the highest and the lowest point respectively, Figures 5.12 and 5.13.

Considering the vertical error at the different heights, no differences have been noticed for the central point  $\gamma = 0^\circ$  ( $p=0.541$ ), the point  $\gamma = -7.5^\circ$  ( $p=0.998$ ) and in the lower point  $\gamma = -15^\circ$  ( $p=0.424$ ). Regarding the horizontal error for the different heights, in all the positions VBAP performed better than Hybrid 15.5, with a lower average value.

Considering only the point on the lower plane, as said before, the vertical localization presents no differences between the methods. The horizontal localization presents lower error for VBAP compared to Hybrid 15.5. Considering the single points on the lower plane, there is no difference for the extreme right point ( $\theta = -30^\circ$ ) in terms of horizontal error ( $p=0.104$ ). For ( $\theta = -15^\circ$ ), there are no differences between methods for the vertical error, but considering the horizontal one, VBAP is the one with the lowest value ( $p=0.000$ ). On the median plane of the loudspeakers, no differences are noticed for the horizontal ( $p=0.340$ ) and vertical error ( $p=0.793$ ). Same result are obtained for an azimuth of  $\theta = 15^\circ$  and  $\theta = 30^\circ$ .

## 5.3 Chapter Conclusions

### Locatedness

The center position produces a better locatedness in the listener with respect to the off-center position.

Considering only the point on the lower plane, for both center and off center, the performances are the same, that is quite similar in respect of what has been found in [2].

Considering all the points, both for an on-center and an off-center position there is no difference between the methods in all the considered positions in terms of Certainty.

### Angular Error

On the center position, no difference between the methods has been detected, both for the horizontal ( $p=0.368$ ) and vertical ( $p=0.069$ ) angular errors. For every method, the average vertical error resulted higher than the horizontal one, and that is quite normal, considering the anatomic position of the ears.

Considering the different heights, the points are worse perceived when vertically localized on the lower plane (avg error of  $6,97^\circ$ ). The best result has been obtained for the highest point (avg error of  $4,58^\circ$ ). This result can be explained by the highly regarded literature result of the overestimation of the height [70], along with the poor concentration of vertically disposed points around the lower plane, which may contribute to a poor familiarisation of localisation in this zone.

In general, for the different methods, no differences are founded in terms of vertical error for the different heights.

Considering now a comparison between VBAP and Hybrid 15.5 for the lower plane answers, the first method results as a better one with respect to the other (statistical significance= $0.005$ ). This is evident only for the extreme points ( $\theta = \pm 30^\circ$ ), but not for the others. This result, called Detent Effect is aligned to what has been found on [2].

Another common result is that the other Hybrid methods (Hybrid 0 and Hybrid 50) perform much worse in terms of horizontal error, and this confirms the choice of Johnston for the best proposed array radius of 15.5 cm [45].

Lastly, considering the horizontal error at the other heights, the only significant difference is a lower error for Hybrid 15.5 in respect of VBAP at the height of  $\gamma = -7.5^\circ$  ( $p=0.059$ ).

Considering now the off-center position, the only difference between methods has been found for the horizontal error between Hybrid 15 and VBAP, with the second one performing better.

Also with these analogies, TID methods seem to be a good path to follow, also because this type of arrays, the spaced ones, are used in physical systems to reach a better spaciousness of the sound, since two different reverb points are recorded.

## Chapter 6

# Conclusions

Throughout the work, various panning methods for different stereo formats have been analysed. Mainly, the approaches have been divided by two main distinctions: type of panning and number of channels.

The methods have been distinguished, by the number of channels, into: 2 channel stereophony (or binaural), surround (as 5.1 or 7.1), and multichannel with more than 9 channels. Every method presents pros and cons, in terms of quality of the rendering and the complexity of the implementation. The conclusion reached has been that the best compromise is a multichannel implementation with a medium number of channels, which guarantees a wide sweetspot with a manageable complexity of implementation. Various examples of reproduction systems, based on the number of channels, have been presented.

Regarding the type of panning, the distinction has been made between perceptual soundfield and physical reconstruction models. If the first one aims to render the soundsource with a psychoacoustic approach (with the help of cues as ICLD, ICTD and HRTF), the second one tries to record the soundfield with particular microphones arrays.

For the physical reconstruction models, different implementation of microphones arrays has been shown, for binaural stereophony, multichannel and systems with a high-number of channels, providing, at the beginning, an overview about microphone characteristics.

Also in this case, it has been underlined that, the increase in the number of channels improves the quality of the recording, but, at the same time, makes the handling of the single channel difficult, especially with a one-to-one correspondence between the number of microphones and the number of loudspeakers.

Perceptual methods are, in addition, divided in different categories, based on the cues they use. Mainly, have been presented ID, TD, TID and TILD methods, with a particular focus on the first and the third ones.

For ID methods, the panning implemented with the tangent panning law has been presented, which allows us to define the relative gains of two loudspeakers in terms of a tangent ratio, function of the source and loudspeaker angles.

For TID methods, have been shown the use of particular psychoacoustic curves as the ones of Williams and Frannsen, which define the couple (delay, level difference) between the two channels to render a source in a certain position.

Two examples (one for each method) have been illustrated, in order to have a comparison with state-of-the-art techniques.

For TID, PSR has been presented, which is a method for the rendering of a source only on the horizontal plane. The approach is based on the Active Intensity concept, and it has been demonstrated that, for rendering a sound source, only two adjacent active loudspeakers at a time are necessary, with more than two loudspeakers (or two not adjacent), fluctuations, and consequent interferences, are present. This is an importance result, since it brings a multichannel problem back to a stereophonic one, significantly simplifying the task.

As previously seen, the model, starting from three free parameters (radius array, angle between the loudspeakers and angular position of the source), determines the directivity pattern of the microphones array, which corresponds to the ICLD of the two loudspeakers, and the delays ICTD of the two channels. For the implementation, the Williams curve is used.

On the other hand, for ID, the VBAP full-sphere method has been presented. Horizontal-plane only version is available, but the one illustrated is a three-dimensional one, which is a 3D vector reformulation of the Tangent Panning Law. The main advantage is the use of a maximum of three loudspeakers for rendering a sound source in two dimensions, inside a space called the active triangle. Adding an arbitrary number of loudspeakers means an increase the number of the triangles, and consequentially, expands the space where the source can be placed. Even if the system is usable with an arbitrary number of loudspeakers, the cons concern the strict position of these (overcome with DBAP) and the localisation blur problem.

Due to this problem, a possible solution has been studied and proposed. The idea is the introduction of time delays in the full-sphere perceptual panning (only for the horizontal dimension), the aim of which is the improvement of the localisation and the achievement of a more natural sound.

The starting point has been, for the horizontal plane, the PSR method, and, an ID original formulation for the vertical one has been added, based on the Tangent Panning Law and on three main assumptions. Due to a factorization for horizontal and vertical angles, surprisingly a closed formula has been founded for the gains of the three loudspeakers.

The original method has been compared with VBAP in a perceptual experiment, in terms of localisation error (both horizontal and vertical) and Locatedness (certainty of the localisation position). In addition to VBAP and Hybrid with an array radius of 15.5 cm (proposed as the best one), also Hybrid with a radius of 0 (for comparing that with VBAP) and 50 cm (which is the limit for the 60° stereophony) have been tested.

Concentrating on VBAP and Hybrid 15.5, the main results of the statistical analysis have been an equivalent Locatedness for the methods in both the Center and Off-Center Position. For the localisation error, the methods are equivalent in every condition (with a point in the centred position where Hybrid performed slightly better in terms of horizontal error), apart from the azimuthal error for the off-centred position, where VBAP performs better.


Due to these similarities, the TID path can be pursued, since these types of methods are the ones used for a better spaciousness in the physical methods, due to a lower Inter-Channel Cross-Correlation.



# Appendix A

# Appendix

## A.1 Consent Form

  
UNIVERSITY OF  
SURREY  
Institute of Sound Recording (IoSR)

**Consent Form for Participants in IoSR Research Studies**

Study title: Spatial Comparison of Full Sphere Panning Methods  
Ethical approval ID<sup>1</sup>: 801367-801358-87955275  
Researcher name: Enrico Fodde email: [enrico.fodde@studenti.polito.it](mailto:enrico.fodde@studenti.polito.it)

tick

☐ I have read the information sheet relating to this study and have been given adequate time to consider it. I understand the nature, location and likely duration of what I will expect to do. I have been given the opportunity to ask questions, and any questions asked have been answered to my satisfaction.

☐ I understand that my participation is voluntary and that I am free to withdraw at any time during the study without giving any reason and without being disadvantaged in any way.

☐ I consent to the storage, analysis and publication of anonymised data gathered from me during the study, and to the use of such data in future studies. I understand that confidentiality and anonymity will be maintained and that I will not be identified personally in any research output.

☐ I understand that I will be able to withdraw my data at any point up to the end of the session in which it is gathered. I understand that, since data (other than this consent form) will be stored anonymously, it will be possible to identify and withdraw my data after that point.

☐ I consent to the storage of this form and to it, and/or its content, being made available for auditing purposes as required. I understand that such storage and processing will be in accordance with prevailing data protection regulations.

If you have any other questions about the experiment, please do ask.

_____ Name of the participant	_____ Signature	_____ Date
----------------------------------	--------------------	---------------

<sup>1</sup> If ethical review was by self-assessment only and did not require a formal Ethics & Governance Application (EGA) then the ID is the Self-Assessment for Governance & Ethics (SAGE) form response ID.



## A.2 Listening Test Instruction

### Listening Test Instructions

Thank you for taking part in this experiment. Your participation is very important and useful for this project and is therefore much appreciated.

There will be two experiment sessions, focusing on one or more attributes per session. The attributes that you will be focusing on today are **LOCALIZATION** and **CERTAINTY**.

Localization	Which is the position of the auditory event.
Certainty	How certain you are about the position of the auditory event.
Auditory Event	Subjective auditory perception due to the exposure of a sound.

Each stimulus will consist of some audio (a speech sample and a bongo sample) through some loudspeakers positioned behind a curtain. For each stimulus, you will be asked to indicate the position of the auditory event and indicate your degree of certainty with respect to the position, on a scale from 0 (I have no idea of the position) to 100 (I am certain of the position).

If the position of the auditory event is not clear in only one position, please indicate the perceived position in average.

Before the actual start of the test, a practice part will be run. The Familiarization Part consists of a series of 24 samples, in two different positions of the seat, that resumes the spectrum of difficulty of the experiment, to help you to take some confidence with the interface and the experiment.

### Test Interface

The interface for the practice test and the real test will be the same and is shown below.

Determine the position of the perceived auditory event and evaluate the degree of certainty of the perceived auditory event.

Perceived position of the auditory event

1	2	3	4
5	6	7	8
9	10	11	12

Degree of certainty of the auditory event

100 - I am certain

75 - I have a slight doubt

50 - I have a doubt

25 - I am really not sure

0 - I have no idea

Sample N: 1 / 96

Play Stop

0

Next →

Click on the play button to start or stop the playback of the sample. Each sample consists of a 7 seconds audio stimulus. For each sample define the perceived position in the grid and establish a score for the certainty of the given answer. Press the NEXT button to confirm the answer and pass to the next sample. Please notice that it will not be possible to go back to the previous samples once you click NEXT.

If you have any questions then please ask me.

If you are happy with all of the above and you would like to participate, please complete the consent form. You are welcome to keep this information sheet.

## A.3 Experiment Guide

1. Open the Patch named Localization.

2. Select the input file.

The file is composed of 98 lines, one for each sample plus a starting and an ending null line (formed by zeros), useful for synchronizing the input and output files. The effective file starts on the second line, with the following line format:

1	#line, #sample	radius_array	azimuth_planewave	elevation_planewave	#audio_sample	#Method
---	----------------	--------------	-------------------	---------------------	---------------	---------

- The first two numbers represent the number of the line in the proposed order (randomized), and the number of the sample in the original one.
- The radius of the array can be 0, 15.5 cm or 50 cm.
- The azimuth takes values between  $-30^\circ$  and  $+30^\circ$  and the elevation values are included between  $-15^\circ$  and  $+3.75^\circ$ .
- For the experiment two audio samples are selected, both cut in a window of approx 10 seconds (considering the silent point nearest to 10 seconds, to avoid audible clicks). Code 1 corresponds to a Voice sample, while 2 corresponds to a Bongo sample.
- As the final parameter, the method code is 1 for the Hybrid one and 2 for VBAP.

3. Press the START button to read the first line.

4. Select the output file.

5. Go in Presentation Mode and start the experiment.

6. After the 96th sample, the program will ask you to insert the code of the participant.

Write the code in the form: code.txt. Es: 001.txt

The output file will contain 96 lines with the following format:

1 #line, #sample radius azim elev #audio_sample #Method x_pos y_pos cert_degree
---

- The position is composed of the x and y positions. Both of them can assume values between 0 and 128.
- The certainty degree is an integer value included between 0 and 100.

# Bibliography

- [1] O.M. Lord Rayleigh and Pres. R.S. «XII. On our perception of sound direction». In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13.74 (1907), pp. 214–232 (cit. on pp. 1, 13).
- [2] Hüseyin Hacıhabiboğlu Enzo De Sena and Zoran Cvetković. «Analysis and Design of Multichannel Systems for Perceptual Sound Field Reconstruction». In: *IEEE Transactions on audio, speech, and language processing* 21.8 (Aug. 2013), pp. 1653–1665 (cit. on pp. 3, 22, 36, 38, 60, 77, 80, 82, 83, 91).
- [3] Jens Blauert. *Spatial Hearing - The Psychoacoustic of Human Sound Localization*. Cambridge, Massachusetts, London, England: The MIT Press, 1983 (cit. on pp. 4, 10, 11, 14–16, 18, 86).
- [4] Hans Lungwitz. *Die Entdeckung der-Seele Allg. Psychobiol. [The discovery of the soul- General Psychobiology]*. Leipzig: Oldenburg: Brücke-Verlag – Kurt Schmiersow, 1923 (cit. on p. 4).
- [5] Vincenzo Lombardo - Andrea Valle. *Audio e Multimedia*. Milano, Italy: Apogeo, 2002 (cit. on pp. 5, 7, 10, 11, 13, 18, 19).
- [6] M. Risoud, J.N. Hanson, F. Gauvrit, C. Renard, P.E. Lemesre, N.X. Bonne, and C. Vincent. «Localisation sonore spatiale». In: *Annales françaises d’Otorhino-laryngologie et de Pathologie Cervico-faciale* 135.4 (Sept. 2018), pp. 251–257 (cit. on pp. 5, 11, 12, 16, 17).
- [7] Castano - Cocco - Floriani - Spinella. *Anatomia Umana*. Milano, Italy: edimeres, 1999 (cit. on p. 6).
- [8] F. Alton Everest. *Manuale di Acustica*. Milano, Italy: Editore Ulrico Hoepli, 1996 (cit. on pp. 6, 7, 19).
- [9] Wikiwand. *Cochlea*. URL: [www.wikiwand.com/en/Cochlea](http://www.wikiwand.com/en/Cochlea) (cit. on p. 8).
- [10] Physics LibreText. *Sound: An Interactive eBook*. URL: [www.compadre.org/books/SoundBook](http://www.compadre.org/books/SoundBook) (cit. on pp. 8, 9).
- [11] Curtis Roads. *The Computer Music Tutorial*. Cambridge, Massachusetts, London, England: The MIT Press, 1996 (cit. on p. 11).

- [12] Bosun Xie. *Head-Related Transfer Function and Virtual Auditory Display: Second Edition*. Plantation (FL), USA: J.Ross Publishing, 2013 (cit. on pp. 11, 13, 14, 17).
- [13] University of Minnesota- Dr. Cheryl Olman. *Interaural time difference*. URL: [pressbooks.umn.edu/sensationandperception/chapter/interaural-time-difference-draft/](http://pressbooks.umn.edu/sensationandperception/chapter/interaural-time-difference-draft/) (cit. on p. 11).
- [14] OpenLearn. *12.3 Interaural time delays: continuous tones*. URL: [www.open.edu/openlearn/science-maths-technology/biology/hearing/content-section-12.3](http://www.open.edu/openlearn/science-maths-technology/biology/hearing/content-section-12.3) (cit. on p. 12).
- [15] Xuan Zhong. «Dynamic Spatial Hearing by Human and Robot Listeners». MA thesis. may: Arizona State University, 2015 (cit. on p. 12).
- [16] J.V. Opstal. *The auditory system and human sound localization behavior*. Boston, MA: Elsevier, 2016 (cit. on p. 12).
- [17] Batteau DW. «The role of the pinna in human localization». In: *Proc R Soc Lond B Biol Sci*. (Aug. 1967), pp. 158–180 (cit. on pp. 13, 14).
- [18] Ville Pulkki. «Spatial sound generation and perception by amplitude panning techniques». MA thesis. Aug: Helsinki University of Technology Laboratory of Acoustics and Audio Signal Processing, 2001 (cit. on pp. 13, 43).
- [19] Hans Wallach. «The role of the head movements and vestibular and visual cues in sound localization». In: *Journal of Experimental Psychology* 27.4 (Oct. 1940) (cit. on p. 13).
- [20] Benjamin Bernfeld. «Attempts for better understanding of the directional stereophonic directional mechanism». In: *44th Audio Engineering Society Convention* (Feb. 1973) (cit. on pp. 13, 20–22, 27, 30).
- [21] P. Morse and K. Ingard. *Theoretical Acoustics, International Series in Pure and Applied Physics*. Princeton: Princeton University Press, 1968 (cit. on p. 15).
- [22] W.A. Yost and X. Zhong. «Sound source localization identification accuracy: bandwidth dependencies». In: *J Acoust Soc Am* 136 (2014), pp. 2737–2746 (cit. on p. 15).
- [23] C. C. Pratt. «Perceptual Spatial Audio Recording, Simulation and Rendering». In: *IEEE Signal Processing Magazine* (May 2017) (cit. on p. 16).
- [24] Robert A. Butler and Richard A. Humanski. «Localization of sound in the vertical plane with and without high-frequency spectral cues». In: *Perception Psychophysics* 51 (1992), pp. 182–186 (cit. on p. 17).
- [25] Zahorik P. «Assessing auditory distance perception using virtual acoustics». In: *Journal of the Acoustical Society of America* 111.4 (2002), pp. 1832–1846 (cit. on p. 17).

- [26] B. C. Moore. *Hearing*. Academic Press, 1995 (cit. on p. 18).
- [27] Hüseyin Hacıhabiboğlu, Enzo De Sena, Zoran Cvetković, James Johnston, and Julius O. Smith III. «Perceptual Spatial Audio Recording, Simulation and Rendering». In: *IEEE Signal Processing Magazine* 17 (May 2017), pp. 1053–1088 (cit. on pp. 18, 20, 21, 33, 36, 47, 57–60).
- [28] H. Fletcher and W. A. Munson. «Loudness, its definition. Measurement and calculation». In: *Journal of the Acoustical Society of America* 5 (Sept. 1933), pp. 82–108 (cit. on p. 18).
- [29] Yôiti Suzuki. «ISO 226:2003 -Precise and Full-range Determination of Two-dimensional Equal Loudness Contours». In: (2003) (cit. on p. 19).
- [30] Ville Pulkki. «Virtual Sound Source Positioning Using Vector Base Amplitude Panning». In: *J. Audio Eng. Soc* 45.6 (June 1997), pp. 1653–1665 (cit. on pp. 22, 43–46, 88).
- [31] N. V. Franssen. «Stereophony». In: *44th Audio Engineering Society Convention* (1964) (cit. on p. 22).
- [32] M. Williams and G. Le Du. «Microphone array analysis for multichannel sound recording». In: *AES 107th Conv.* (Sept. 1999). Preprint 4997 (cit. on p. 22).
- [33] Enzo De Sena, Zoran Cvetković, Hüseyin Hacıhabiboğlu, Marc Moonen, and Toon van Waterschoot. «Localization Uncertainty in Time-Amplitude Stereophonic Reproduction». In: *IEEE/ACM Transactions on audio, speech, and language processing* 28 (2020), pp. 1000–1015 (cit. on pp. 23, 24).
- [34] Shure. *Mic Basics: What is Frequency Response?* URL: [www.shure.com/en-US/performance-production/louder/mic-basics-frequency-response](http://www.shure.com/en-US/performance-production/louder/mic-basics-frequency-response) (cit. on p. 25).
- [35] University of Texas and Austin. *Workshop Recap: Audio Recording and Editing*. URL: [www.dwrl.utexas.edu/2016/10/17/recap-audio-workshop/](http://www.dwrl.utexas.edu/2016/10/17/recap-audio-workshop/) (cit. on p. 26).
- [36] envato tuts+. *6 Stereo Microphones Techniques which can use*. URL: [music.tutsplus.com/it/tutorials/6-stereo-miking-techniques-you-can-use-today--audio-204](http://music.tutsplus.com/it/tutorials/6-stereo-miking-techniques-you-can-use-today--audio-204) (cit. on p. 27).
- [37] DPA Microphones. *Stereo recording techniques and setups*. URL: [www.dpamicrophones.com/mic-university/stereo-recording-techniques-and-setups](http://www.dpamicrophones.com/mic-university/stereo-recording-techniques-and-setups) (cit. on pp. 28–31).
- [38] Elektronauts. *Mid-Side recording with an active mic and a passive mic*. URL: [www.elektronauts.com/t/mid-side-recording-with-an-active-mic-and-a-passive-mic-tips/39036](http://www.elektronauts.com/t/mid-side-recording-with-an-active-mic-and-a-passive-mic-tips/39036) (cit. on p. 28).

- [39] Berlin Georg Neumann GmbH. «The Dummy Head, Theory and practice». In: () (cit. on p. 32).
- [40] Berlin Peus Stephan Georg Neumann GmbH. «Natural Listening with a dummy head». In: (July 1985) (cit. on p. 32).
- [41] International Telecommunication Union. «Report ITU-R BS.2159-8- Multichannel sound technology in home and broadcasting applications». In: (July 2019) (cit. on pp. 33, 59).
- [42] Wikiwand. *Decca tree*. URL: [www.wikiwand.com/en/Decca\\_tree](http://www.wikiwand.com/en/Decca_tree) (cit. on p. 34).
- [43] David M. Howard and Jamie Angus. *Acoustics and Psychoacoustics*. New York and London: Taylor Francis, 2017 (cit. on pp. 35, 50, 52).
- [44] Radiocommunication Sector International Telecommunication Union. «Recommendation ITU-R BS.775-3. Multichannel stereophonic sound system with and without accompanying picture». In: (Aug. 2012) (cit. on p. 35).
- [45] J. D. Johnston and Y. H. Lam. «Perceptual soundfield reconstruction». In: *AES 109th Conv* (Sept. 2000). Preprint #2399 (cit. on pp. 36, 77, 92).
- [46] Jörn Nettingsmeier. «The Why and How of With-Height Surround Sound». In: (Jan. 2012) (cit. on p. 43).
- [47] Ashley Andrew-Jones, Zoran Cvetković, Hüseyin Hacıhabiboğlu, and Enzo De Sena. «Time-Intensity Panning In The Median Plane». In: *Junior Audio Eng. Society* 1.1 (Mar. 2021) (cit. on pp. 43, 44).
- [48] B. Martin A. Tregonning. «The vertical precedence effect: Utilizing delay panning for height channel mixing in 3D audio». In: (2015) (cit. on p. 43).
- [49] Pascal Baltazar Trond Lossius and Theo de la Hogue. «DBAP - Distance-Based Amplitude Panning». In: (2011) (cit. on p. 48).
- [50] Joshua D. Reiss Dimitar Kostadinov and Valeri Mladenov. «Evaluation of Distance Based Amplitude Panning for spatial audio». In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (Mar. 2011) (cit. on p. 48).
- [51] Nathan Wolek and Julien Rabin. *VBAP package for Max*. URL: [github.com/nwolek/vbap](https://github.com/nwolek/vbap) (cit. on p. 48).
- [52] SoundField. *Ambisonics: an introduction*. URL: <https://www.soundfield.com/#/learn/ambisonics> (cit. on pp. 49, 50, 55).
- [53] Andrea Criniti. «Tecnologie di spazializzazione audio per la ripresa e la riproduzione multicanale». MA thesis. Jul: Politecnico di Torino, 2021 (cit. on p. 49).



- [54] J. Daniel and S. Moreau. «Further study of sound field coding with higher order ambisonics». In: *in Proc. 116th Conv. Audio Engineering Society* (May 2004). Preprint #6017 (cit. on pp. 49, 51).
- [55] DPA Microphones. *Immersive sound/ object-based audio and microphones*. URL: [www.dpamicrophones.com/mic-university/immersive-sound-object-based-audio-and-microphones](http://www.dpamicrophones.com/mic-university/immersive-sound-object-based-audio-and-microphones) (cit. on pp. 50, 57, 58).
- [56] *Ambisonics*. URL: [w2.mat.ucsb.edu/240/D/notes/Ambisonics.html](http://w2.mat.ucsb.edu/240/D/notes/Ambisonics.html) (cit. on p. 51).
- [57] D. de Vries A. J. Berkhout and P. Vogel. «Acoustic control by wave field synthesis». In: *J. Acoust. Soc. Amer.* 93.6 (1993), pp. 2764–2778 (cit. on p. 53).
- [58] D. de Vries A. J. Berkhout and P. Vogel. «Vibration Analysis of Edge and Middle Exciters in Multiactuator Panels». In: *AES Convention Paper 8548* (Oct. 2011) (cit. on p. 53).
- [59] Basilio Pueo, José Vicente Rico1, and José Javier Lòpez. «Vibration analysis of edge and middle exciters in multiactuator panels». In: *Audio Engineering Society Convention Paper 8548* (Oct. 2011) (cit. on p. 53).
- [60] Michael Makarski, Anselm Goertz, Stefan Weinzierl, and Christoph Moldrzyk. «Development of loudspeakers for wave field synthesis systems». In: *VDT International Convention* (Nov. 2008) (cit. on p. 54).
- [61] Rode - Soundfield. «SPS200 Software Controlled Microphone user guide v 1.02». In: () (cit. on pp. 55, 56).
- [62] SoundField. *Soundfield microphone basics*. URL: [https://www.soundfield.com/#/learn/microphone\\_basics](https://www.soundfield.com/#/learn/microphone_basics) (cit. on p. 55).
- [63] mh acoustics. *Eigenmike microphone*. URL: <https://mhacoustics.com/products> (cit. on p. 56).
- [64] Jakob Vennerød. «Binaural Reproduction of Higher Order Ambisonics». MA thesis. Jun: Norwegian University of Science and Technology, 2014 (cit. on p. 56).
- [65] Schoeps Mikrofone Helmut Wittek. «Development and application of a stereophonic multichannel recording technique for 3D Audio & VR». In: (2016) (cit. on pp. 68, 69).
- [66] F. Rumsey H. Wittek and G. Theile. «Perceptual enhancement of wavefield synthesis by stereophonic means». In: *J. Audio Eng. Soc.* 55.9 (Sept. 2007), pp. 723–751 (cit. on p. 70).
- [67] L. S. R. Simon and R. Mason. «Time and level localization curves for a regularly-spaced octagon loudspeaker array». In: *Proc. AES 128th Conv.* (May 2010). preprint #8079 (cit. on pp. 70, 76, 78, 80).

- [68] ITU-R International Telecommunication Union- Radiocommunication Sector. «Recommendation ITU-R BS.1116-3». In: (2015) (cit. on p. 70).
- [69] HHB UK. *University of Surrey opens 22.2 listening room*. URL: [www.hhb.co.uk/university-of-surrey-opens-22-2-listening-room/](http://www.hhb.co.uk/university-of-surrey-opens-22-2-listening-room/) (cit. on p. 71).
- [70] Hyunkook Lee. «Investigation on the Phantom Image Elevation Effect». In: *139th Audio Engineering Society Convention* (Oct. 2015) (cit. on pp. 88, 91).