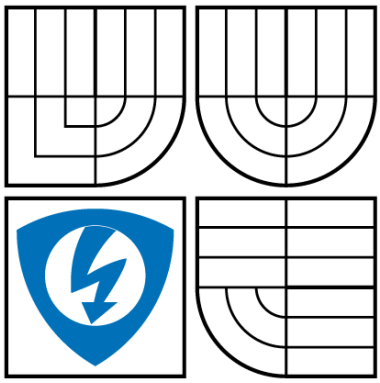
Brno University of Technology
Faculty of Electrical Engineering and Communication
Department of Telecommunication



LOCALIZATION AND RENDERING OF SOUND SOURCES IN ACOUSTIC FIELDS

by

Hasan Khaddour

A thesis submitted to
the Faculty of Electrical Engineering and Communication, Brno University of Technology,
in partial fulfillment of the requirements for the degree of Doctor.

PhD programme: Teleinformatics

Brno 2015

Thesis Supervisor:
Ing. Jiří Schimmel, Ph.D.
Department of Telecommunication
The Faculty of Electrical Engineering and Communication
Brno University of Technology
TECHNICKA 10
616 00 BRNO
Czech Republic

Abstract

This doctoral thesis deals with sound source localization and acoustic zooming. The primary goal of this dissertation is to design an acoustic zooming system, which can zoom the sound of one speaker among multiple speakers even when they speak simultaneously. The system is compatible with surround sound techniques.

In particular, the main contributions of the doctoral thesis are as follows:

1. Design of a method for multiple sound directions estimations.
2. Proposing a method for acoustic zooming using DirAC.
3. Design a combined system using the previous mentioned steps, which can be used in teleconferencing.

Keywords:
Multiple sound sources localization, DirAC, sound rendering, acoustic zooming.

Abstrakt

Disertační práce se zabývá lokalizací zdrojů zvuku a akustickým zoomem. Hlavním cílem této práce je navrhnout systém s akustickým zoomem, který přiblíží zvuk jednoho mluvčího mezi skupinou mluvčích, a to i když mluví současně. Tento systém je kompatibilní s technikou prostorového zvuku.

Hlavní přínosy disertační práce jsou následující:

1. Návrh metody pro odhad více směrů přicházejícího zvuku.
2. Návrh metody pro akustické zoomování pomocí DirAC.
3. Návrh kombinovaného systému pomocí předchozích kroků, který může být použit v telekonferencích.

Klíčová slova:
Lokalizace zdrojů zvuku, DirAC, reprodukce zvuku, akustický zoom.

DECLARATION

I declare that I have elaborated my doctoral thesis on the theme of *Localization and Rendering of Sound Sources in Acoustic Fields* independently, under the supervision of the doctoral thesis supervisor and with the use of technical literature and other sources of information which are all quoted in the thesis and detailed in the list of literature at the end of the thesis.

Brno: .....

.....
(Author's signature)

Faculty of Electrical Engineering
and Communication
Brno University of Technology
Purkynova 118, CZ-61200 Brno
Czech Republic
http://www.six.feec.vutbr.cz

PODĚKOVÁNÍ

Výzkum popsaný v této doktorské práci byl realizován v laboratořích podpořených z projektu SIX; registrační číslo CZ.1.05/2.1.00/03.0072, operační program Výzkum a vývoj pro inovace.

Brno ....
(podpis autora)

Acknowledgments

I would like to express my deepest gratitude to my supervisor Jiří Schimmel, MSc., Ph.D for his help and priceless advices. He helped me throughout my study with many problems and professional advancements, and he has been a constant source of encouragement and insight during my research.

I would like to thank all the members of the Department of Telecommunications at Brno University of Technology, and special thanks go to Tishreen University in Syria for help and support.

I would like to take this opportunity to thank all my friends whom I consider as a second family to me, and to thank unforgettable people in the Czech Republic who helped and supported me in many ways during my stay in this country.

Last but not least, my greatest thanks go to my parents, brother and sisters, for their unwavering love, infinite patience and encouragements. I highly appreciate everything they did and went through to keep me going.

Contents

Abbreviations

r	the radius of the head
θ	the azimuth
ΔD	the extra path the sound travels to reach the further ear
f	frequency
$H_{\|R\|}$	HRTFs for the right ear
$H_{\|L\|}$	HRTFs for the left ear
ϕ	the elevation
d	the distance from the sound source to the head
$p_{\|R\|}$	the value of pressure at the right ear
$p_{\|L\|}$	the value of pressure at the left ears
p_0	the sound pressure value at the centre of the head
t	time
$h_{\|R\|}$	HRIRs for the right ear
$h_{\|L\|}$	HRIRs for the left ear
$\hat{h_{\|R\|}}$	the measured HRIRs at the right ear
$\hat{h_{\|L\|}}$	the measured HRIRs at the left ear
h_0	the sound source impulse response
h_0^{-1}	the inverse filter of h_0
$y_f(n)$	the output of a beamforming array
Dl_m	the delay applied to a signal in a beamforming array
$x_m(n)$	the signal from the microphone number m
El	the output energy of a beamforming array
$R_{x_{m_1},x_{m_2}}$	the cross-correlation between the signals x_{m_1} and x_{m_2}
τ_m	the delay of arrival for that microphone number m
d_m	the distance between two microphones
c	the sound speed
f_{max}	the maximum frequency detected without aliasing problem
$Y_i(t)$	the energy measured by microphone m_i
K	the number of sound sources
$S_k(t)$	the energy emitted by the sound source number k
t_{ki}	the time delay between a sound source (k) and a microphone (i)
r_k	the coordinates of the sound source number (k)
r_i	the coordinates of the microphone (i)

a	an energy decay factor
g_i	the gain factor of the microphone number k
$\varepsilon_i(t)$	the cumulative effects of the modelling error
D_k	the time delay of arrival for different k
ς_k	an attenuation factor for different k
$b_k(t)$	a noise signal for different k
$s_1(t)$	a sound signal
$x_k(t)$	a sound signal for different k
k	the number of the microphone
g_k	the acoustic impulse response
$\hat{\tau}$	the estimated time delay of arrival
E	the expectation
$R_{x_1x_2}(t)$	the cross-correlation function
$G_{x_1x_2}(f)$	cross power spectral density
$G_{y_1y_2}(f)$	The cross power spectrum between the filter outputs
$\psi_g(f)$	the general frequency weighting
$R^{(g)}_{y_1y_2}(\tau)$	the cross-correlation between $y_1(t)$ and $y_2(t)$ when GCC is used
$R^{(p)}_{y_1y_2}(\tau)$	the cross-correlation between $y_1(t)$ and $y_2(t)$ when PHAT is used
$R^{(ML)}_{y_1y_2}(\tau)$	the cross-correlation between $y_1(t)$ and $y_2(t)$ when ML is used
$\psi_p(f)$	the weighting filter for PHAT method
$\hat{\tau}_p$	time delay of arrival estimated by PHAT method
$\psi_{ML}(f)$	the weighting filter for ML method
$\hat{\tau}_{MLp}$	time delay of arrival estimated by ML method
$\lvert\gamma_{12}(f)\rvert$	the magnitude squared coherence
D_m	the distance between two microphones
ϑ	the direction of arrival
G_{L1} , G_{L2}	gains factors applied to the loudspeakers
ϱ	the angle from the listener to the sound source
ϱ_0	the half angle between the loudspeakers
ϱ_s	the angle of the phantom sound source
ϱ_{ss}	the angle of the phantom sound source calculated by sine law
G_{Ls1}, G_{Ls2}	gain factors calculated using sine law
G_{Lt1}, G_{Lt2}	gain factors calculated using tangent law
ϱ_{st}	the angle of the phantom sound source calculated by tangent law
G_{n_L}	the gain applied to the loudspeaker n_L
N_L	the number of the loudspeakers
L_c, R_c, T_c, Q_c	C-format signals
X_a, Y_a, Z_a, W_a, R_a	signals of second-order Ambisonic
S_a, T_a, U_a, V_a	signals of second-order Ambisonic
S_i	a signal
α	an angle pointing to the sound source
β	an angle pointing to the sound source
G_{xi}, G_{yi} , G_{wi}	decoder coefficients
X_{a1}, Y_{a1}, W_{a1}	encoded audio signals
$\mathbf{G}_{2l}$	gain vector in 2D plane
$\mathbf{I}_m$, $\mathbf{I}_n$, $\mathbf{I}_k$	the vectors defining the direction of the loudspeakers m, n, k
$\mathbf{L}_{mn}$	the Cartesian unit vector in 2D
$\mathbf{P}$	the panning vector of the virtual source

$\mathbf{G}_{3l}$	gain vector in 3D plane
$\mathbf{L}_{mnk}$	the Cartesian unit vector in 3D
Z_0	the characteristic acoustic impedance of air
v	the particle velocity
E	the instantaneous energy density
p	the sound pressure
W	the B-format signal $w(t)$ in the STFT domain
I	the instantaneous intensity vector
Ψ	the diffuseness
$\Psi(n_1)$	the diffuseness in the time frame number n_1
$W(n_1)$	the B-format signal W in the time frame number n_1
$W_1(m_1)$, $W_2(m_1)$, $W_{3a}(m_a)$	window functions
$\mathbf{D}(n_1)$	the instantaneous direction of arrival
a_1, b_1, a_2, b_2, a_3, b_3	real variables
$g(n_a, i_a, k_a)$	gain factor
i_a	represents subband
n_a	represents time
k_a	the loudspeaker channel
$x(t)$	a continuous time-domain signal
$X(f)$	a signal in the continuous frequency-domain
f	frequency
$x(n)$	a discrete sequence of time-domain sampled values
n	the time sample
$X(m)$	a discrete Fourier transform (DFT) of the signal $x(n)$
m	the index of DFT output in the frequency domain
N	the number of samples
$X_h(\tau, \omega)$	short time Fourier transform of the signal $x(t)$
ω	the angular frequency
$h(t-\tau)$	the window function
τ	the time shift
$X_h[m, k]$	short time Fourier transform of the signal $x[n]$
k	the frequency bin
$G_x(\tau, w)$	the continuous Gabor transform of a signal
$G_h[m, n]$	the discrete Gabor transform of a signal
L	the length of the signal
a	the number of channel
$g[l-an]$	a window function
M	the decimation in time
$\mathbb{R}$	the set of the real numbers
A	a real variable
B	a real variable
σ	the standard deviation of the distribution
$g_{\mathrm{GA}}(t)$	A Gaussian window
$g_{\mathrm{re}}(t)$	rectangular window
$g_{\mathrm{tr}}(t)$	Bartlett window
$g_{\mathrm{ham}}(t)$	Hamming window
$g_{\mathrm{han}}(t)$	Hanning window
$g_{\mathrm{bl}}(t)$	Blackman window

$G_{re}(f)$	the Fourier transform of the rectangular window
ρ_t	the radius of the window in the time domain
ρ_f	the radius of the window in the frequency domain
Δf	the diameter of the window in the time domain
Δt	the diameter of the window in the time domain
q_d	the Ambisonic order
Q_d	the highest Ambisonic order
$x(t), y(t), z(t),\ w(t),$	B-format signals
$\overrightarrow{I}(t)$	the acoustic intensity
$p(t)$	the sound pressure
$\overrightarrow{v}(t)$	the velocity vector
$I_x,\ I_y,\ I_z$	components of the acoustic intensity
$\overrightarrow{x},\ \overrightarrow{y},\ \overrightarrow{z}$	the unit vectors in the 3D plane
c	the speed of the sound
ρ_0	the density of the air
f_l	the fluid compressibility
$I_x(t,f),\ I_y(t,f),\ I_z(t,f)$	components of the intensity vector
$X(t,f),\ Y(t,f),\ Z(t,f)\ ,\ W(t,f)$	Fourier transform of the B-format signals
$F(\alpha)$,$F(\beta)$	the number of the frequency bins
W_{rms}	root-mean-square value of the B-format signal $w(t)$
$b(n)_{\text{rms}}$	root-mean-square value of the noise signal b_n
$\overline{P_W}$	the average power of W_{rms}
$\overline{P_n}$	the average power of $b(n)_{rms}$
n_a	time frame
k_a	the loudspeaker channel
$g(m,n)$	a gain factor
g_{max}	the maximum gain factor
g_{min}	the minimum gain factor
Υ	the half of the angle in which we zoom the sound
S	the response rate
N	the total number of PB words
N_C	number of correctly recognized PB words
N_I	number of incorrectly recognized PB words

Miscellaneous Abbreviations

RADAR	radio detection and ranging
SONAR	sound navigation system
ITD	interaural time difference
ILD	interaural level difference
HRTFs	head related transfer functions
HRIR	head-related impulse response
RWPHAT	reliability-weighted phase transform
TDOA	time delay of arrival
DOA	direction of arrival
CC	cross correlation method
GCC	generalized cross correlation method
PHAT	phase transform method
ML	maximum likelihood method
SCOT	the smoothed coherence transform
SNR	signal to noise ratio
IQR	the interquartile range
ASIO	audio streaming Input/output
DFT	discrete Fourier transform
STFT	short-term Fourier transform
DirAC	directional audio coding
VBAP	vector base amplitude panning
2D	two dimensional
3D	three dimensional
WFS	wave field synthesis
SIRR	spatial impulse response rendering
IC	interaural coherence cues
MDCT	modified discrete cosine transform
RMS	root-mean-square value
PEASS	perceptual evaluation methods for audio source separation
PEMO-Q	Perception Model for Quality assessment
OPS	perceptual score
TPS	target-related perceptual score
IPS	interference-related perceptual score
APS	artifacts-related perceptual score
Pb	phonetically balanced
CACR	Chance-Adjusted Correct Response rate

List of Figures

List of Tables

Introduction

This dissertation combines two important parts of the acoustic discipline; namely, localization and rendering of sound sources. Both parts have been intensively investigated in the past decade. Although those two fields, localization and rendering, are studied separately, they are connected to each other. Where the remarkable ability of human beings to observe the surround environments can be seen as a common factor between these two fields i.e., the same cues, which the people use to localize the sound sources, are used in sound source localization methods and also are attempted to be re-created in the surround sound systems.

Surround sound systems have been the focus of attention for many years. New methods for spatial sound rendering are constantly appearing. They can be mainly used in cinemas, music, and video games. The spatial sound techniques not only provide the possibility of orientation, but also improve the pleasure of the listening to the sound. The aim of the modern spatial sound rendering methods is to bring the pleasure and the sensation of the musical places, such as the theaters and the opera houses, to the domestic sound systems.

The source localization methods are also used in many applications. They are primarily used in RADAR (Radio Detection and Ranging) and underwater SONAR (Sound Navigation and Ranging). Sound source localization in the air, however, is a quite new application of this technique compared to SONAR. Even so, it found its role in many applications in the modern technologies, for instance, in surveillance, speech recognition and teleconferencing.

This work deals with both localization and rendering methods, it aims at proposing a new method for sound source direction estimation, combining this method with an acoustic zooming technique and designing a new system that provides the possibility of sound source direction estimation and acoustic zooming in the same time. It should be noted that evaluation is an important part of each system. Therefore, throughout this work, some sound localization methods were evaluated and compared to each other, listening tests were also performed in order to compare the performance of existing sound spatial rendering techniques, and to choose the rendering technique with the best results to be used in the proposed acoustic zooming system.

Structure of the Thesis

The rest of the dissertation is organized into seven chapters as follows:

1. Chapter 1 *Background and State of the Art*: Introduces the reader to the theoretical

background and surveys the current state of the art of the sound source localization methods, and also provides an overview of surround sound rendering techniques.

2. Chapter 2 *Time-Frequency Representation*: Surveys some time-frequency representation methods and introduces the reader to the concept of the accuracy and optimal localization in time and frequency domain.

3. Chapter 3 *Objectives of Dissertation*: Presents briefly the goals of this work.

4. Chapter 4 *A Comparison and Evaluation of Localization and Rendering Methods*: Introduces simulation and experimental results of some of sound source localization methods and also investigates the performance and the accuracy of several sound rendering techniques.

5. Chapter 5 *Estimation of the Direction of Arrival of Multiple Speakers*: Presents a method for multiple sound sources direction estimation and provides simulation and experimental results of this method.

6. Chapter 6 *Acoustic Zooming*: Proposes a system for direction estimation and acoustic zooming based on DirAC and shows the results of the listening tests for this system.

7. Chapter 7 *Conclusion*: Summarizes the results of our research, suggests possible topics for further investigation, and concludes the dissertation.

Chapter 1

Background and State of the Art

This chapter is devoted to sound source localization methods and surround sound systems. Background and state of the art about the sound source localization methods are introduced in the first part of this chapter, whereas surround sound systems are presented in the second part.

1.1 Spatial Hearing and Sound Source Localization Methods

In this section we study the spatial hearing mechanism. The cues, which are used by the human beings to localize the sound sources, are presented as well. This section also discusses the sound source localization methods and introduces the techniques invented for this goal. It also presents the relation between the spatial hearing mechanism and sound source localization methods.

1.1.1 Spatial Hearing Mechanism

Needless to say, it is very important to study and understand the mechanism of the spatial hearing in order to design a good spatial sound rendering method, since the ultimate objective of sound spatial rendering is to give the illusion of directionality and spaciousness to the listener. Even more, the cues, that people use for spatial hearing, could be used as well as cues for sound source localization methods.

In the real environment, the sound reaching the ears is not a pure sound coming directly from the sound source, it also contains the sound reflecting to the ears from the surrounding environment e.g., walls, roof, floor, furniture etc. Therefore, the duty of the auditory system is not only to localize the direct sound from the sound source, but also to distinguish it from the reflected sound signals. However, these reflections enrich the sound and make it *colorful*.

The people's ability of localizing sound sources comes from some physical phenomena that happen when the sound signal travels towards the ears. This ability relies primarily

on the tiny differences in the time of arrival and on the differences in the intensity of the sound signals at each ear. The cues, which the people use to localize the sound sources, can be divided into [1]:

1. Interaural time difference (ITD).
2. Interaural level difference (ILD).
3. Monaural spectral cues [2].
4. The changing in cues that happens when the listener moves his head.
5. The ratio between the direct and the reverberant energy.
6. The familiarity with the sound source.

Although the primary cues for estimating the so-called lateral angle are both ITDs and ILDs, all of the presented cues are used for optimum sound source localization. However, as we will see later, some of these cues are dominant in some cases more than the others. In the next paragraph, the most used cues in the spatial hearing will be introduced.

1.1.1.1 Interaural Time Difference

Interaural time difference (ITD) is the difference in the time of arrival of sound signal at the two ears. It should be noted that this difference in time does not exceed a small fraction of a millisecond [3]. A model given by Woodworth in 1938 explained how interaural path-length differs [4]. This model assumed that the additional path, which the sound travels to the further ear, can be divided into two segments: a straight path ($r \sin\theta$), and a curved path around the skull ($r\theta$), see Figure 1.1. Given that ΔD is the whole extra path that the sound has to travel to reach the more distant ear from the sound source, we obtain

$$\Delta D = r \sin\theta + r\theta = r(\sin\theta + \theta) \tag{1.1}$$

where r is the radius of the head and θ is the azimuth of the sound source measured in radians.

As can be seen from eq. (1.1), ITD differs depending on the angle of the sound source. We can find the maximum possible ITD using eq. (1.1) to be 673 μs, this maximum occurs when the sound source is at the angle $\pi/2$ radians regarding the listener i.e. the sound source is to one side or another beside the head of the listener, and assuming that the sound speed is $344\frac{m}{s}$. Thanks to this difference in ITDs, the brain is able to *calculate* these angles, and localize the sound sources.

However, this model assumed that the ITD is irrelevant to the frequency of sound, which was proven to be incorrect, especially for low frequencies [5]. The thresholds, at which the people are able to detect the ITD, differ depending on the sound frequency. In [6] it was shown that the discrimination thresholds for ITD may be less than $10\,\mu s$ when pure sound tones in the range $700 - 1000\,\text{Hz}$ were used. The article presented in [7]

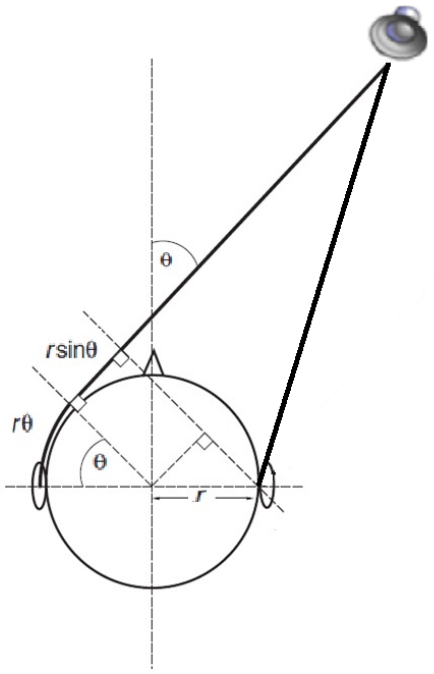


Figure 1.1: Interaural time difference (Woodworth' model).

showed the dependency of the thresholds for detection of ITDs on frequencies. For instance, it is claimed that the ITD was not measurable at 1500 Hz and above, and the threshold reached its minimum ($10\mu s$) at 1000 Hz. Hence, ITD plays the major role of sound source localization at the low-frequency hearing [8].

1.1.1.2 Interaural Level Difference

Interaural level difference (ILD) is the difference in the intensity of the sound signal when it reaches the eardrums. This difference is caused by the attenuation during the travelling in the air.

In the nineteenth century, Lord Rayleigh investigated the spatial hearing mechanism. He presented his theory in [9], which assumed that the low frequencies are localized depending on ITD and the high frequencies are localized depending on ILD, that was also presented in [10].

At high frequencies, a head's diameter becomes greater than the half length wavelength of the sound, causing a sound shadow. Therefore, head's shadow attenuates the sound signal in the ear opposite the sound source, and ILD becomes detectable at high frequency starting at 1000 Hz and it increases with frequency [10].

In the frequency range from 1500 to 2000 Hz, both ILD and ITD are not detectable enough [11]. Therefore, ambiguity arises in the mentioned frequency range. Fortunately, most natural sounds contain frequencies outside this range, which enables to localize them using both ITDs and ILDs.

Recalling that ITDs and ILDs happen because of the difference in path from the sound source to each ear, ITDs and ILDs become theoretically zero when the sound source is in the mid line (equidistant between the ears). In this case, the sound locations cannot be determined depending on the binaural cues. Although ILDs and ITDs give important information about the sound source positions, whether the sound comes from left or right (localization in the horizontal plane), they do not provide information about the vertical coordinates of the sound source. Even more, changing the sound source positions on the circumference of a circle, which has a constant diameter and a center on the half line between the ears, produces identical ILDs and ITDs. The area, where ILDs and ITDs are identical, is called the *cone of confusion* [12]. As can be seen in Figure 1.2, the sound coming from points **A** and **B** would have theoretically identical ITDs and ILDs, which produces a confusion about the location of these sound sources, unless there are other cues. In such case, the problem can be solved by moving the head a little bit because interaural differences will dynamically change with the movement.

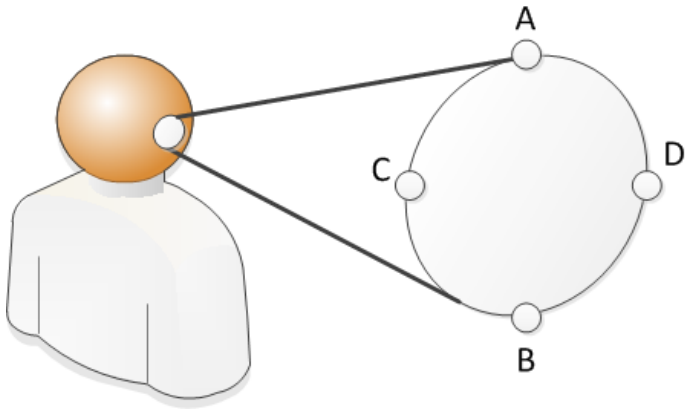


Figure 1.2: Cone of confusion.

1.1.1.3 Head-Related Transfer Functions

Head-related transfer functions (HRTFs) are the transfer functions from the source of the sound to the ear canal [13]. HRTFs show how the sound signals are affected by the whole path from the sound source to the ears especially by the listener's head, shoulders, torso and pinnae [14]. This effect happens due to the reflection and diffraction of the signal. Since HRTFs are the spectral filtering of a sound signal before it reaches the ear drum, they are expressed in frequency domain. The inverse Fourier transform of HRTFs is called head-related impulse response (HRIR) [1], [15]. It should be noted that HRTFs are a function of azimuth, elevation and frequency. They are individual for each person, and they differ for each ear and they also differ with angles in both vertical and horizontal

plane. HRTFs can be expressed for each ear as [15]

$$H_{|L|}(d,\theta,\phi,f) = p_{|L|}(d,\theta,\phi,f)/p_0(d,f)$$
$$H_{|R|}(d,\theta,\phi,f) = p_{|R|}(d,\theta,\phi,f)/p_0(d,f) \tag{1.2}$$

where $H_{|R|}$ and $H_{|L|}$ are the HRTFs for the right and left ear, respectively, d is the distance from the sound source to the head, θ is the azimuth in the range from 0° to 360°, ϕ is the elevation with variation range from $-90°$ to 90°, f is the frequency, $p_{|R|}$ and $p_{|L|}$ are the value of acoustic pressure at the right and the left ears and p_0 is the sound pressure value at the position of the center of the head when the subject is absent.

As mentioned earlier, HRIRs are the inverse Fourier transform of HRTFs. Hence, they can be derived from eq. (1.2) as [15]

$$h_{|L|}(d,\theta,\phi,t) = \hat{h_{|L|}}(d,\theta,\phi,t) * h_0^{-1}(d,t)$$
$$h_{|R|}(d,\theta,\phi,t) = \hat{h_{|R|}}(d,\theta,\phi,t) * h_0^{-1}(d,t) \tag{1.3}$$

where t is the time, $\hat{h_{|R|}}$ and $\hat{h_{|L|}}$ are the measured HRIRs at the right and the left ears, respectively, h_0 is the sound source impulse response which is measured when the subject was absent, h_0^{-1} is the inverse filter of h_0 and $*$ is the convolution operator.

1.1.2 Sound Source Localization

Sound source localization methods were intensively investigated over the past decade. The traditionally dominant application of this discipline is underwater acoustic or so-called SONAR (sound navigation system) [16]. However, this discipline found its applications in the era of telecommunications and modern science. This includes, but not limited to, interaction between the human and the machines [17], surveillance, speech recognition [18], video conferencing [19] and many other applications.

As already mentioned, people are capable of localization the sound sources depending on some physical phenomena. The existing sound source localization methods try somehow to mimic the natural ability of mammals to localize the sound sources. They can be loosely categorized into two groups:

1. Active sound source localization methods: The system renders an acoustic wave which will be reflected by the targets and then detected by the sensors [20].

2. Passive sound source localization methods: The system listens to the sound coming from the sound source, and then localizes it.

However, the sound source localization methods can be categorized from a different point of view. Depending on the way that used for estimating the sound source localization, they can be divided into:

1. Steered beamforming based methods.
2. Energy based localization methods.
3. Time difference of arrival based methods.

1.1.2.1 Steered Beamforming Based Methods

Beamforming techniques can be used in many disciplines, they are used for enhancing signals from a desired direction, detecting the signals, and also for estimating the direction of arrival of a signal.

The one can look at the beamforming techniques as a spatial filtering, where the signals coming from wanted direction are enhanced and the noise including reverberation and unwanted signals are attenuated. The simplest beamforming technique is so-called delay-and-sum. This technique can be seen as two processes: synchronization process and weight-and-sum process, see Figure 1.3. Whereas the former delays or (advances) the output of the sensors to synchronize the wanted signals coming from a specific direction, the later aligns the resulted output signals and then sums them producing an output with an enhanced wanted signal [21].

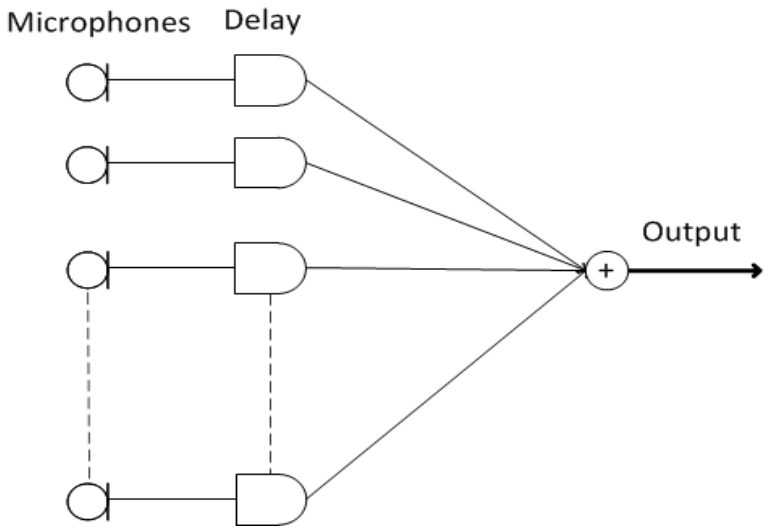


Figure 1.3: Delay-and-sum beamforming.

The idea of sound source localization using this methodology is to steer the beamform in all directions and look for the maximum output of the sensor array [22]. Assuming we have M microphones located at different positions, the output of delay-and-sum beamformer is then given as

$$y_f(n) = \sum_{m=0}^{M-1} x_m(n - Dl_m) \tag{1.4}$$

where $y_f(n)$ is the output, $x_m(n)$ is the signal from the microphone number m and Dl_m is the delay that is applied to this microphone.

According to [22], the output energy of this array can be calculated as a sum of cross-correlation between the microphone pairs

$$El = Kl + 2\sum_{m_1=0}^{M-1}\sum_{m_2=0}^{m_1-1} R_{x_{m_1,x_{m_2}}}(\tau_{m_1} - \tau_{m_2}) \tag{1.5}$$

where $Kl = \sum_{m_1=0}^{M-1}\sum_{m_2=0}^{L-1} x_m^2(n-\tau_m)$ and it can be ignored because it is nearly constant with respect to the τ [22], $x_m(n)$ is the signal from the m^{th} microphone, $R_{x_{m_1,x_{m_2}}}$ is the cross-correlation between the signals x_{m_1} and x_{m_2} and τ_m is the delay of arrival for that microphone.

It can be seen from eq. (1.5) that the output achieves its maximum when the microphone signals are in phase and they are added constructively. In this case, the beamform indicates the direction of the sound source. In order to avoid spatial aliasing, the distance between the microphones should be less than the half of the shortest sound wave that we want to capture, this applies that the distance between the microphones $d_m \leq \frac{c}{f_{max}}$ where c is the sound speed, and f_{max} is the maximum frequency detected without an aliasing problem.

However, the energy peaks using this method are wide, which causes a poor resolution [23]. Moreover, source response overlap appears in case we have multiple sound sources [22]. Other disadvantage of this technique comes from the big number of the needed microphones. For instance, a microphone array of eight microphones is used in [22].

However, the accuracy and the performance of this technique can be improved. For instance, in [24] they use a steered beamformer based on the reliability-weighted phase transform (RWPHAT) combined with a particle filter-based tracking algorithm using an array of 8 microphones.

1.1.2.2 Energy Based Methods

Energy based localization methods are inspired by a physical fact that sound signal is attenuated while it travels in the air to reach the further microphone. The same principle is mentioned when we talked about ILD, see Section 1.1.1.2. By using multiple sensors (microphones) distributed in a certain way, the relation between decrease of the sound energy and the position of the sound source can be estimated. Even though the principle is easy, so many factors should be taken into account, for instance, the surrounding environment, the shape of the source, the distribution of the sensors, and the frequency of the sound.

In case we have multiple sound sources and multiple microphones, the energy measured by microphone m_i can be written as [25]

$$Y_i(t) = g_i \sum_{k=1}^{K} \frac{S_k(t - t_{ki})}{|r_k(t - t_{ki} - r_i|^a} + \varepsilon_i(t) \tag{1.6}$$

where K is the number of sound sources, $S_k(t)$ denotes the energy emitted by the sound source number k, t_{ki} is the time needed for the sound propagation between the sound source number (k) and the microphone (i), r_k and r_i denotes the coordinates of the sound source number (k) and the microphone (i), respectively, a is an energy decay factor, g_i is the gain factor of the microphone number k and $\varepsilon_i(t)$ is the cumulative effects of the modelling error.

Assuming that we have two microphones, the ratio of the measured energies defines a set of points, where the sound source could be existed. This set has a shape of a circle. Adding a third microphone creates a new circle, at which the sound source could be. Crossing these two circles decrease the possible locations of the sound source into two points.

Another acoustic energy decay model was given in [26]. A disadvantage of the energy based localization methods is that they use a large number of low-cost, low-power sensors [27].

In acoustic, the energy based method can work only for sources that are close to the sensors. The distant sources have almost plane wavefront which has intensity independent of distance, which means that difference in sound pressure measured by individual sensors is bellow measurement precision and uncertainties. There is also attenuation in the air but it is negligible when sensors are not distant enough (several meters).

1.1.2.3 Time Delay of Arrival

Among the above mentioned categories, time delay of arrival (TDOA) is the most used methodology for the passive sound source localization. This methodology is inspired by ITD, see Section 1.1.1.1. For TDOA estimation in two-dimensional plane, two sensors (microphones) at least are needed to pick-up the sound signal considering only the space in front of microphones, and at least three microphones are needed in the three dimensional plane.

Signal Model

Time delay of arrival estimation is highly related to the environment's conditions. When we talk about TDOA, two different environments are considered i.e., free-field environment and reverberant environment. Whereas the former environment ensures that the microphones receive the direct-path signals only, in the latter, the microphones receive both the direct-path signals and the reverberant-path signals.

Even more, the problem of time delay of arrival can bee seen from two different points depending on the ratio between the distance of the sound source from the microphones and the dimensions of the used microphone array. Thus, the sound source can be determined to be in the array's far-field or in the array's near-field.

When the sound source is located in the array's far-field, only the direction of arrival of the sound can be estimated depending on the time delay of arrival. However, when the sound source is considered to be in the array's near-field, it is possible to estimate both its direction and location depending on the time delay of arrival. Figure 1.4 illustrates

the single-source near-field case, given a single sound source producing a sound signal $s(t)$, and two microphones m_1 and m_2 and the distance between them is D_m. As a matter of convention, we will choose the microphone m_1 to be the reference microphone.

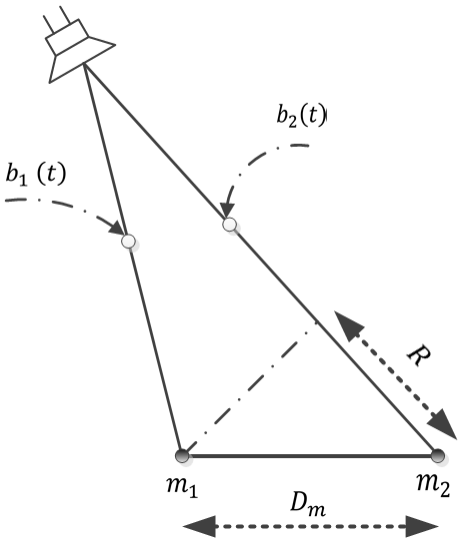


Figure 1.4: Time delay of arrival

The sound signals $x_1(t)$ and $x_2(t)$, which are captured by the microphones, can be written as

$$\begin{aligned} x_1(t) &= s_1(t) + b_1(t) \\ x_2(t) &= \varsigma s_1(t - D_1) + b_2(t) \end{aligned} \tag{1.7}$$

where ς and D_1 are the attenuation factor and the time delay of arrival of the signal $x_2(t)$ with respect to the signal $x_1(t)$, respectively, $b_1(t)$ and $b_2(t)$ are additive noise and $s_1(t)$ is the sound signal which is rendered by the sound source.
The model can be generalized in case of a single sound source and n used microphones as

$$x_k(t) = \varsigma_k s_1(t - D_k) + b_k(t) \tag{1.8}$$

where ς_k and D_k are the attenuation factor and the time delay of arrival of the signal $x_k(n)$ which is picked-up by the microphone m_k with respect to the microphone m_1, and $k = 1...n$ is the number of the microphone.

However, we did not take into account the multi-path effect in the last model, neither the possibility of the correlation between the noise signals and the sound signal. Therefore, another model is needed to describe the real environment with reverberation and correlation between the noise signals and the sound. Such model was given in [28], and it is expressed as

$$x_k(t) = g_k * s(t) + b_k(t) \tag{1.9}$$

where $*$ is the convolution operator, g_k is the acoustic impulse response between the sound source $s(t)$ and the microphone number k.

In practical cases, the room impulse response changes with position of the source (also with position of the microphone but we expect that it is constant).

Time Delay of Arrival Estimation Algorithms

In this section, several time delay estimation algorithms are described, including cross-correlation method (CC), generalized cross-correlation method (GCC), phase transform (PHAT) and maximum likelihood (ML).

1. **Cross-Correlation Algorithm**
   Cross-correlation method is the simplest way to estimate the time delay of arrival between two signals. Assuming we have a signal model as described in eq. (1.7), the time delay D_1 can be calculated as the value τ for which the cross-correlation function $R_{x_1x_2}$ is maximized [29], where the cross-correlation function is given as

   $$R_{x_1x_2}(t) = E[x_1(t)x_2(t-\tau)] \tag{1.10}$$

   where E is the expectation. Then the estimated time delay is given as

   $$\hat{\tau} = \arg\max_{\tau} R_{x_1x_2}(t). \tag{1.11}$$

2. **Generalized Cross Correlation Algorithm**
   The generalized cross-correlation (GCC) algorithm was proposed by Knapp and Carter in 1976 [29]. It is one of the most used algorithm for estimation the time delay of arrival. As in CC, the time delay of arrival is estimated as the lag time that maximizes the cross-correlation function. However, the microphones' signals are filtered first, see Figure 1.5. Using the filters, when they are chosen properly, helps with estimating the time delay of arrival. As can be seen from Figure 1.5, the microphone signals are filtered first by the filters (H_1 and H_2), the resultant signals are then delayed and multiplied, they are then integrated for a variety of delays until peak output is obtained [29].

   Assuming we have a free-field two-microphones model presented in eq. (1.7), and the sound signal $s_1(t)$ and the noise signals $b_1(t)$ and $b_2(t)$ are mutually independent. Under these conditions, the generalized cross-correlation between the signals $x_1(t)$ and $x_2(t)$ is the cross-correlation between the signals $y_1(t)$ and $y_2(t)$ and it is related to their cross power spectral density as [29]

   $$R^{(g)}_{y_1y_2}(\tau) = \int_{-\infty}^{\infty} \psi_g(f) G_{x_1x_2}(f) \mathrm{e}^{j2\pi f\tau} \,\mathrm{d}f \tag{1.12}$$

   where $\psi_g(f)$ is the general frequency weighting and it is given as

   $$\psi_g(f) = H_1(f)H_2^*(f) \tag{1.13}$$

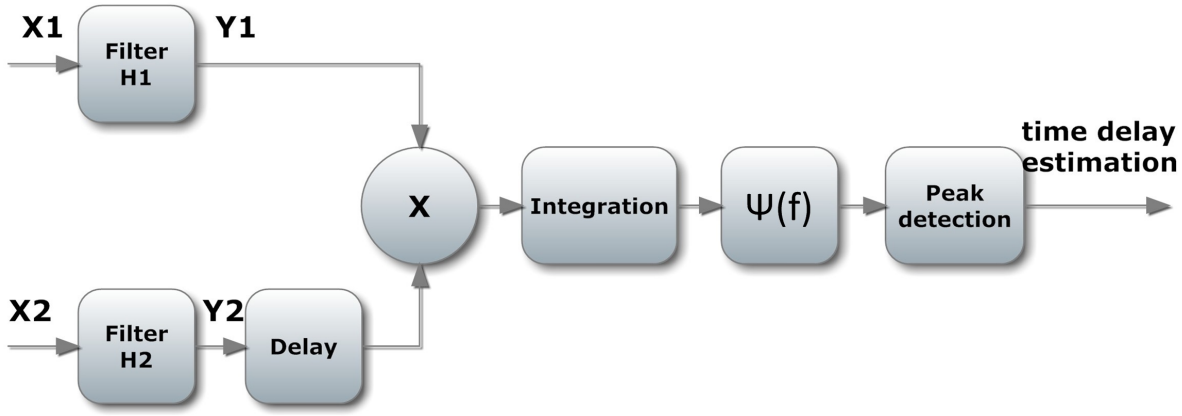


Figure 1.5: Generalized cross-correlation algorithm's diagram.

where $*$ denotes the complex conjugation. The cross power spectrum between the filter outputs is expressed as [29]

$$\begin{aligned} G_{y_1y_2}(f) &= H_1(f)H_2^*(f)G_{x_1x_2}(f) \\ G_{y_1y_2}(f) &= \psi_g(f)G_{x_1x_2}(f). \end{aligned} \tag{1.14}$$

However, in practice only a value $\hat{G}_{x_1x_2}(f)$ can be estimated from generalized $G_{x_1x_2}(f)$. Therefore, the eq. (1.12) can be written as [29]

$$\hat{R}_{y_1y_2}^{(g)}(\tau) = \int_{-\infty}^{\infty} \psi_g(f)\hat{G}_{x_1x_2}(f)\mathrm{e}^{j2\pi f\tau}\,\mathrm{d}f. \tag{1.15}$$

Choosing the weighting filter $\psi_g(f)$ affects the time delay estimation performance. For instance, we can get the classical cross-correlation method by choosing the weighting filter equal to one. Other methods are derived from generalized cross-correlation family by choosing different weighting filters, for instance, the Roth processor, the phase transform (PHAT), the smoothed coherence transform (SCOT), Eckart filter and the Hannan-Thomson (maximum likelihood) filter. The role of the weighting function is to give sharper and larger peak in the cross-correlation between the signals obtained from the microphones.

The GCC methods are usually used for time delay of arrival estimation due to their accuracy and low computational requirements [30]. In the following paragraphs, we describe two methods that are derived from generalized cross-correlation (GCC) algorithm by choosing different weighting filter; namely, phase transform and maximum likelihood.

(a) **Phase Transform Method**
Considering eq. (1.12), it is clear that the time delay of arrival information is

presented in phase more than in amplitude of the cross-spectrum [21]. Therefore, we can choose the weighting filter to be equal to [29]

$$\begin{aligned}\psi_p(f) &= \frac{1}{|H_1(f)H_2^*(f)|} \\ &= \frac{1}{|G_{x_1x_2}(f)|}.\end{aligned} \tag{1.16}$$

Applying that to eq. (1.15), we get [29]

$$\hat{R}_{y_1y_2}^{(p)}(\tau) = \int_{-\infty}^{\infty} \frac{\hat{G}_{x_1x_2}(f)}{|G_{x_1x_2}(f)|} \mathrm{e}^{j2\pi f\tau}\,\mathrm{d}f. \tag{1.17}$$

In the ideal case, when the noise signals are not correlated to each other or to the sound signals, and when $\hat{G}_{x_1x_2}(f) = G_{x_1x_2}(f)$, we get [29]

$$\begin{aligned}\hat{R}_{y_1y_2}^{(p)}(\tau) &= \int_{-\infty}^{\infty} \frac{\hat{G}_{x_1x_2}(f)}{|G_{x_1x_2}(f)|} \mathrm{e}^{j2\pi f\tau}\,\mathrm{d}f \\ &= \int_{-\infty}^{\infty} \mathrm{e}^{j2\pi fD} \mathrm{e}^{j2\pi f\tau}\,\mathrm{d}f.\end{aligned} \tag{1.18}$$

The time delay of arrival can be then estimated as

$$\hat{\tau}_p = \arg\max_{\tau \in D} \hat{R}_{y_1y_2}^{(p)}(\tau). \tag{1.19}$$

As a result, PHAT algorithm has a better performance than CC method, and it is computationally efficient. Even more, it is able to avoid spreading of the peak of the correlation function [31]. The PHAT filter has the effect of removing all energy content from the cross spectrum.

(b) **Maximum Likelihood Method**
The maximum likelihood method is estimated from the generalized cross-correlation method by choosing the weighting function as [29]

$$\psi_{ML}(f) = \frac{1}{|G_{x_1x_1}(f)|} . \frac{|\gamma_{12}(f)|^2}{[1 - |\gamma_{12}(f)|^2]} \tag{1.20}$$

where the magnitude squared coherence $|\gamma_{12}(f)|$ is given as [29]

$$\gamma_{x_1x_2}(f) = \frac{\hat{G}_{x_1x_2}(f)}{\sqrt{G_{x_1x_1}(f)G_{x_2x_2}(f)}}. \tag{1.21}$$

Applying that to eq. (1.12) we get [29]

$$R_{y_1y_2}^{(ML)}(\tau) = \int_{-\infty}^{\infty} \hat{G}_{x_1x_2}(f) \frac{1}{G_{x_1x_1}(f)|} . \frac{|\gamma_{12}(f)|^2}{[1 - |\gamma_{12}(f)|^2]} \mathrm{e}^{j2\pi f\tau}\,\mathrm{d}f. \tag{1.22}$$

The time delay of arrival is then estimated as the value which achieves a peak in eq. (1.22), and it is expressed as [29]

$$\hat{\tau}_{ML} = \arg\max_{\tau \in D} \hat{R}^{(ML)}_{y_1 y_2}(\tau). \tag{1.23}$$

The ML weighting function attenuates the signals fed into the correlator in the spectral region, where the SNR is the lowest value which improves the accuracy of this method. However, in the above mentioned situation, it was assumed that the signals $x_1(t)$, $x_2(t)$ and $b(t)$ are Gaussian.

1.2 Sound Rendering Techniques

A part of this work is devoted to spatial sound rendering. Thus, this section presents a short historic overview of surround sound techniques and a state of the art of the currently used methods.

Although the surround sound systems have grown enormously over the last decades, the early systems are as important as the modern ones. Even more, the old systems are still in use, thanks to their simplicity, and their good performance, specially for non-demanding listeners.

In order to create the perceptual feeling of the spatial hearing sound inside a loudspeaker setup, three localization cues should be created. These cues are:

- The azimuth angle in the horizontal plane.
- The elevation angle in the vertical plane.
- The distance of the sound source.

Creating these three cues properly can achieve the illusion of spatial hearing to the listener [32]. However, the reliable reproduction sound systems should also achieve the following conditions [33]

1. All audible frequency should be included in the frequency range of the system.
2. The dynamic range must be large enough in order to prevent distortion.
3. The reproduced sound must contain the original spatial sound patterns.
4. The system should approximate the reverberation characteristics of the original sound in the reproduced sound.

1.2.1 Early Surround Sound Systems

Although the surround sound systems have been a goal of audio engineers since the early part of the twentieth century, commercial restrictions were the reason of using only two channels feeding two loudspeakers in the front of the listeners [34]. Thus, two-channel stereophonic reproduction is the most used recording and rendering approach of domestic use providing *some spatial content*.

1.2.1.1 Loudspeaker Stereo

Two-loudspeaker stereo is mostly used in domestic environment thanks to its features and prices. It can provide a good phantom image for the listener in the sweet spot, and it is relatively cheap to implement.

The universal optimum configuration for two-loudspeaker stereo is a equilateral triangle, with the loudspeakers located in the two vertexes of the triangle and the listener seated in the third vertex [35], [32]. Therefore, the loudspeakers are located at $\pm 30°$ from the sweet spot, and their membranes should face it.

Blumlein suggested using a different level in each loudspeaker, in order to reproduce a phantom sound source using two loudspeakers [36]. However, the gain, applied to the signals in this case, is multiplied directly by the sound signals. Considering that this gain will influence the amplitude of the signal rather than its intensity, and that the amplitude of the sound is proportional to the square of its intensity, the gains applied to the loudspeakers should achieve [32]

$$G_{L1}{}^2 + G_{L2}{}^2 = \text{constant} \tag{1.24}$$

where G_{L1} and G_{L2} are the gains applied to the first and the second loudspeaker, respectively. This requirements can be satisfied by using the trigonometric identity

$$\sin^2 \varrho + \cos^2 \varrho = 1 \tag{1.25}$$

where ϱ is the angle from the listener to the sound source in the horizontal plane. Thus, the gain of one loudspeaker can be chosen to be proportional to $\sin \varrho$ and the second gain to $\cos \varrho$ [32].

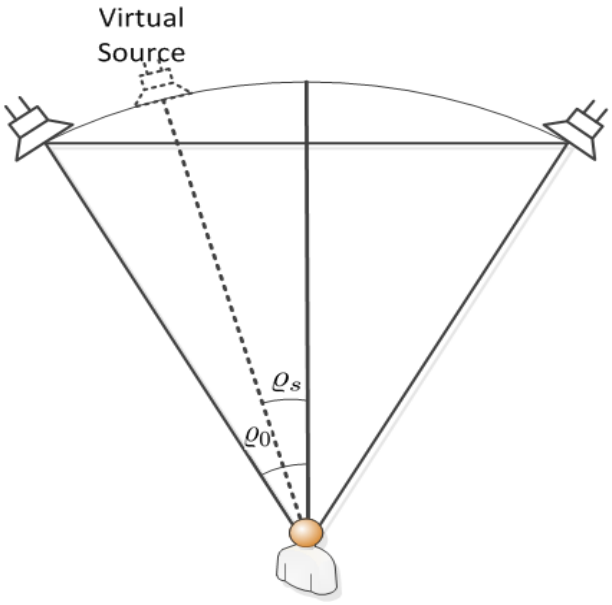


Figure 1.6: Two-channel loudspeaker setup.

Using this way, it is possible to reproduce the sound source to be in any position between the loudspeakers by feeding the loudspeakers with the same signal with difference in amplitude or with small delay in time. Several laws can be used in order to calculate the gain fed to each loudspeaker. The *sine law* states the ratio between the signals fed to each signal as a function of the angle the sound is supposed to be coming from as [37], [38]

$$\frac{\sin \varrho_{ss}}{\sin \varrho_0} = \frac{G_{Ls1} - G_{Ls2}}{G_{Ls1} + G_{Ls2}} \tag{1.26}$$

where G_{Ls1}, G_{Ls2} are gain factors applied to the first and second loudspeaker, respectively, calculated using sine law, ϱ_0 is the half angle between the two loudspeakers regarding the position of the listener, ϱ_s is the angle of the phantom sound source located between the two loudspeakers and ϱ_{ss} denotes this angle when it is calculated using sine law, and $-\varrho_0 \leq \varrho_{ss} \leq \varrho_0$, see Figure 1.6. However, this law has its limitation, where the listener is supposed to be pointing directly forward, and the frequency does not exceed 500 Hz [32].

Other panning laws can be used to calculate the gains applied to each loudspeaker. The *tangent law*, for instance, calculates the gains as [39]

$$\frac{\tan \varrho_{st}}{\tan \varrho_0} = \frac{G_{Lt1} - G_{Lt2}}{G_{Lt1} + G_{Lt2}} \tag{1.27}$$

where G_{Lt1}, G_{Lt2} are gain factors applied to the first and second loudspeaker, respectively, calculated using tangent law, and ϱ_{st} is the angle of the phantom sound source located between the loudspeakers, and $-\varrho_0 \leq \varrho_{st} \leq \varrho_0$.

Both eq. (1.26) and eq. (1.27) do not provide enough information in order to calculate the gain for each loudspeaker. Therefore, another relation is needed for calculating the gains. Eq.(1.24) can be used as an additional equation, which can solve this problem. However, it can be generalized in another form in case of involving more loudspeakers, in order to keep the intensity of the sound constant. The generalized version of eq. (1.24) can be written as [32]

$$\sqrt[p]{\sum_{n_{\mathrm{L}}=1}^{N_{\mathrm{L}}} G_{n_{\mathrm{L}}}^{p}} = 1 \tag{1.28}$$

where the value of p depends on the listening room acoustic, whereas $p = 1$ in the anechoic room listening, $p = 2$ in the real room listening [40], $G_{n_{\mathrm{L}}}$ is the gain applied to the loudspeaker n_{L} and N_{L} is the number of the loudspeakers.

1.2.1.2 Three-Channel System

Three-channel system consists of three loudspeakers; namely, left, right and center. The loudspeakers are located in the front of the listener and arranged equidistantly. Using the center loudspeaker enables the outer loudspeakers to be placed further out to the sides. However, in order to be compatible with the two-channel loudspeakers, the angle between the outer loudspeakers is chosen to be 60°. The three-channel stereo has some advantages

over two-channel stereo, it has a wider front sound stage, and it enables a wider range of listening positions. Even more, it provides a clearer dialog in the middle of the screen in cinemas [35].

1.2.1.3 Four-Channel System

The four-channel system, sometimes called 3-1 system, has a fourth *surround* channel in additional to the three channels presented in the three-channel systems. The additional channel feeds one loudspeaker or more, located in the backside or to the sides of the listener [34]. Thus, the additional channel can provide the wrap-around effects [35]. However, this system is not capable of creating the sensation of envelopment of spaciousness [35].

1.2.1.4 Five-Channel System

The five-channel system is widely used in surround sound applications. It overcomes the disadvantages of four-channel system and provides the sensation of the room ambiance. Three channels of these five-channel system are used as in three-channel system, the additional two channels are used to create the supporting ambiance [34].

This system is also called 5.1 system, where an additional channel is used to transmit low-frequency effect, and it has the term ".1" channel [35].

The loudspeakers layout is defined in [41]. The three front channels are located as in three-channel system i.e., center loudspeaker at 0° and two loudspeakers located at $\pm 30^\circ$, and the surround loudspeakers are located at approximately $\pm 100^\circ$ to $\pm 120^\circ$.

1.2.2 Present Spatial Sound Reproduction Techniques

In this section, we present some modern spatial reproduction-recording techniques; namely, Ambisonic, vector based amplitude panning, wave field synthesis and directional audio coding.

1.2.2.1 Ambisonic

Ambisonic is a surround sound recording and reproduction system. Several researchers shared the invention of its theoretical part [42], [43], [44]. However, Ambisonic was mainly introduced by Gerzon [45]. Ambisonic deals with reconstruction of the sound wave and it aims at recreating the sound field to as large area as possible. Its optimum goal is to recreate the localization cues, which are heard in the natural hearing, using a small number of loudspeakers and channels. However, the larger the rendered area is, the higher the order of Ambisonic is needed [46].

Several signal formats can be used in Ambisonic system. These signal formats are [35]:

1. A-format signals which are used for microphone pick-up.

2. B-format signals which are used for studio equipment and processing.

3. C-format signals which are used for transmission.

4. D-format signals which are used for decoding and reproduction.

Both A-format and B-format signals will be explained in chapter 5. C-format signals consist of four signals; namely, L_c, R_c, T_c and Q_c. The number of C-format signals needed for transmission is a matter of the directional resolution we aim at. Hence, two, three or four C-format signals can be used [34].

D-format signals are used in the synthesis stage, they can be derived from either B-format signals or from C-format signals. Distributed D-format signals can be used for any loudspeaker layout and for unlimited number of loudspeakers [35]. However, the limitation in this case is about the minimum number of loudspeakers, which should be equal to the number of Ambisonic signals or bigger [47].

Ambisonic can be divided into two groups; namely, first-order Ambisonic and higher-order Ambisonic (HOA). The decoding equation for the first-order Ambisonic is B-format signals equation. The higher-order Ambisonic provides a bigger sweet spot in which the sound field is accurately reproduced. However, increasing the order of the Ambisonic requires increasing the number of channels needed to transmit the sound signal which means increasing of the needed number of loudspeakers as well [48]. For instance, in case of second-order Ambisonic, the encoding equation are [49]

$$\begin{aligned} X_a &= \cos(\alpha)\cos(\beta) \\ Y_a &= \sin(\alpha)\cos(\beta) \\ Z_a &= \sin(\beta) \\ W_a &= \frac{1}{\sqrt{2}} \\ R_a &= 1.5\sin^2\beta - 0.5 \\ S_a &= \cos(\alpha)\sin(2\beta) \\ T_a &= \sin(\alpha)\sin(2\beta) \\ U_a &= \cos(2\alpha)\cos^2(\beta) \\ V_a &= \sin(2\alpha)\cos^2(\beta) \end{aligned} \tag{1.29}$$

where $X_a, Y_a, Z_a, W_a, R_a, S_a, T_a, U_a, V_a$ are the signals of second-order Ambisonic, α and β are the angle of the sound source in the horizontal and vertical plane. Again, for horizontal plane, $\beta = 0$, which means that the signals Z_a, R_a, S_a and T_a are eliminated, and the other signals are the only used channels in this case.

The previous equation shows the cost of using higher-order Ambisonic. However, it might achieve larger sweet spot.

Ambisonic Decoder

Gerzon has defined a decoder to be Ambisonic if it achieves the following conditions [50]

1. It should maintain the energy vector at least up to 4 kHz.
2. The velocity vector should be near unity at low frequency up to 400 Hz.
3. The energy vector magnitude should be maximized at middle and high frequency i.e., in interval $700 - 4000$ Hz.

As mentioned before, the format of the signal in the transmission part does not depend on the loudspeaker layout. Hence, each loudspeaker layout has its special decoder. However, the loudspeaker layout should be as regular as possible [48]. The design of the decoder depends on the loudspeaker layout. Loudspeaker layout can be divided into three groups [47]

- Regular polygons and polyhedrons such as square, hexagon, cube and dodecahedron.
- Irregular arrays but with speakers in diametrically opposite pairs such as rectangle.
- General irregular arrays such as loudspeaker layout according to ITU-R BS.775 recommendation.

The loudspeaker lay out plays the major rule in designing the Ambisonic decoder. There are many attempts to design an Ambisonic decoder for regular and irregular loudspeaker setup [51], [52]. The gain for each loudspeaker can be written as a sum of Ambisonic signals multiplied by the gain for each loudspeaker as [52]

$$S_i = G_{xi}X_{a1} + G_{yi}Y_{a1} + G_{wi}W_{a1} \tag{1.30}$$

where S_i is the signal fed to the i^{th} loudspeaker, G_{xi}, G_{yi} and G_{wi} are the decoder coefficients for the i^{th} loudspeaker applied to the encoded audio signals X_{a1}, Y_{a1} and W_{a1}, respectively. The previous equation is applied to loudspeaker in the horizontal plane.

1.2.2.2 Vector Base Amplitude Panning

Vector base amplitude panning (VBAP) is a method for spatial sound reproduction. It was invented by Pulkki [53]. It can be used in both two-dimensional (2D) and three-dimensional (3D) plane to calculate the gains for each loudspeaker. It can be seen as a *generalized tangent law* in the two-dimensional plane [40]. VBAP can work with any loudspeaker setup, the gains can be calculated depending on this method to ensure a good rendering of the virtual sound using the surrounding loudspeakers.

Using VBAP in Two-Dimensional Plane

When VBAP is used in 2D plane, a loudspeaker pair is specified with two unit-length vectors ($\mathbf{I}_m$ and $\mathbf{I}_n$), where $\mathbf{I}_m = [\mathrm{I}_{m1}\ \mathrm{I}_{m2}]^T$ and $\mathbf{I}_n = [\mathrm{I}_{n1}\ \mathrm{I}_{n2}]^T$ are unit-length vectors pointing to the direction of the loudspeakers m and n from the listener's position in 2D

plane, respectively. Supposing that the vector $\mathbf{P}$ indicates the direction of the virtual sound source, it can be written as a linear weighted sum of the loudspeaker vectors [53]

$$\mathbf{P} = G_{lm}\mathbf{I}_m + G_{ln}\mathbf{I}_n \tag{1.31}$$

where G_{lm} and G_{ln} are the gain factors for the loudspeaker m, n, respectively, and it is calculated as

$$\mathbf{G}_{2l} = \mathbf{P}^T\mathbf{L}_{mn}^{-1} \tag{1.32}$$

where $\mathbf{G}_{2l} = [G_{lm}\ G_{ln}]^T$ and $\mathbf{L}_{mn} = [\mathbf{I}_m\ \mathbf{I}_n]$. Figure 1.7 shows the case when VBAP is used with two loudspeakers to create a virtual sound between them [54].

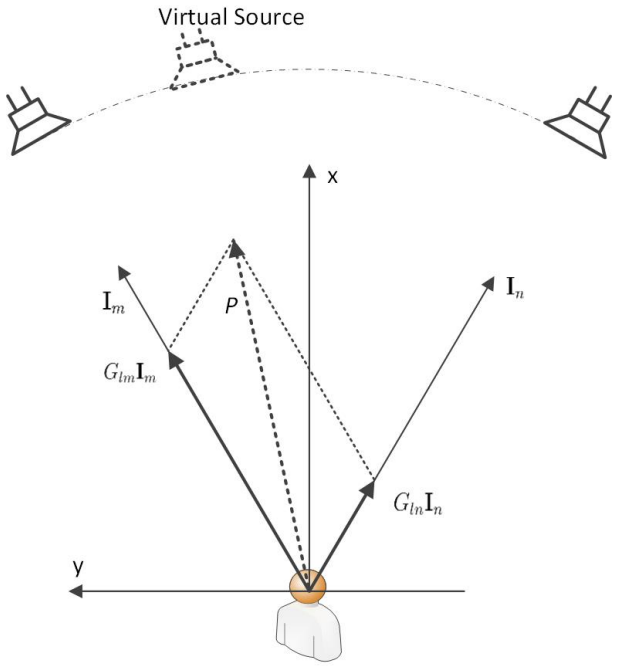


Figure 1.7: Two-dimensional VPAP with a loudspeaker pair.

Using VBAP in Three-Dimensional Plane

VBAP can be used in three-dimensional plane. In this case, the Cartesian unit vector $\mathbf{L}_{mnk} = [\mathbf{I}_m\ \mathbf{I}_n\ \mathbf{I}_k]$, where the vectors $\mathbf{I}_m$, $\mathbf{I}_n$ and $\mathbf{I}_k$ define the direction of the loudspeakers m, n, and k, respectively, and the Cartesian vector $\mathbf{I}_m = [\mathrm{I}_{m1}\ \mathrm{I}_{m2}\ \mathrm{I}_{m3}]$ points to the direction of the loudspeaker m. Hence, the gain factors can be calculated as [40]

$$\mathbf{G}_{3l} = \mathbf{P}^T\mathbf{L}_{mnk}^{-1} = [\mathrm{p}_m\ \mathrm{p}_n\ \mathrm{p}_k]\begin{bmatrix}\mathrm{I}_{m1} & \mathrm{I}_{m2} & \mathrm{I}_{m3}\\ \mathrm{I}_{n1} & \mathrm{I}_{n2} & \mathrm{I}_{n3}\\ \mathrm{I}_{k1} & \mathrm{I}_{k2} & \mathrm{I}_{k3}\end{bmatrix}^{-1} \tag{1.33}$$

where $\mathbf{G}_{3l} = [G_{lm}\ G_{ln}\ G_{lk}]^T$ are the gain factors for the loudspeakers m, n, k, and $\mathbf{P}^T = [\mathrm{p}_m\ \mathrm{p}_n\ \mathrm{p}_k]$ is the panning vector of the virtual source.

However, when more loudspeakers are used, a triplet-wise method is performed. The chosen loudspeakers of a triplet should meet the following conditions [40]

1. They are not in the same plane with the listeners.
2. The triangles made by different triplets should not overlap.
3. The triangles of the triplets should have as short sides as possible.

1.2.2.3 Wave Field Synthesis

Wave field synthesis (WFS) is a sound reproduction method aiming at reproduction spatial sound over a large area. The concept of (WFS) was originally presented by Snow in [55]. However, the terminology (wave field synthesis) has appeared later in [56]. WFS depends on Huygens' principle, which states that each point of a wave front can be considered as a source of a new wave [57]. Thus, the principle of WFS is derived from Huygens' principle by replacing each *new wave* in the surface by a *loudspeaker*. That can be done by recording the sound wave by a number of microphones and reproducing the reordered signals by the same number of loudspeakers arranged next to each other. The listener in front of the loudspeaker array can perceive a virtual sound correctly localized even when he moves in the rendered area [58].

When only plane waves are created, wave field synthesis can produce the same acoustic image for many listeners in the listening area. The sound field reproduction is valid not only at a particular point, but also at any point within the whole area. It should be noted that the listening area, compared with other surround sound systems techniques, is very large. In theory, the synthesis of the wave field arises from the summation of an infinite number of loudspeaker signals. In practice, however,the loudspeaker array will always have a finite length. The finite array can be seen as a window, through which the primary (virtual) source is either visible, or invisible, to the listener. The conventional WFS reproduces the sound in the horizontal plane only. However, new three dimensional WFS has been investigated in the literature [59].

The main limitation of WFS is that it needs a huge number of loudspeakers in order to render the sound field in a large area. The 3D WFS requires more loudspeakers than the conventional 2D WFS [59]. The article presented in [60] introduced a new technique for 3D WFS with a limit number of loudspeakers (24 loudspeakers were used in that article).

The second limitation of WFS is the artifacts. WFS suffers from the following artifacts [61]

1. Spatial aliasing artifact caused by secondary sources, which leads to localization inaccuracy.
2. The limitation of used loudspeakers causes a limitation in the area of correctly reproduced wave field and also causes circular waves propagated from the secondary sources.

3. Amplitude errors caused by the secondary sources.

1.2.2.4 Directional Audio Coding

Directional audio coding (DirAC) is a method for multichannel audio reproduction of spatial sound [62]. DirAC can work with different loudspeaker configurations, and it can transmit spatial aspects of audio with a low bit rate. DirAC is found to be a suitable method for sound reproduction in this work, and it is used in the listening tests presented in the next chapters. Therefore, it is discussed in details in this chapter.

DirAC is based on spatial impulse response rendering (SIRR) [63]. SIRR states that reproducing the spatial impression of an existing performance venue does not require to perfectly reconstruct the original sound field [64]. Both DirAC and SIRR are based on the following assumptions [65],

1. Direction of arrival (DOA) transforms into interaural time difference (ITD), interaural level difference (ILD), and monaural cues.

2. Diffuseness transforms into interaural coherence cues (IC).

3. The timbre of the sound depends on the monaural spectrum together with ITD, ILD and IC.

4. The direction of arrival (DOA), diffuseness and spectrum of sound measured in one position with the temporal and spectral resolution of human hearing determine the auditory spatial image the listener perceives.

The first three assumptions say that the correct reproducing of the directional of arrival guarantees the correct perceiving of the interaural time difference and interaural level difference. The fourth assumption depends on the fact that the auditory system is able to decode only one spatial cue within one critical frequency band at one instant time [13]. Therefore, there is no need to store a separate audio channel for each loudspeaker and one audio channel will be sufficient for recreating the spatial sound with information about direction and diffuseness at different critical bands.

DirAC can be divided into three parts; namely, analysis, transmission and synthesis, see Figure 1.8 [66].

DirAC Analysis

The analysis part is performed in the frequency domain. Several time-frequency transforms can be used to transmit the sound signal into frequency domain. For instance, short time Fourier transform [67], modified discrete cosine transform (MDCT) [68], Gabor transform [69] and filter banks [70]. These transformations differ in time-frequency resolution and computational efficiency, as it will be seen in the next chapter. However, in this work we will use both STFT and Gabor transform.

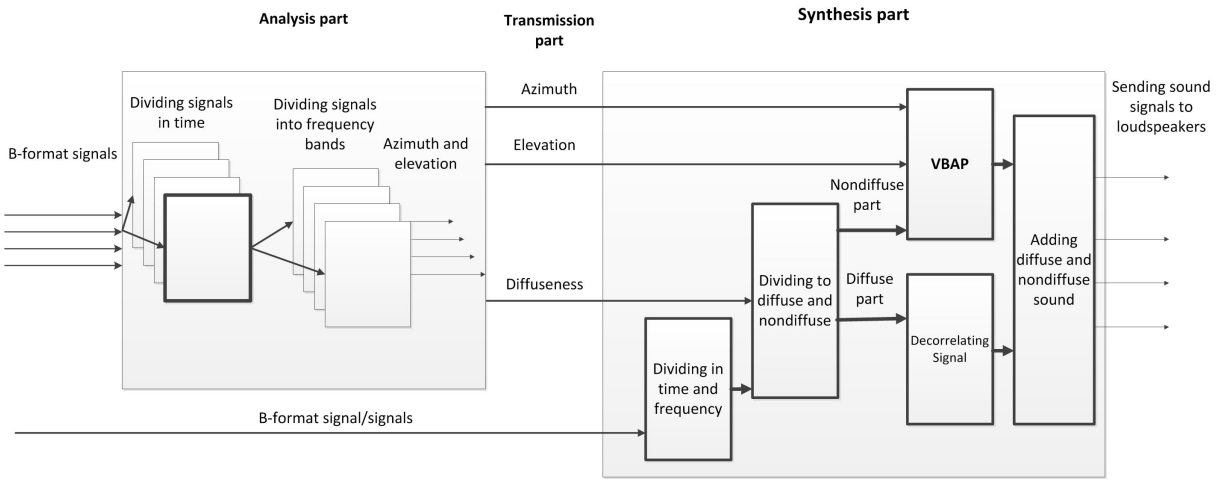


Figure 1.8: DirAC diagram.

DirAC analysis part aims at estimating the direction of arrival (DOA) and the diffuseness, which can be derived from the sound pressure and the particle velocity [62]. The input signals for this stage can be the recorded signals of any microphones for which the pressure and velocity can be estimated. For instance, B-format signals are suitable signals for DirAC. However, an other microphone array can be used [71], [72].

Recalling that B-format signals consists of four signals, $x(t), y(t), z(t)$ and $w(t)$, and $w(t)$ signal is obtained by omni-directional microphone. The sound pressure, expressed in STFT domain, can be obtained as [66]

$$p = \sqrt{2}W \tag{1.34}$$

where p is the pressure and W is the B-format signal $w(t)$ in the STFT domain.

The instantaneous energy density E can be computed as

$$E = \frac{1}{2}\rho_0 \left(\frac{p^2}{Z_0^2 + u^2} \right) \tag{1.35}$$

where ρ_0 is the density of the air, Z_0 is the characteristic acoustic impedance of air and u is the particle velocity.

The diffuseness can be then defined as [62]

$$\Psi = 1 - \frac{\|\langle \mathrm{I} \rangle\|}{c\langle E \rangle} \tag{1.36}$$

where Ψ is the diffuseness, c is the speed of sound, I is the instantaneous intensity vector and $\langle . \rangle$ means short-time average.

When B-format signals are used as input signals, the diffuseness can be calculated by

equation [62]

$$\Psi(n_1) = 1 - \frac{\sqrt{2}\left\| \sum_{m_1=a_1}^{b_1} W(n_1+m_1)v(n_1+m_1)W_1(m_1) \right\|}{\sum_{m_1=a_1}^{b_1} [|W(n_1+m_1)|^2 + |v(n_1+m_1)|^2/2]W_1(m_1)} \tag{1.37}$$

where $\Psi(n_1)$ denotes the diffuseness in the time frame number n_1, $\|.\|$ denotes vector norm, $W(n_1)$ is the B-format signal W in the time frame number n_1, $W_1(m_1)$ is a window function defined between constant time values $a_1 \leq 0$ and $b_1 > 0$ and v is the particle velocity vector.

The second task of DirAC analysis is to compute the *instantaneous* direction of arrival, which can be computed at each frequency channel using the formula [62]

$$\mathbf{D}(n_1) = -\sum_{m_1=a_2}^{b_2} W(n_1+m_1)v(n_1+m_1)W_2(m_1) \tag{1.38}$$

where $W_2(m_1)$ is a window function for short-time averaging defined between constant time values $a_2 \leq 0$ and $b_2 > 0$.

DirAC Transmission

DirAC provides the opportunity of transmission the sound with both low-bit-rate and high-bit-rate systems. For the high-bit-rate, all microphone signals are transmitted, and the analysis can be performed after transmission. However, in teleconferencing applications, the data rate should be kept low. Therefore, the analysis part is done before the transmission, and only one channel of audio is transmitted, and the diffuseness and the direction of arrival are transmitted as a meta data [72].

DirAC Synthesis

The audio signal(s) in the synthesis part is first divided into non-diffuse and diffuse streams in each frequency channel. That can be done by multiplying the input signal(s) with two time-variant factors derived from diffuseness [62]. The diffuse part is estimated from the audio signal by multiplying it by $\sqrt{\Psi}$, whereas the non-diffuse part is calculated as the result of multiplying the audio signal by $\sqrt{1-\Psi}$ [62].

After obtaining the diffuse and non-diffuse streams, they are processed separately. The non-diffuse part can be panned using different rendering methods, such as VBAP, Ambisonic and wave field synthesis. When only one audio channel is transmitted, the gain factor is computed using the information of the loudspeaker setup, and the direction of arrival that was estimated in the DirAC analysis.

In order to avoid audible artifacts, which could happen because of the temporal variations, the second phase of temporal averaging in directional reproduction has to be computed. Thus, the gain factors calculated by VBAP has to be temporal-domain averaged

as [62]

$$g(n_a, i_a, k_a) = \frac{\sum_{m_a=-M_a/2}^{M_a/2} g(n_a + m_a, i_a, k_a)[1 - \Psi(n_a + m_a, i_a)]W_{3a}(m_a)}{\sum_{m_a=a_3}^{b_3} [1 - \Psi(n_a + m_a, i_a)]W_{3a}(m_a)} \tag{1.39}$$

where $g(n_a, i_a, k_a)$ is the gain factor, i_a represents subband, n_a is the time, k_a is the loudspeaker channel, and $W_{3a}(m_a)$ is a window function defined between constant time values $a_3 \leq 0$ and $b_3 > 0$.

The diffuse part of the sound is decorrelated for each loudspeaker. The goal of the decorrelation is decreasing the coherence between loudspeaker signals [62]. The decorrelated signals are then applied to the corresponding loudspeakers.

Chapter 2

Time-Frequency Representation

A part of this research focuses on the impact of the signal resolution in time and frequency domain on the accuracy of the some sound source localization and the rendering methods. Thus, an overview about the signal representation and the resolution in time and frequency domain is introduced in this chapter.
This chapter contains the following subjects

1. Time-frequency analysis.
2. Windowing.
3. Signal reconstruction and inverse transform.

2.1 Time-Frequency Analysis

Time-frequency transformations represent the transformed signal using independent frequency and time variables. Thus, working with the frequency components regarding the time slots is possible. In following, a background about time-frequency analysis methods is presented including Fourier transform, short-time Fourier transform, and Gabor transform.

2.1.1 Fourier Transform

The well-known Fourier transform is the most used way to transform a signal to the frequency domain. It provides the information about the existence of the frequency components of the signal regardless of their time duration. In other words, Fourier transform tells if a certain frequency exists in the whole signal regardless to the time of its existence. Thus, the Fourier applications can be seen as a one-dimensional application.

The continuous form of Fourier transform is given as [73]

$$X(f) = \int_{-\infty}^{\infty} x(t) \mathrm{e}^{-\mathrm{j}2\pi f t} \mathrm{d}t \tag{2.1}$$

where $x(t)$ is a continuous time-domain signal, $X(f)$ is the signal in the continuous frequency-domain, t is the time and f is the frequency.

The discrete Fourier transform can be written as [74]

$$X[m] = \sum_{n=0}^{N-1} x[n] \mathrm{e}^{\frac{-\mathrm{j}2\pi nm}{N}} \tag{2.2}$$

where $x[n]$ is a discrete sequence of time-domain sampled values of the continuous signal $x(t)$, $X[m]$ is the discrete Fourier transform (DFT) of the signal, m is the index of DFT output in the frequency domain and N is the number of samples.

2.1.2 Short-Time Fourier Transform

One can say without a doubt that Fourier transform is a strong tool for signal processing and analysis. Although the mentioned transform provides information about the frequencies contained in the signal, it has a problem representing the frequencies varying with time.

One of the most popular time-frequency representation tool is so-called short-time Fourier transform (STFT). STFT is simply the output of the Fourier transform applied to a signal within a sliding window, when this window is moved along the signal. STFT can be expressed in the continuous time as [75]

$$\mathbf{STFT}_h x(t) \equiv X_h(\tau, \omega) = \int_{-\infty}^{\infty} x(t) h(t-\tau) \mathrm{e}^{-\mathrm{j}\omega t} \mathrm{d}t \tag{2.3}$$

where $\omega = 2\pi f$ is the angular frequency , $h(t-\tau)$ is the window function and $x(t)h(t-\tau)$ is the resulted signal and it is centered at the time τ.

The discrete version of the previous equation can be expressed as [76]

$$\mathbf{STFT}_h x[n] \equiv X_h[m, k] = \sum_{n=0}^{N-1} x[n] h[n-m] \mathrm{e}^{\frac{-\mathrm{j}2\pi kn}{N}} \tag{2.4}$$

where k is the frequency bin, n is the time sample, h is the used window and N is the number of the window samples.

STFT considers the non-stationary to be *almost* stationary signal inside the window, which makes it possible to apply Fourier transform to the signal inside the window.

2.1.3 Gabor Transform

Gabor transform was presented in [77] by D. Gabor. The major idea of Gabor transform is to use a window function for extracting the information from the Fourier transform during the time existence of this window. It can be seen as a special way for time-frequency representation which portions the time-frequency plane into equally sized regions in the

conventional form of Gabor transform. The so-called Gabor atom is used to represent the time-frequency spread, regardless of the time and frequency ordering, see Figure 2.1 [78].

The mentioned Gabor's point of view has advantage over STFT by enabling a mathematical description of the window's properties, although they are mathematically equivalent [78].

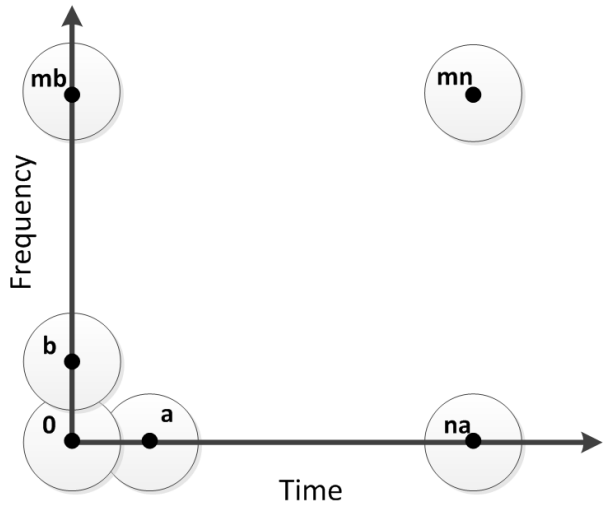


Figure 2.1: Gabor's atoms.

The continuous Gabor transform can be expressed as [79]

$$G_x(\tau,\omega) = \int_{-\infty}^{\infty} x(t)\overline{h(t-\tau)}\mathrm{e}^{-\mathrm{j}\omega t}\mathrm{d}t \tag{2.5}$$

where h is a window function, $x(t)$ is a finite energy signal, the operator $\overline{(.)}$ returns the conjugate part, t is the time and τ is the shift in time.

The discrete version of Gabor transform can be written as [80]

$$G_h[m,n] = \sum_{l=0}^{L-1} x[l]\overline{g[l-an]}\mathrm{e}^{\frac{-\mathrm{j}2\pi ml}{M}} \tag{2.6}$$

where x is a signal of length L, g is a window function, a and M are equivalent to the number of channel and the decimation in time respectively in the Fourier modulated filter bank.

As can be seen from the previous equation, the Gabor transform is a function of two variables, the first one is the time variable τ, this variable is the center of the window function in the time domain. The second one is the frequency variable. Hence, the Gabor transform represents a map of the signal in two domains (time domain and frequency domain).

2.2 Windowing

Windowing is a method applied to the signal in order to improve the time-frequency representation. This technique restricts the signal to ensure that the wanted relevant oscillatory signal features would appear. Choosing the window influences the resolution of the time-frequency representation.The longer the window we used, the less the resolution in time we get and the more the resolution in frequency. Even more, the window should be chosen carefully to ensure the possibility of synthesis the signal again.

The window can be seen generally as an even function whose values are real, positive and concentrated around the point, where it achieves its maximum. Three issues should be taken into account when a window is applied to the signal

1. The shape of the window.
2. The edges of the window.
3. The size of the window.

2.2.1 Window Function

Each signal $g(t)$ that fulfills a certain condition can be considered as a window. This condition requires that $g(t) \in L^2(\mathbb{R})$, $\| g(t) \|_2 \neq 0$ and $tg(t) \in L^2(\mathbb{R})$ [76].

In following, we present the most used window with the time-frequency representation. Figure 2.2 illustrates the spectral resolution of several windows using a sinusoid signal with 100 Hz frequency. It should be noted that the mathematical formulas of these windows are given, so the windows are centered around zero with ability to change the center and the amplitude of each window.

The Gaussian Window

The Gaussian window is one of the most used window in the time-frequency representation thanks to its properties. The best property of this window is its Fourier transform, where its Fourier transform is also a Gaussian window [76]. The Gaussian window can be described as [76]

$$g_{\mathrm{GA}}(t) = A\mathrm{e}^{-Bt^2} \tag{2.7}$$

where $A, B > 0$ are variables. However, when Gabor transform is used, the parameters of the Gaussian window are chosen as [79]

$$g_{\mathrm{GA}}(t) = \mathrm{e}^{-t^2/2\sigma} \tag{2.8}$$

where σ is the standard deviation of the distribution. The value of σ determines the width of the window. Thus, it influences the resolution of the transform. For instance, when $\sigma = 1$ the frequency resolution is too poor [79].

Rectangular Window

Rectangular window can be described by the equation [79]

$$g_{\text{re}}(t) = \begin{cases} 1, & \text{for } -\frac{A}{2} \leq t \leq \frac{A}{2} \\ \\ 0, & \text{otherwise} \end{cases} \tag{2.9}$$

where $A > 0$ is a real variable.

The rectangular window has the narrowest main lobe. Therefore, it provides the best resolution. However, the main lobe is surrounded by big slide-lobes, which can be seen as a disadvantage of this window.

Bartlett (Triangular) Window

This window can be described using the equation [79]

$$g_{\text{tr}}(t) = \begin{cases} 2t, & \text{for } -A/2 \leq t \leq 0 \\ 2(1-t), & \text{for } 0 \leq t \leq A/2 \\ 0, & \text{otherwise.} \end{cases} \tag{2.10}$$

This type of windows has a narrow spectral peak and a large side lobe, see Figure 2.2.

Hamming Window

Hamming window has a useful advantage over other windows regarding its spectrum. The spectrum of the main lobe of this window drops quickly in comparison to other windows . Hamming window can be described using the equation [79]

$$g_{\text{ham}}(t) = \begin{cases} 0.54 - 0.46\cos 2\pi t, & \text{for } 0 \leq t \leq 1 \\ \\ 0, & \text{otherwise.} \end{cases} \tag{2.11}$$

Hanning Window

The Hanning window belongs to the Hamming window family with a difference about the edges of this window, where Hanning window has zero values at the tails. Thus, it causes some circumstances involving not using the whole data. However, this problem is solved by using overlapping. The Hanning window is given by the equation [79]

$$g_{\text{han}}(t) = \begin{cases} 0.5(1 + \cos(2\pi t)), & \text{for } 0 \leq t \leq 1/2 \\ \\ 0, & \text{otherwise.} \end{cases} \tag{2.12}$$

This window is also called the raised cosine window.
By using the formula $\cos(2t) = \cos^2(t) - \sin^2(t) = 2\cos^2(t) - 1$, Hanning window can be defined as

$$g_{\text{han}}(t) = \begin{cases} \cos^2(\pi t), & \text{for } 0 \leq t \leq 1 \\ \\ 0, & \text{otherwise.} \end{cases} \tag{2.13}$$

Blackman Window

Blackman window can be described using equation [79]

$$g_{\text{bl}}(t) = \begin{cases} 0.42 - 0.5\cos 2\pi t + 0.08\cos 4\pi t, & \text{for } 0 \leq t \leq 1 \\ \\ 0, & \text{otherwise.} \end{cases} \tag{2.14}$$

The Blackman window has the best of both Hamming and Hanning window. As can be seen in Figure 2.2, the main lobe has a sharp drop, whereas the side-lobes are narrow and drop off rapidly.

2.3 Time-Frequency Localization

When a signal is represented in the time-frequency plane, a trade off must be done between the resolution in time domain and the resolution in frequency domain. These properties can be clearly seen when a rectangular window is used. The Fourier transform of a rectangular window presented in eq. (2.9) is given by equation

$$G_{\text{re}}(f) = \int_{-\infty}^{+\infty} g_{\text{re}}(t)\mathrm{e}^{-\mathrm{j}2\pi ft}\mathrm{d}t = \int_{-A}^{A} e^{-\mathrm{j}2\pi ft}\mathrm{d}t = 2A[\frac{\sin 2\pi fA}{2\pi fA}] \tag{2.15}$$

where $G_{\text{re}}(f)$ is the Fourier transform of the rectangular window $g_{\text{re}}(t)$ which has a magnitude = 1 and width $2A$ and centered around 0.

As can be seen from the previous equation, the spread of the window in frequency domain depends on the parameter A. The bigger the parameter A is, the broader the window in the time domain is, and the narrower the window in the frequency domain is and vice versa. Assuming that the window has an infinity width in time domain, the width of the window in frequency domain will approach zero, leading to a Dirac pulse. That can be generalized for all windows. In order to achieve the best resolution of a given window, the length of this window should be chosen carefully. Whereas a longer window provides less localization in time, it ensures more discrimination in frequency.

2.3.1 Heisenberg Principle

The uncertainty principle elucidates the possibility of performing measurement on a system without disturbing it. The famous Heisenberg uncertainty principle, which was presented

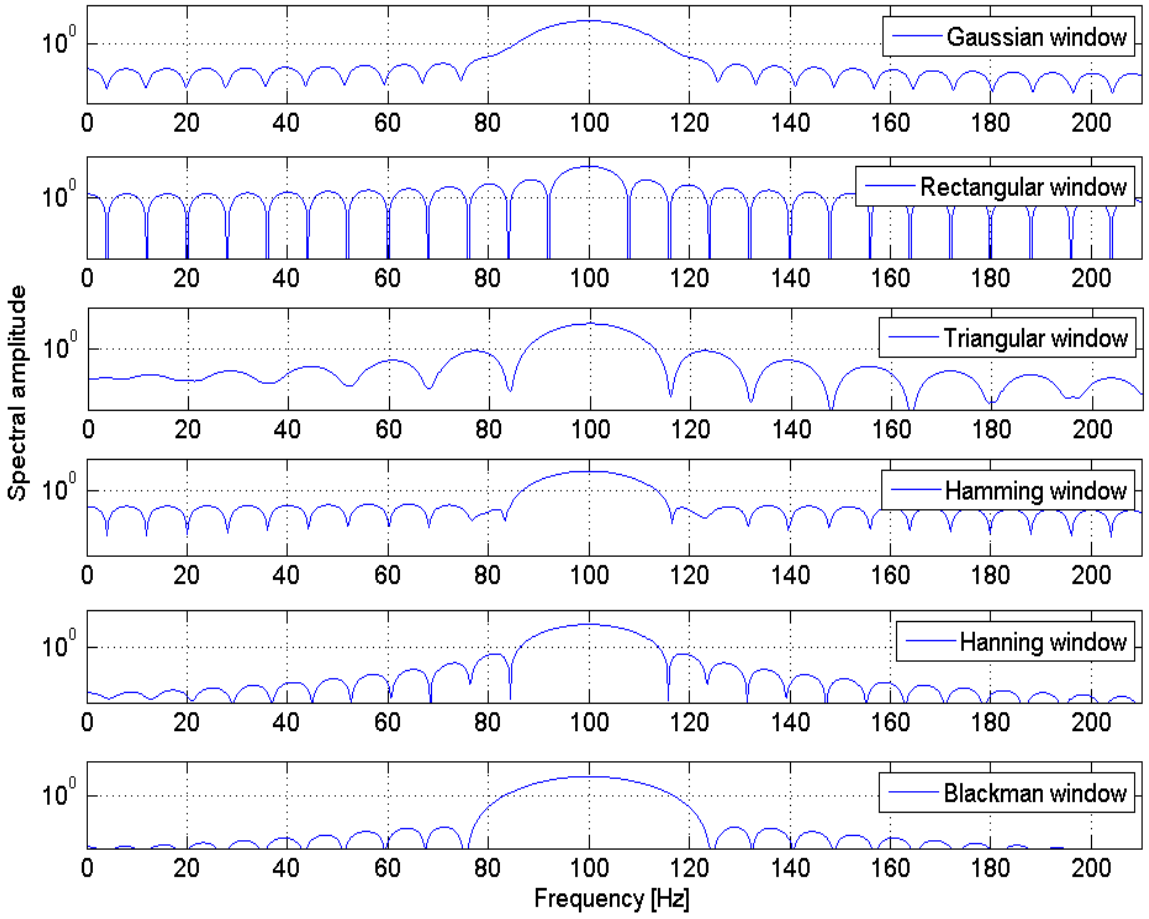


Figure 2.2: Spectral representation of the weighting windows.

by Heisenberg in 1927 [81], formulated the relationship between the position and the momentum in the quantum mechanics. It states that it is impossible to determine the exact position and momentum of a particle simultaneously.

However, the Heisenberg uncertainty principle can be generalized to include the signal processing discipline, describing the relation between the resolution in time and frequency when a signal is transformed into frequency domain. It imposes a lower limit on the area of the rectangle that represents the time-frequency localization in two dimensional plan, i.e., $\Delta t, \Delta f$, see Figure 2.3.

Supposing that $x(t) \in L^2(\mathbb{R})$ is a window function, $X(w) = F[x](w)$ is the radial Fourier transform of $x(t)$, $\Delta(t) = 2\rho_t$ is defined as a measure of time duration of the window function $x(t)$ and $\Delta(f) = 2\rho_f$ is a measure of the bandwidth of its Fourier transform, see

Figure 2.3. Heisenberg's inequality can be expressed as [76]

$$\rho_t \rho_f \geq \frac{1}{2}. \tag{2.16}$$

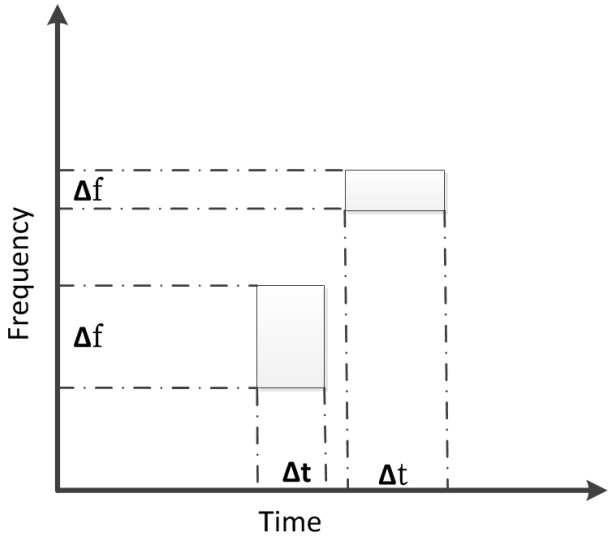


Figure 2.3: Resolution in time and in frequency.

Another way to see the previous equation is that a signal and its Fourier transform cannot be both sharply localized simultaneously [82].
It should be noted that the equality in eq. (2.16) can be held if and only if $x(t) = ae^{-bt^2}$ for $a \in \mathbb{C}$ and $b \geq 0$ [76]. The previous condition holds when a Gaussian window is used. That can be easily proven recalling that the radius of the window in time and frequency domain ρ_t and ρ_f respectively are given by equations [76]

$$(\rho_t)^2 = \frac{1}{\|x\|_2^2} \int_{-\infty}^{\infty} (t - \langle t \rangle)^2 |x(t)|^2 \mathrm{d}t, \tag{2.17}$$

and

$$(\rho_f)^2 = \frac{1}{\|X\|_2^2} \int_{-\infty}^{\infty} (w - \langle w \rangle)^2 |X(w)|^2 \mathrm{d}w \tag{2.18}$$

where $\|.\|_P$ is the norm in the Banach space $L^P(\mathbb{R})$ and the expectation value $\langle t \rangle$ and $\langle w \rangle$ of the time t and the frequency w are given by equations

$$\langle t \rangle = \frac{1}{\|x\|_2^2} \int_{-\infty}^{\infty} t |x(t)|^2 \mathrm{d}t, \tag{2.19}$$

and

$$\langle w \rangle = \frac{1}{\|X\|_2^2} \int_{-\infty}^{\infty} w |X(w)|^2 \mathrm{d}w. \tag{2.20}$$

2.4 Signal Reconstruction and Inverse Transform

The inverse transform aims at transforming the signal from the frequency domain into the time domain. In following we present the inverse transform and the ability of reconstruction the signal into time domain after modifying the signal in the frequency domain.

2.4.1 Inverse Fourier Transform

The inverse Fourier transform is given by equation [76]

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X(f) \mathrm{e}^{\mathrm{j}2\pi f t} \mathrm{d}f. \tag{2.21}$$

The inverse version finds the function in time domain from its Fourier function supposing the existence of such function. A sufficient condition for its existence can be written as

$$\int_{-\infty}^{\infty} |x(t)| \mathrm{d}t < \infty. \tag{2.22}$$

The previous condition ensures the existence of both the Fourier transform of a function and its inverse. However, this condition is sufficient but it is not necessary [67].

The inverse version of the discrete Fourier transform for a discrete signal $x(n)$ and its discrete Fourier transform $X(m)$ on the interval $[0, N-1]$ is given as [76]

$$x(n) = \frac{1}{N} \sum_{m=0}^{N-1} X(m) \mathrm{e}^{\frac{2\pi \mathrm{j} n m}{N}}. \tag{2.23}$$

2.4.2 Inverse Short-Time Fourier Transform

The reconstruction of the signal into time domain, when STFT is used, requires overlapping between the windows. The window length should be chosen as a compromise between the time and frequency resolution, as it was discussed before.

The inverse continuous STFT can be written as [83]

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} X_h(\tau, \omega) \mathrm{e}^{\mathrm{j}\omega t} \mathrm{d}\omega \mathrm{d}\tau. \tag{2.24}$$

We assumed in the previous equation that the window is scaled to one

$$\int_{-\infty}^{\infty} h(\tau) \mathrm{d}\tau = 1. \tag{2.25}$$

By applying a different window to eq. (2.24) we get

$$x(t)h(t-\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X_h(\tau, \omega) \mathrm{e}^{\mathrm{j}\omega t} \mathrm{d}\omega. \tag{2.26}$$

The inverse discrete STFT can be written as [84], [76]

$$x(n) = \frac{1}{N||h||_2^2} \sum_{m=0}^{N-1} \sum_{k=0}^{N-1} X_h(m,k) h(n-m) \mathrm{e}^{2\pi \mathrm{j} k \frac{n}{N}} \tag{2.27}$$

where $X_h(m,k)$ is the discrete STFT of a discrete signal $x(n)$ which has period $N > 0$, $h(n)$ is a window function and $||h||_2$is a l^2 -norm of $h(n)$ in the interval $[0, N-1]$.

The previous equation ensures the reconstruction of the signal from the frequency domain to the time domain. However, to ensure a good reconstruction, the overlapping should be chosen carefully. An article presented in [85] suggested using half-overlapping square-root for both analysis and synthesis part.

2.4.3 Inverse Gabor Transform

The continuous version of inverse Gabor transform is [76]

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} G_x(\tau, \omega) \mathrm{e}^{\mathrm{j}\omega t} \mathrm{d}\omega \right) \mathrm{d}\tau \tag{2.28}$$

where $x(t)$ is the signal in the time domain, and its Gabor transform is $G_x(\tau, \omega) \in L^1(\mathbb{R})$ and $g(t) = g_{\tau,\sigma}(t)$ is a Gaussian window with $\sigma > 0$ which is the standard derivation and τ is the mean of this window.

The inverse discrete Gabor transform can be written as [86]

$$x(l+1) = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} G_h(m+1, n+1) \mathrm{e}^{\frac{n\pi \mathrm{j} m l}{M}} g(l - an + 1) \tag{2.29}$$

where g is a window function, a is the length of the time shift and $G_h(m,n)$ is an array of Gabor coefficient of size MxN.

Chapter 3

Objectives of Dissertation

The main goal of the dissertation is to design and test an acoustic zooming system, which can locate and zoom the sound of one speaker among the multiple speakers.

In order to design such system, this work had to go gradually through several steps, which can be summarized as follows:

- Explore the existing surround sound techniques: The goal of this step is to choose the best sound rendering method, which can be used in the proposed acoustic zooming system.
- Investigate sound source direction estimation methods: This step introduces several sound source localization techniques, shows their advantages and disadvantages, and studies the absolute angle error of these techniques.
- Design a new sound source direction estimation method and evaluate the proposed method: The goal of this step is to suggest a new method, which can be used in the proposed acoustic zooming system in order to estimate the direction of multiple speakers.
- Implement an acoustic zooming system using DirAC and sound source direction estimation method: The purpose of this step is to propose a system which can estimate the direction of arrival of multiple speakers and also choose the sound of one of them and zoom it while attenuating the other sounds.
- Perform subjective and objective quality assessments to evaluate the proposed acoustic zooming system: Evaluation is an essential part of each system. Therefore, this step aims at evaluating the proposed system by performing the listening tests and applying objective measurements using several time-frequency transforms.

Chapter 4

A Comparison and Evaluation of Localization and Rendering Methods

This chapter investigates the accuracy of some sound source localization methods, and also provides the listening tests to evaluate and compare the performance of sound reproduction techniques.

To compare the performance of sound source localization methods, the methods were first simulated in Matlab, and then the measurements were performed in the laboratory to affirm the simulation results.

The evaluation of sound rendering methods is proceeded by performing the listening tests in laboratory. The listening tests were designed to compare the average absolute angle error of these methods when the listener is moved out of the center of the loudspeaker array.

4.1 Comparison of Sound Source Localization Methods

In order to evaluate the time delay of arrival based methods, we simulated the three methods mentioned earlier in section 1.1.2.3, namely, cross-correlation (CC), phase transform (PHAT) and maximum likelihood (ML), and we studied the impact of the existence of the noise signal on their performance [87]. The experimental results are also presented in this section.

4.1.1 Simulation Results

The simulation process was carried-out in Matlab. A normal male voice was chosen as a sound source. The spectral density of this signal is plotted using spectrogram in Matlab and it is shown in Figure 4.1.

Gaussian noise was used as an additive noise, this noise signal is pseudo-random noise with a normal distribution and mean value of zero and standard deviation of one. Two Gaussian noise signals were generated in Matlab and added to the voice signal. The noise

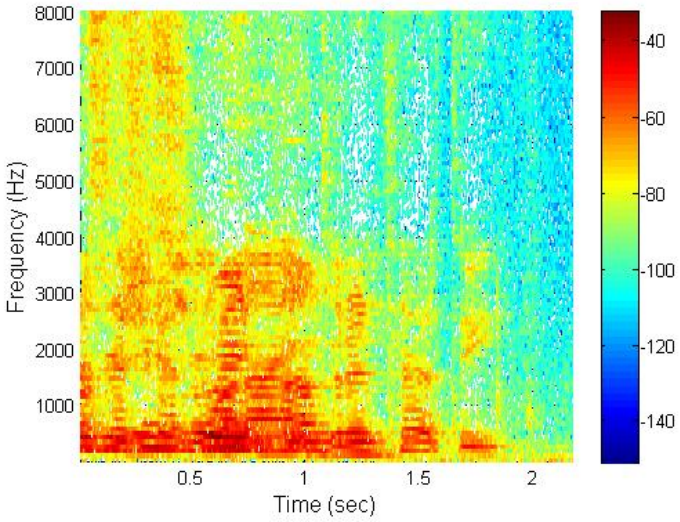


Figure 4.1: Spectral density distribution of the sound signal.

signals are additive, normally distributed and not correlated to each other. One of the resultant signals was then delayed and both signals were then applied to the methods.

Figure 4.2 illustrates the simulation results when no additive noise was used. It should be noted that the same estimated time delay can lead to different directions of the speaker depending on the distance between the two used microphones (D_m), as it will be shown in the section 4.1.2. The peak position indicates the time delay estimation. Although all methods were able to estimate the time delay of arrival perfectly when no noise was used, phase transform method (PHAT) has the sharpest mean peak with no additional peaks, which makes it the best method for time delay estimation among the others when no noise is used.

Figure 4.3 shows the simulation results when the noise signals are added to the original signal. SNR in this simulation was -12 dB. The same time delay of arrival, which was used in the previous simulation, is used here. As can be clearly seen, the three methods were able to estimate the time delay of arrival correctly. Compared to the results when no additive noise was used, additive peaks appear in the new graphs. However, the highest peak still indicates the right time delay estimation. It can be also seen that the PHAT method achieved a sharper peak than the other methods. The maximum likelihood method also achieved a sharp peak, but it has several additional smaller peaks. Although cross-correlation has the widest peak, it still can estimate the real time delay in the simulation conditions. It should be also noted that PHAT method has the biggest peak denoting the time delay estimation compared to the additive peaks around it, whereas the other methods have bigger additional peaks during the time interval caused by the additive noise [88].

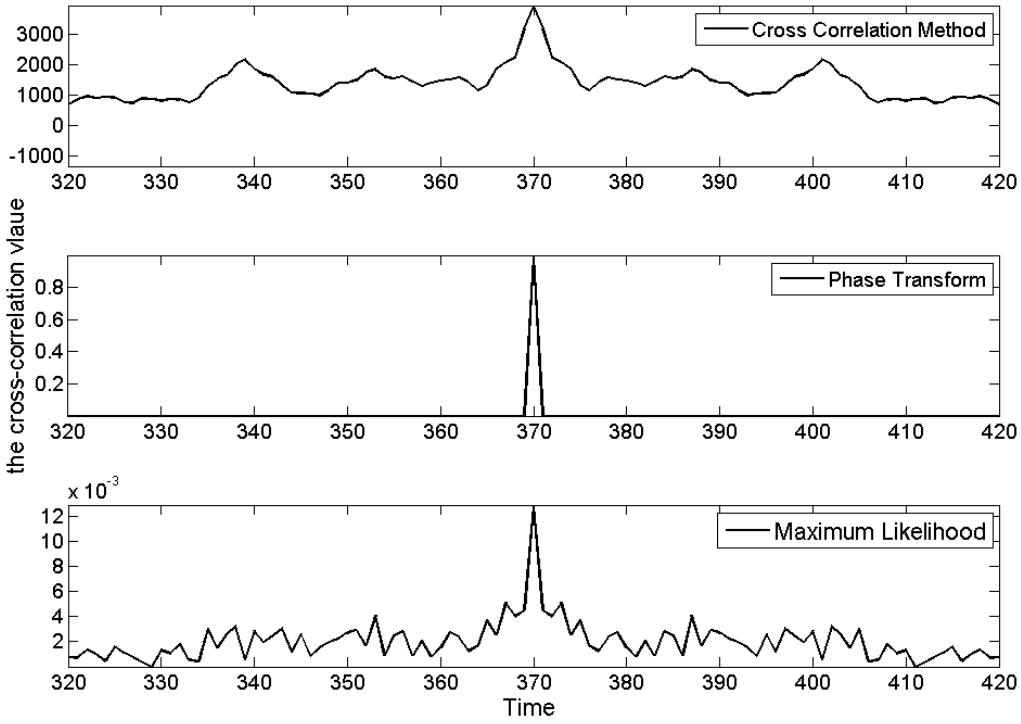


Figure 4.2: The simulation results of CC, PHAT, and ML methods with noise absence.

4.1.2 Experimental Results of Time Delay of Arrival Estimation Methods

In order to verify the simulation results, we tested the three mentioned methods in a real environment.

Microphone Arrays Used for Time Delay Estimation

It is well known that we need two microphones to estimate the time delay of arrival, which leads to the estimation of the direction of arrival of the sound source, and at least three microphones are needed to localize the sound source (direction and distance). However, in case of far-field situation, the best we can get is the direction of arrival of the sound source.

In our experiment, two different microphone arrays were used; the first array is illustrated in Figure 1.4, whereas the second one is illustrated in Figure 4.4. The distance between the microphones should be chosen carefully. It is obvious that theoretically the further the microphones are from each other, the greater time delay of arrival and the easier to estimate it. However, this is valid only under some conditions. The distance

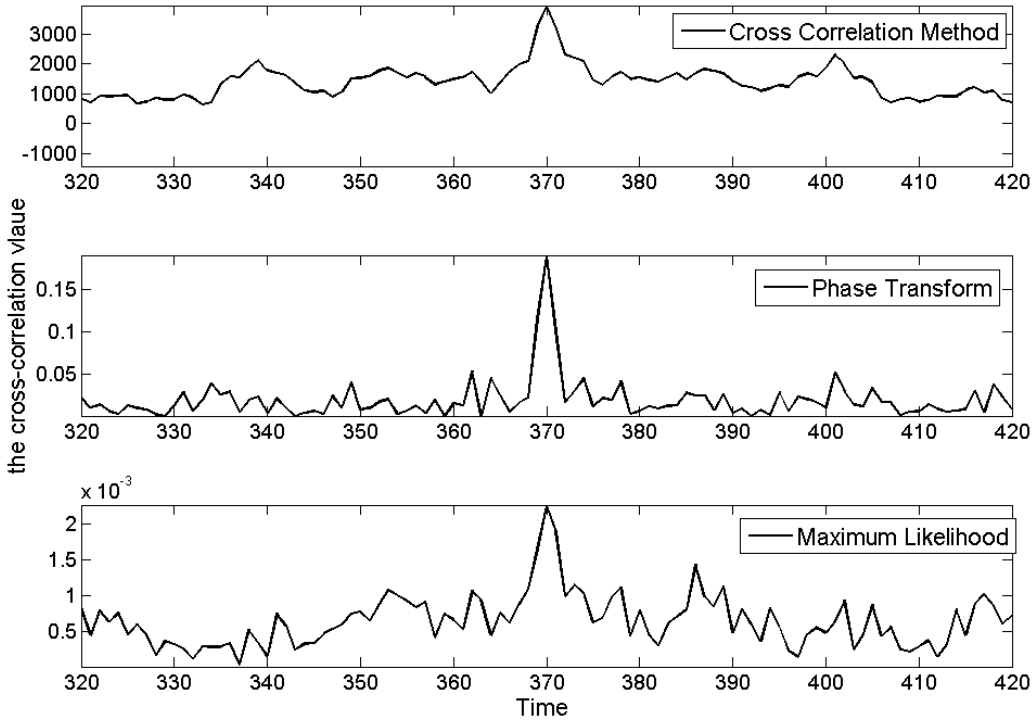


Figure 4.3: The simulation results of CC, PHAT, and ML methods with an additive Gaussian noise.

between the microphones should not exceed a half of the shortest sound wave contained in the sound signal

$$D_m \leq \frac{\lambda_{\min}}{2} \tag{4.1}$$

where D_m is the distance between two microphones and $\lambda_{\min}$ is the shortest sound wave presented in the signal. Otherwise, a spatial aliasing could happen. However, this restriction should be considered for the methods that use phase difference estimation of narrow-band signals. When we use two microphones, the time delay of arrival can be calculated using one of the above mentioned methods, and then we get the direction of arrival as

$$\vartheta = \arccos \frac{R}{D_m} = \arccos \frac{c\Delta t_{11}}{D_m} \tag{4.2}$$

where ϑ is the direction of arrival, Δt_{11} is the time delay of arrival between the microphones, D_m is the distance between the microphones, c is the sound speed and R is the extra distance the sound travels to the furthest microphone see Figure 1.4.

When three microphones are used, we can get different microphone arrays' shape, depending on the geometry of the distribution of the microphones. However, we chose to

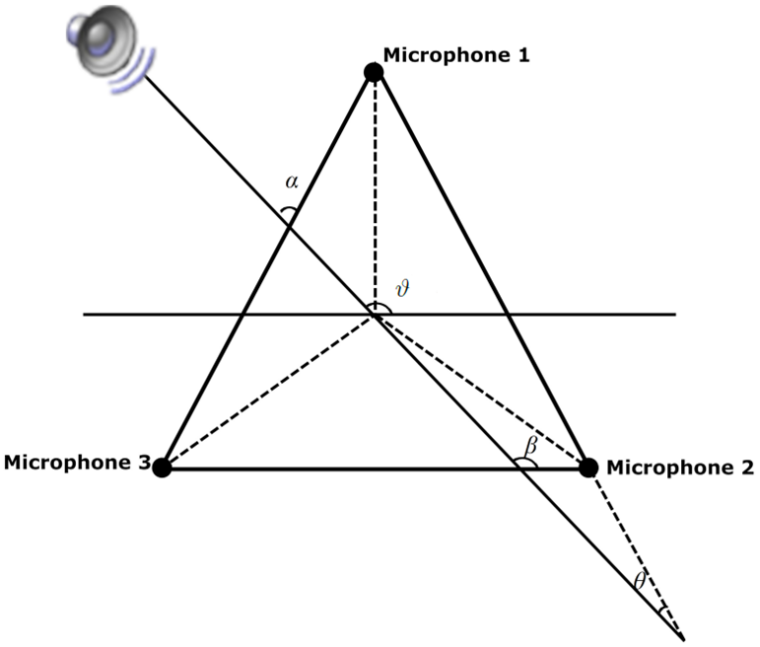


Figure 4.4: Microphone array with three microphones.

put the microphones in the heads of equilateral triangle, see Figure 4.4. In this case, three different TDOAs are estimated, we estimate TDOA between the microphones in the pairs (m_1, m_3), (m_3, m_2) and (m_2, m_1), which leads to three different DOAs values.

$$\begin{aligned} \alpha &= \arccos \frac{c\Delta t_{13}}{D} \\ \beta &= \arccos \frac{c\Delta t_{32}}{D} \\ \theta &= \arccos \frac{c\Delta t_{21}}{D} \end{aligned} \tag{4.3}$$

where α, β, θ are the angles pointing to the sound source, see Figure 4.4, Δt_{13} is TDOA between the microphones (m_1, m_3), Δt_{32} is TDOA between the microphones (m_3, m_2) and Δt_{21} is TDOA between the microphones (m_2, m_1).
From these different angles we can get the angle ϑ that represents the DOA for the center of the microphone array as

$$\begin{aligned} \vartheta &= \alpha + 60 \\ \vartheta &= -\alpha + 60 \\ \vartheta &= \beta \\ \vartheta &= -\beta \\ \vartheta &= \theta + 120 \\ \vartheta &= -\theta + 120. \end{aligned} \tag{4.4}$$

The angle ϑ is then estimated as the average of the two closest angles to each other calculated by eq. (4.4).

Experimental Results

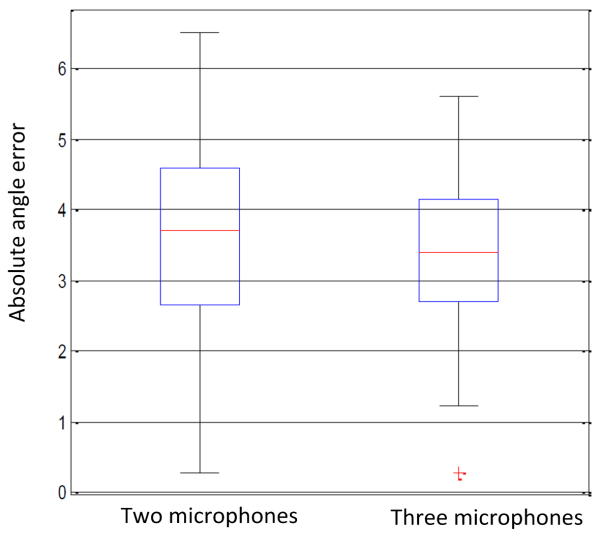
The experiment was performed in an acoustic laboratory presented in appendix A as follows: In the first part of our experiment, two microphones were located in the middle of the laboratory; the distance between the microphones was chosen to be 16.5 cm. In the second part, we used three microphones located as shown in Figure 4.4. In the both parts, the sound was recorded in thirty six different angles in the front side of the microphone array. These angles were equidistantly separated (i.e., 5 degrees from each other). The microphones, which were used in the experiments, are MiniSPL microphones from NTi Audio Company, which is in accordance with IEC 61672 class 2 microphone. This microphone has omni-directional polar pattern [89]. After having the sounds recorded, we applied them on the three mentioned methods. The results are illustrated using boxplot where the center red line presents the median, the upper blue edge of the box is 75% percentile, the lower blue edge indicates 25% percentile, and the whiskers present the extreme data points. In the following paragraphs, we will present the results for these methods:

1. **Cross Correlation Method**
   When two microphones were used, this method achieved a median error bigger than 3.5 degrees, and the biggest error was more than 6 degrees. As can be seen in Figure 4.5a, adding the third microphone improved the results of this method with more than 0.3 degree for the median error and more than one degree for the biggest error. The interquartile range has also been improved.

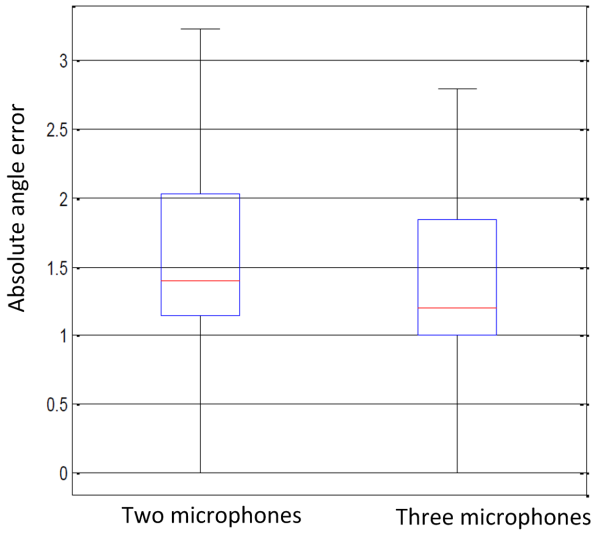
2. **Phase Transform Method**
   The results show that the PHAT method achieved the good results with a median error less than 1.5 degrees when two microphones were used, and it achieved better result in the case of three microphones. However, using the third microphone had a small effect on the center quartile range of this method, as can be seen in Figure 4.5b.

3. **Maximum Likelihood Method**
   The results show that adding the third microphones slightly improved the results when the maximum likelihood method was used. However, the median error was almost the same, see Figure 4.5c.

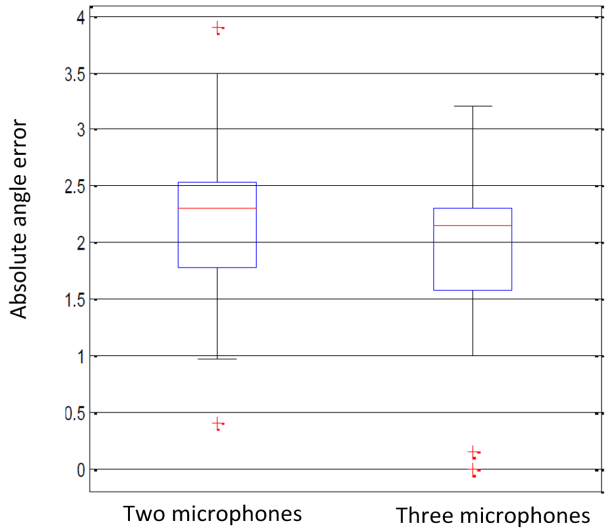
Figure 4.6a shows the results of the three methods when two microphones are used. As can be clearly seen the PHAT method achieved the best results with median error less than tow degrees, with almost one degree different from ML method. Even more, the interquartile range (IQR) between the first quartile and the third quartile is also small (approximately one degree). These results qualify PHAT to be the most accurate method among the tested methods. However, ML method achieved also the good results compared to the other methods with median error about two degrees. As can be also seen, its IQR



(a) CC method.



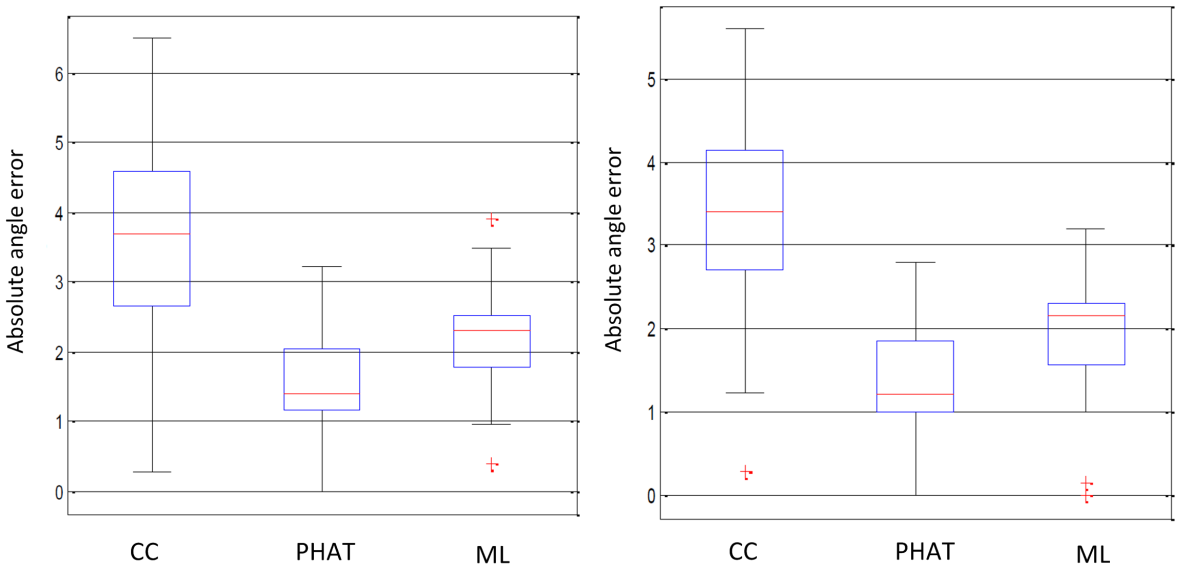
(b) PHAT method.



(c) ML method.

Figure 4.5: The average absolute angle error when CC, PHAT and ML methods are used.

was also about one degree. The worse results in the experiments were achieved when CC method was used, with median error more than three degrees and with almost two degrees IQR. However, the experimental results were expected regarding the simulation results



(a) Two microphones. (b) Three microphones.

Figure 4.6: The average absolute angle error when two and three microphones are used.

that have been shown in section 4.1.1. The error in the results is caused because of the noise, reverberant signals, and the correlation between these signals and the original sound signal.

As can be seen in Figure 4.6b, having added the third microphone did not actually improved the direction of arrival estimation when the sound sources were located in the front side of the microphone array. However, it provides extra information that can be used for sound source localization, i.e. the distance of the microphone array.

This information about the sound source position can be derived using so-called *triangulation* once the time delay of arrival is estimated [90].

4.2 Localization Blur of 2D Ambisonic and VBAP

A part of this work is devoted to design a rendering system with acoustic zooming. Therefore, an investigation about the most suitable rendering method has been made. Regarding the overview, which has been introduced in chapter 1, the choice has been shortened to Ambisonic and VBAP [91]. Three Ambisonic decoders have been tested; namely, energy decoder (max r_E), velocity decoder and in-phase decoder. These decoders differ in the weighting factor w$[q_d]$, which can be given according to Table 4.1, where q_d is the actual Ambisonic order and Q_d the highest Ambisonic order [92].

decoder	velocity	max r_E	in-phase
weights w$[q_d]$	1	$\cos\left(\frac{q_d\pi}{2Q_d+2}\right)$	$\frac{Q_d!^2}{(Q_d+q_d)!(Q_d-q_d)!}$

Table 4.1: Decoder weights.

In following, we present our evaluation of these methods, and we explain the conditions at which this evaluation has been done.

4.2.1 Description of the Experiment

The experiment was performed in an acoustic laboratory, see appendix A.

Loudspeaker Array

A regular hexagon array was used in our experiment, with array radius of 2.5 m. Figure 4.7 shows this array with the listener's positions, where the colored points indicate the listener's positions.

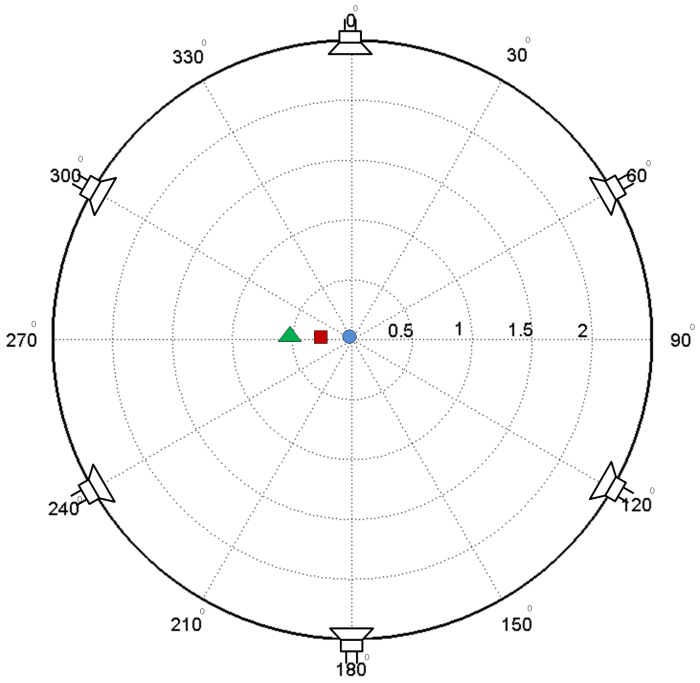


Figure 4.7: The used positions of the listener and loudspeakers.

Stimulus

An amplitude-keyed pink noise has been chosen as a stimulus, because of its characteristics, whereas it has equal energy in all octaves bands in whole audio frequency range. The stimulus was divided into four periods; each period has a rise time and fall time of 100 ms with 200 ms of an attenuated noise in between. The periods are separated by 100 ms of silence [93].

Listening Test

A psychoacoustic method of equivalent stimuli with hidden reference was chosen for the experiment [94]. The experiment has been carried out as follows. At first, the listeners have been chosen without any hearing impairment, in age from 25 to 35 years. Each listener was seated at the position of measurement with closed eyes, listening to the Ambisonic signal, without any prior knowledge about the sound position. However, the loudspeaker layout was known to the listener. Then another loudspeaker is used to check the position where the listener thinks the sound comes from. The operation was repeated many times till the listener was sure about the position (a listener heard two sounds from the same direction), then the place was marked. At the end of each measurement, the angles were measured carefully to be compared with the angles they should be. The experiment for each decoder was carried out in three positions, the first position was at the center of loudspeakers (sweet spot), then the position of the listener was shifted to the left by 0.25 m, and the third position was 0.5 m far to the left from the center.

4.2.2 Experimental Results

The results are illustrated using box plots. The boxes show the average absolute angle error between the position where we wanted the virtual sound to be, and the position that the listener perceives the sound to be coming from. The results are shown for each decoder in the three positions the listen was seated in.

Ambisonic Panning Using Energy Decoder

The results show that the localization of energy decoder is very good at the sweet spot, but worsens greatly by moving outside of the sweet spot, see Figure 4.8. The median error and interquartile range at positions 0.25 m and 0.5 m are almost unchanged.

Ambisonic Panning Using In-Phase Decoder

The in-phase decoder gave very consistent results. While the average localization accuracy worsens by moving away from the sweet spot, the interquartile range remains almost unchanged at all positions, see Figure 4.9.

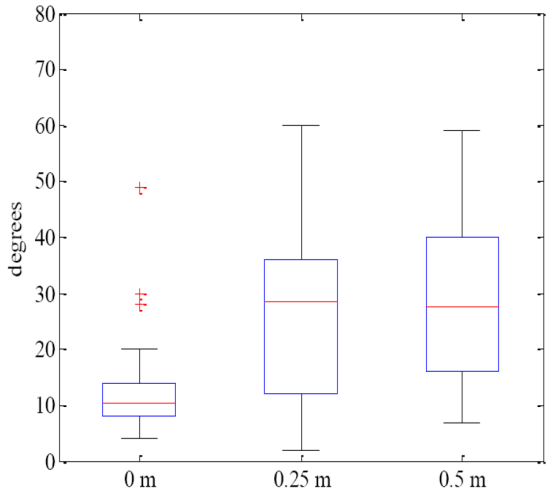


Figure 4.8: The average absolute angle error when max r_E decoder is used.

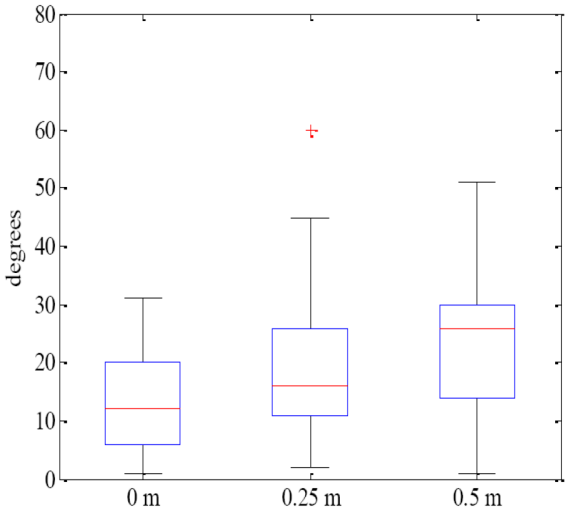


Figure 4.9: The average absolute angle error when in-phase decoder is used.

Ambisonic Panning Using Velocity Decoder

Figure 4.10 shows the localization error of the velocity decoder. While the sudden raise of the median error by moving away from the sweet spot is similar to the energy decoder, the interquartile range is greater at the sweet spot and 0.5 m away from the sweet spot. The decrease of interquartile range at the position 0.25 m away from the center is yet to be explained, but may be due to the out-of-phase components which are attenuated when using the energy decoder.

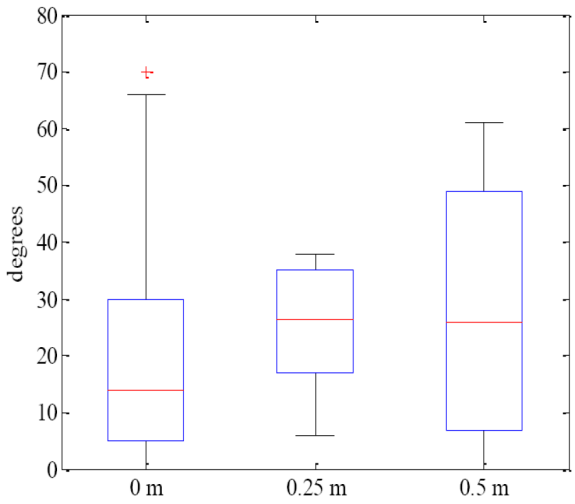


Figure 4.10: The average absolute angle error when velocity decoder is used.

Vector Base Amplitude Panning

The performance of VBAP method is very good in compression to the Ambisonic decoders. The localization error rises only slightly when moving out of the center as can be seen in Figure 4.11.

Performance of Decoders

At sweet spot, the Ambisonic decoders have similar median error, see Figure 4.12. The interquartile range is the best for the energy decoder and the worst for the velocity decoder with the in-phase decoder in the middle. But the best localization is achieved using the vector base amplitude panning method.

The performance of the positioning methods decreased at the position 0.25 m away from the center, see Figure 4.13. Among Ambisonic decoders, the in-phase decoder was the least affected by the movement. The VBAP method has the best localization accuracy at this position.

At the position 0.5 m away from the sweet spot the median error of the Ambisonic decoders is almost identical to each other, see Figure 4.14. The velocity decoder has a

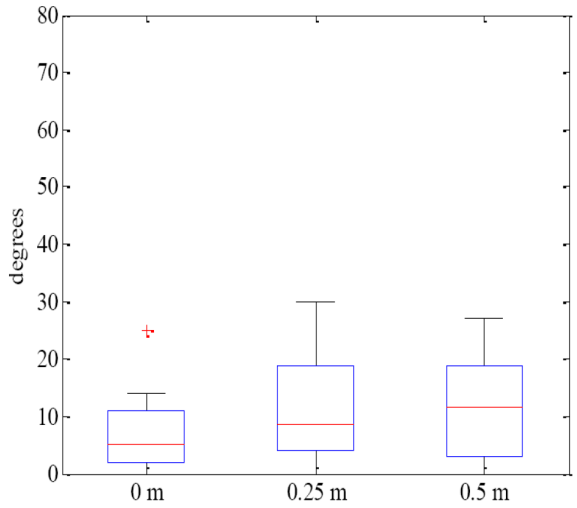


Figure 4.11: The average absolute angle error when VBAP method is used.

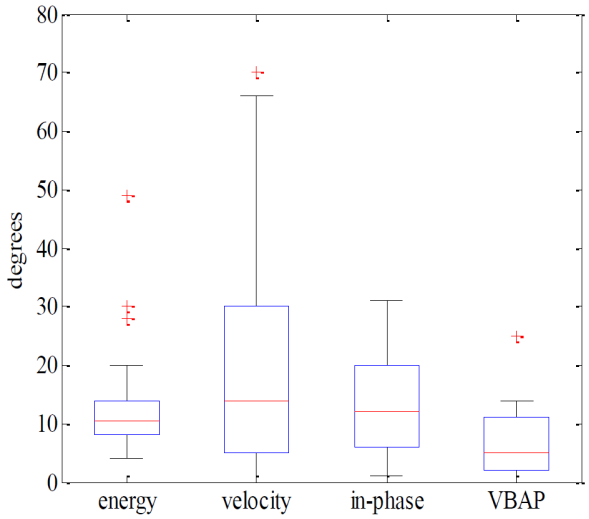


Figure 4.12: The average absolute angle error of the three Ambisonic decoders and VBAP in central position.

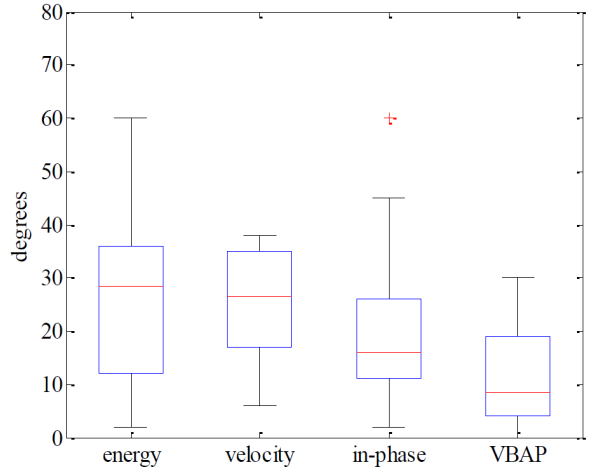


Figure 4.13: The average absolute angle error of the three Ambisonic decoders and VBAP at 0.25 m far from the center.

significant interquartile range meaning the measured values had the most spread at this position. The in-phase decoder's interquartile range remains unchanged. The VBAP method proved to be the best positioning method for this position.

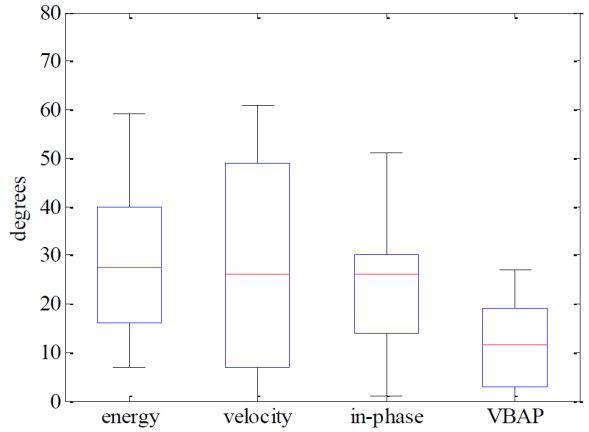


Figure 4.14: The average absolute angle error of the three Ambisonic decoders and VBAP at 0.5 m far from the center.

The results show that VBAP has a better localization performance at all positions than first-order Ambisonic. Both methods achieved their best at the sweet spot. However, the listeners tend to judge the direction of the VBAP virtual sound source to be more in the direction of the loudspeakers as they are sitting far from the sweet spot, whereas the Ambisonic virtual sound source become wider and sometimes the listeners can recognize two or more correlated sounds coming from different distances. VBAP produces a sharper virtual sound source than the one produced by Ambisonic [95].

4.3 Sub-Conclusion

This chapter studied the accuracy of several sound source localization methods. The simulation results showed that PHAT method has the best performance among the studied methods, which was proven by the experimental results as well. Both ML and CC methods were able to estimate the direction of the simulated sound source under noise. However, ML achieved better results in the experiments.

The performance of three Ambisonic decoders have been studied and compared to the performance of VBAP at several positions near the center of loudspeaker array. The listening tests showed the localization blur of VBAP was the best, whereas the localization blur of the first-order Ambisonic decoders worsened outside the sweet-spot.

Chapter 5

Estimation of the Direction of Arrival of Multiple Speakers

This chapter gives a theoretical and practical concept of a new method for sound source direction estimation. The new method is called Energetic Analysis Method. It can be used for sound source direction estimation in both two dimensional and three dimensional plane. It can be also used for tracking a speaker in the horizontal plane. Simulation and the experimental results of this method approved its accuracy. Thus, we will used it later as a compatible method with acoustic zooming system.

The first section of this chapter introduces B-format signal, and discusses the possibility of using the principle of this signals in sound direction estimation directly. The second section presents the energetic analysis method principle, its simulation and experimental results, and the possibility of using it in tracking a mobile target.

5.1 B-Format Signals

B-format signals consist of four signals; namely, $x(t)$, $y(t)$, $z(t)$ and $w(t)$, see Figure 5.1. The signal $w(t)$ represents the pressure signal in a point and it corresponds to the output of an omni-directional microphone. The signals $x(t)$, $y(t)$ and $z(t)$ represent the velocity components of the sound field in the same point, and they correspond to the output of figure-of-eight microphones. Whereas $x(t)$ and $y(t)$ carry information about the sound field in the horizontal plane, the signal $z(t)$ carries information about the sound field in the vertical plane.

B-format signals can be expressed as [96]

$$\begin{aligned} x(t) &= s(t)\cos(\alpha)\cos(\beta) \\ y(t) &= s(t)\sin(\alpha)\cos(\beta) \\ z(t) &= s(t)\sin(\beta) \\ w(t) &= \frac{1}{\sqrt{2}}s(t) \end{aligned} \tag{5.1}$$

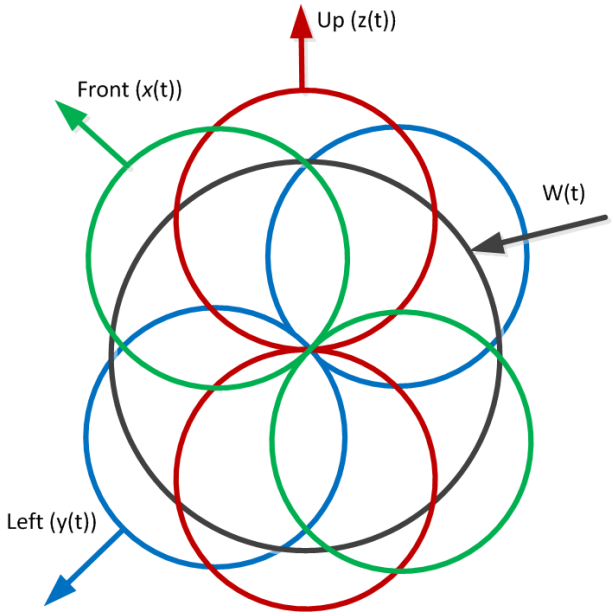


Figure 5.1: Polar patterns of B-format components in horizontal plane.

where $x(t)$, $y(t)$, $z(t)$ and $w(t)$ are the B-format signals, $s(t)$ is the sound signal, α is the angle in the horizontal plane and β is the angle in the vertical plane, see Figure 5.2.

It should be noted that $\beta = 0$ when the sound source is located in the horizontal plane. In the horizontal plane, the previous equation can be simplified as

$$\begin{aligned} x(t) &= s(t)\cos(\alpha) \\ y(t) &= s(t)\sin(\alpha) \\ w(t) &= \frac{1}{\sqrt{2}}s(t). \end{aligned} \tag{5.2}$$

B-format equations suppose that the microphones are centered in the same point, which is practically impossible to achieve. However, the microphones are placed as close as possible to each other. When B-format signals are recorded, a second problem appears because of the poor polar diagrams of the microphones [97].

In order to achieve better results, a solution which was developed by Gerzon in [97] suggests using a tetrahedral microphones. The mentioned microphone contains four cardioid or (near-cardioid) capsules mounted as close as possible to each other at the vertexes of a regular tetrahedron. The practical version of this principle is Soundfield microphone, see Figure 5.3.

The signals estimated directly from Soundfield microphone are A-format signals, the

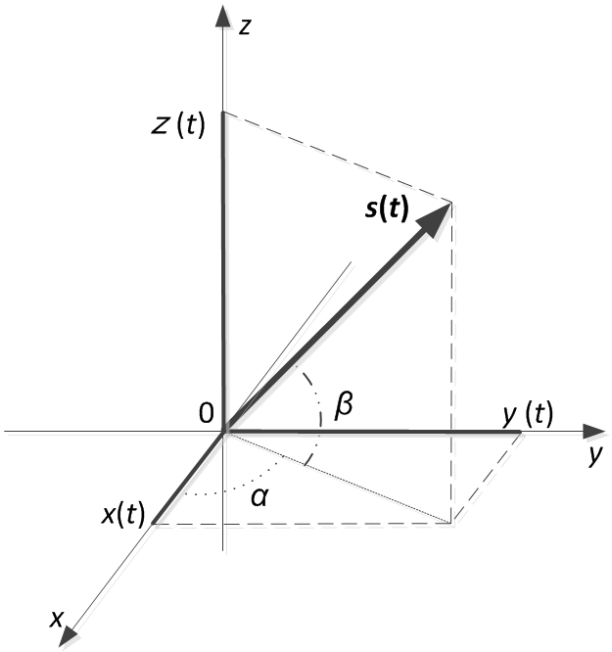


Figure 5.2: The used coordinate system.

B-format signals are then derived from A-format signals using the equations [35]

$$\begin{aligned} x(t) &= 0.5((lf(t) - lb(t) + (rf(t) - rb(t)) \\ y(t) &= 0.5((lf(t) - rb(t) - (rf(t) - lb(t)) \\ z(t) &= 0.5((lf(t) - lb(t) + (rf(t) - rb(t)) \\ w(t) &= 0.5((lf(t) + lb(t) + (rf(t) + rb(t)) \end{aligned} \tag{5.3}$$

where $x(t)$, $y(t)$, $z(t)$ and $w(t)$ are B-format signals, and $lf(t)$, $rf(t)$, $lb(t)$ and $rb(t)$ correspond to the signals recorded by the capsules left-front, right-front, left-back and right-back respectively.

5.1.1 Sound Source Localization Using B-format Signal

Recalling the equations that represent B-format signals, it can be seen that the direction of a single sound source can be derived directly from these equations [98]. The direction

Figure 5.3: Soundfield microphone.

of arrival can be estimated from B-format equations as

$$\begin{aligned} \alpha &= \arctan \frac{y(t)}{x(t)} \\ \beta &= \arctan \frac{z(t)}{\sqrt{x(t)^2 + y(t)^2}} \end{aligned} \tag{5.4}$$

where α and β are the angles denoting the sound source in both horizontal and vertical planes respectively and $x(t)$, $y(t)$, $z(t)$ are B-format signals.

The previous assumption holds in a simulation environment when no additive noise is present. However, a presence of noise complicates the situation.

This technique has been investigated in a real environment, where the voice of one speaker was recorded using two figure-of-eight microphones and one omni-directional microphone.The microphones were placed as close as possible to each other and in a way that their directionality match the theoretical place of B-Format signals. The microphones were then located in the middle of the laboratory, and the experiment was performed for twenty different places. An external 6-channel audio interface was used with ASIO (audio streaming input/output) driver technology, which ensures in-phase recording of all channels.

The experimental results of this technique showed the possibility of using B-format signals directly in order to estimate the direction of one speaker. The angle error did not exceed 5 degrees. This error was caused mostly by the reverberation in the room and the inability to place the microphones in a single point as they supposed to be according to B-format principle. However, the results were acceptable for estimation the direction of one speaker [98].

5.2 Energetic Analysis Method

Energetic analysis method is a technique for sound source direction estimation, based on analyzing B-format signals [99]. This method is able to estimate the direction of multiple speakers in both two and three dimensional plane.

5.2.1 Principle of Energetic Analysis Method

Energetic analysis method is derived from an acoustic principle which states that the sound source direction is the opposite direction of the intensity vector of the sound. This principle is used in some sound rendering methods, such as in DirAC [62]. However, it is modified in our method to ensure the stability of the sound direction estimation process.

Acoustic Intensity

The acoustic intensity $\overrightarrow{I}(t)$ is a vector quantity, which defines the direction of the acoustic energy. In the time domain, it is called the instantaneous acoustic intensity and it can be written as [100]

$$\overrightarrow{I}(t) = p(t)\overrightarrow{v}(t) \tag{5.5}$$

where $\overrightarrow{v}(t)$ represents the particle velocity vector, $p(t)$ is the sound pressure and $\overrightarrow{I}(t)$ is the acoustic intensity and it can be written as

$$\overrightarrow{I}(t) = I_x\overrightarrow{x} + I_y\overrightarrow{y} + I_z\overrightarrow{z} \tag{5.6}$$

where I_x, I_y, I_z are the components of the acoustic intensity and $\overrightarrow{x}$, $\overrightarrow{y}$, $\overrightarrow{z}$ are unit vectors in the three dimensional plane.

The energy density of the sound wave is given by equation [100]

$$e = \frac{1}{2}\rho_0|\overrightarrow{v}(t)|^2 + \frac{1}{2}f_l p(t)^2 \tag{5.7}$$

where $p(t)$ is the sound pressure, ρ_0 is the density of the air and f_l represents the gas compressibility and it is given by the equation

$$f_l = \frac{1}{\rho_0 c^2} \tag{5.8}$$

where c is the speed of the sound.

The energy density can be written depending on the sound intensity as [100]

$$\begin{aligned}\frac{\partial(e)}{\partial(t)} &= -\overrightarrow{\nabla}.\overrightarrow{I} \\ &\equiv -\left(\frac{\partial(I_x)}{\partial(x)} + \frac{\partial(I_y)}{\partial(y)} + \frac{\partial(I_z)}{\partial(z)}\right)\end{aligned} \tag{5.9}$$

where the first part of this equation is used for all coordinate systems, and the second one is used for a Cartesian coordinate.

The previous equation clearly shows the principle of the energetic analysis method. It shows that the intensity vector points to the region of increase of the energy density, which is the principle which this method is based on.

In case of steady state fields, average of the instantaneous intensity over a period T is defined as [100]

$$\overrightarrow{I}(\omega) = \frac{1}{T}\int_0^T p(t)\overrightarrow{v}(t)\mathrm{d}t \tag{5.10}$$

where $T = \frac{1}{f}$ is the time period, f is the frequency of the excitation signal and ω is the angular frequency.

By using complex variable notation, eq. (5.10) can be written as

$$\overrightarrow{I}(\omega) = \frac{1}{2}\mathrm{Re}(p(\omega)\overrightarrow{v}(\omega)^*) \tag{5.11}$$

where Re indicates the real part and $*$ means the complex conjugate.

Analyzing the Sound Signals

The input signals for this method are the B-format signals presented in eq. (5.1). The signals $x(t)$, $y(t)$ and $w(t)$ are used when the method aims at estimating the direction of the sound sources in the horizontal plane. An extra signal $z(t)$ is needed for the localization in the vertical plane.

Applying eq. (5.11) to B-format signals, and recalling that the signal $w(t)$ stands for the pressure and the signals $x(t)$, $y(t)$ and $z(t)$ stand for the velocity, we obtain equations representing the acoustic intensity from B-format signals as [101]

$$\begin{aligned} I_x(t,f) &= \frac{1}{\sqrt{2}Z_0}\mathrm{Re}(X(t,f)W^*(t,f)) \\ I_y(t,f) &= \frac{1}{\sqrt{2}Z_0}\mathrm{Re}(Y(t,f)W^*(t,f)) \\ I_z(t,f) &= \frac{1}{\sqrt{2}Z_0}\mathrm{Re}(Z(t,f)W^*(t,f)) \end{aligned} \tag{5.12}$$

where $I_x(t,f)$, $I_y(t,f)$, $I_z(t,f)$ are the components of the intensity vector, Z_0 is the acoustic impedance of the air, $X(t,f)$, $Y(t,f)$, $Z(t,f)$ and $W(t,f)$ are Fourier transforms of the B-format signals.

Thereby, the instantaneous intensity vector can be written as

$$\overrightarrow{I}(t,f) = [I_x(t,f), I_y(t,f), I_z(t,f)] \tag{5.13}$$

where $\overrightarrow{I}(t,f)$ is the intensity vector in the three dimensional plane.

Calculation of the Angles

As it was mentioned before, the intensity vector points to the direction of the net flow of sound energy, while the direction of arrival is supposed to be opposite to this direction. After calculating the components of the intensity vector using eq. (5.12), the azimuth for each frequency bin in each time frame can be derived as [101]

$$\alpha(t,f) = \begin{cases} \arctan\left[\dfrac{-I_y(t,f)}{-I_x(t,f)}\right] & \text{for } \; I_y(t,f) \geq 0 \\ \arctan\left[\dfrac{-I_y(t,f)}{-I_x(t,f)}\right] - 180^\circ & \text{for } \; I_y(t,f) < 0. \end{cases} \tag{5.14}$$

The elevation can be calculated as

$$\beta(t,f) = \arctan\left[\frac{I_z(t,f)}{\sqrt{I_x(t,f)^2 + I_y(t,f)^2}}\right]. \tag{5.15}$$

It should be noted that calculating the azimuth relies only on two components of the intensity vector, whereas calculating the elevation needs the three components. Therefore, the error in the estimated elevation may be bigger than the error in the azimuth.

Estimation of the Direction of Arrival

In order to estimate the direction of arrival from recorded B-format signals, at the beginning we need to represent the B-format signals in the time-frequency domain using one of the mentioned transforms in Chapter 2. Then we get the direction of arrival for each point in the time-frequency plane by applying eq. (5.14) and eq. (5.15) to each frequency bin, see Figure 5.4.

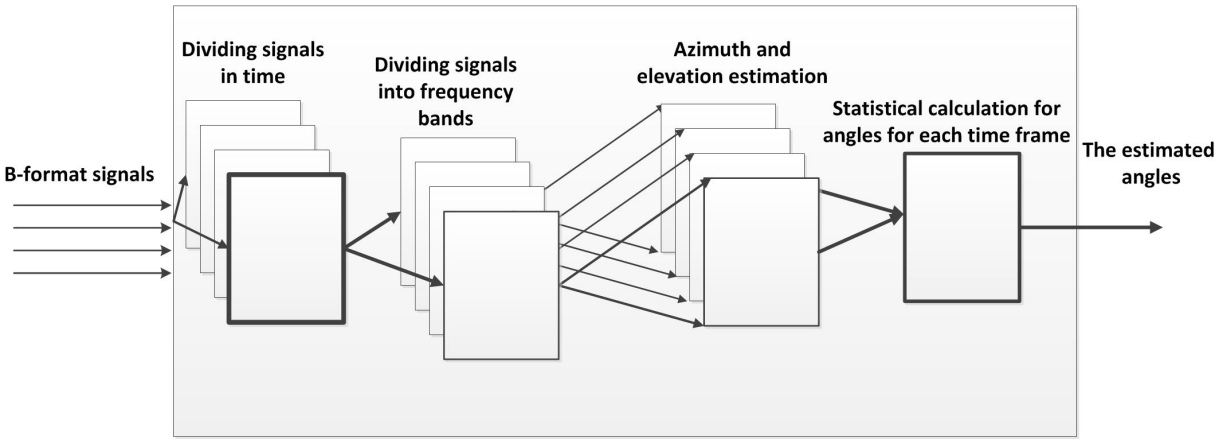


Figure 5.4: Energetic analysis method diagram.

It is assumed here that only one dominant speaker exists in each frequency bin, even it might be not completely correct. However, this assumption is acceptable since the sound of the speakers cannot be identical, which leads to the differences in the spectral density of their sound. After calculating the direction of arrival related to each frequency bin, we define the estimated direction of the sound source as

$$\alpha_{est} = \arg\max_{\alpha} F(\alpha) \tag{5.16}$$

where $F(\alpha)$ represents the number of the frequency bins pointing to the direction α and it is calculated for each angle as

$$F(\alpha) = \sum_{k=0}^{K} (\alpha(t,k)|\alpha) \tag{5.17}$$

where $\alpha \in [-180°, 180°]$ is the azimuth, K is the number of the frequency bins, t represents the time and $\alpha(t,k)|\alpha$ contains the cases where the function $\alpha(t,k)$ points to the direction α.

The elevation is also calculated in a similar way as

$$\beta_{est} = \arg\max_{\beta} F(\beta), \tag{5.18}$$

and $F(\beta)$ is given in this case as

$$F(\beta) = \sum_{k=0}^{K} (\beta(t,k)|\beta) \tag{5.19}$$

where $\beta \in [-90°, 90°]$ is the elevation, K is the number of the frequency bins, t represents the time and $\beta(t,k)|\beta$ contains the cases where the function $\beta(t,k)$ points to the direction β.

5.2.2 Simulation Results in Horizontal Plane

The method is simulated in Matlab using different time-frequency transforms, namely Gabor transform, STFT and zero padding. The simulation is carried out in three virtual environments: with no additive noise, with virtual random noise and with real noise. The simulation supposed the existing of five speakers who speak simultaneously in different positions around the microphones in each scenario. The simulation started by generating B-format signals using eq. (5.2) regarding the positions where the speakers are assumed to be, and then the noise was added to the signals.

Simulation Results with Noise Absence

The first simulation has been carried out without an additive noise. We assumed the existence of five speakers at the positions $(-150°, -70°, 0°, 100°, 160°)$ around the micro-

phone. The STFT transform was applied using squared Hanning window with 256 samples and 50% overlapping, where the length of the window was chosen as a compromise between the resolution in the time and frequency domain as explained in the previous chapter. The same parameters were applied when Gabor transform and zero-padding were used. However, when zero-padding was used, longer duration frame was obtained by zero padding the frame. Thus, the total number of samples in each time slot is exceeded to the next power of two. The total number of the padded zeros in the simulation was 3 window-size zeroes. Figure 5.5 shows the simulation results in this case. As can be seen, the method was able to estimate the direction of multiple speakers, where the peaks denote the estimated directions. When no additive noise is added, the method is perfectly able to estimate the direction of multiple speakers in the horizontal plane.

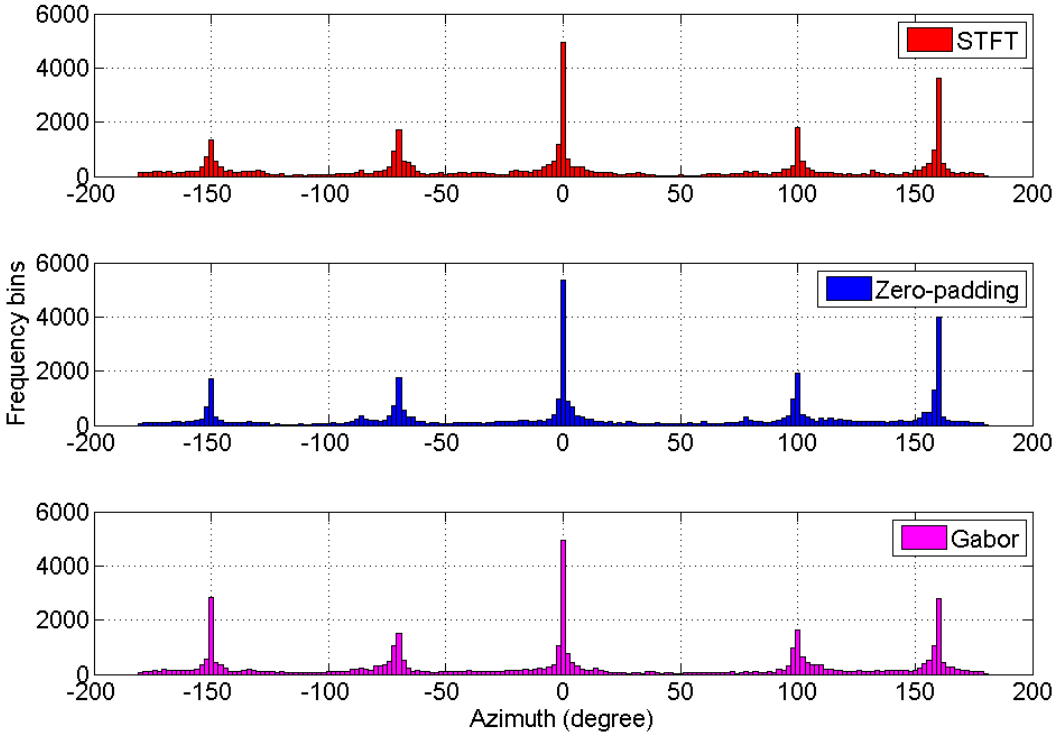


Figure 5.5: Simulation result in absence of noise.

Simulation Results with an Additive Noise

The simulation was carried out using the same parameters for each time-frequency transform. The only difference in this simulation from the previous one is adding an additive noise. This noise signal was generated in Matlab, and it is pseudo-random noise with a normal distribution and mean value of zero and standard deviation of one. The SNR in this

simulation is −20 dB. The noise signal sources are assumed to be around the microphones

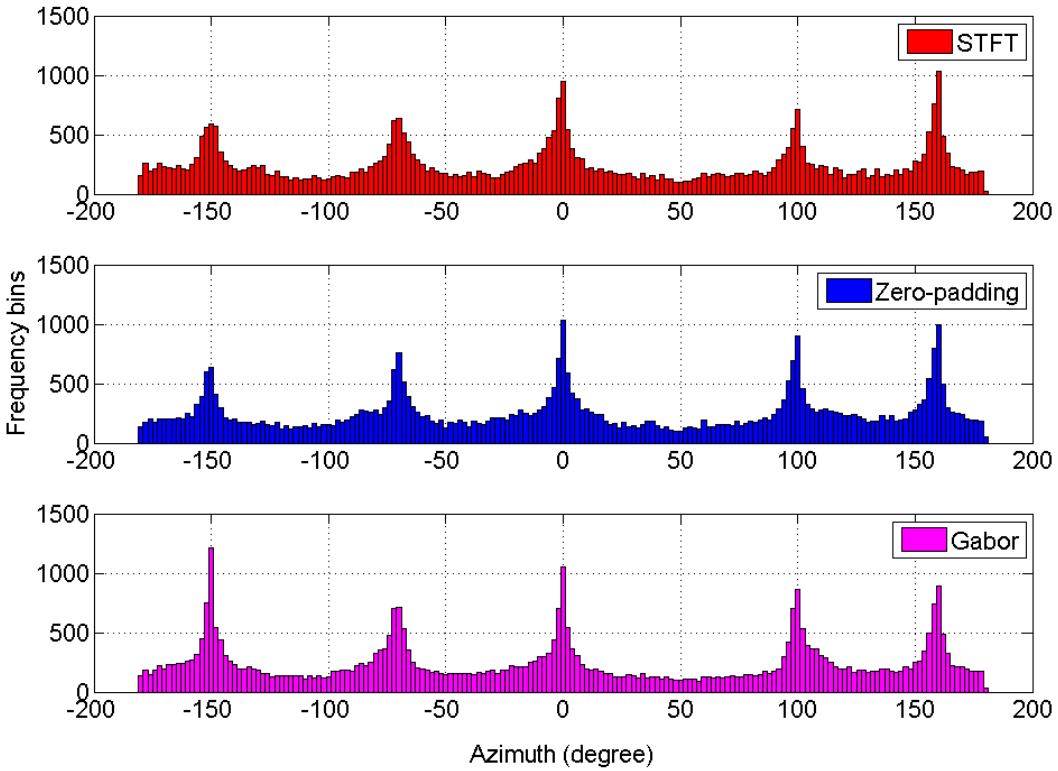


Figure 5.6: Simulation result with the presence of an additive noise.

in the horizontal plane, and to be equidistantly separated (i.e., 4 degrees from each other in the horizontal plane). As can be seen in Figure 5.6, the method was able to estimate the direction of the sound sources when the all time-frequency transforms were used. However, the existence of the noise signal affected the performance of the method. It can be seen that there are some frequency bins pointing to the direction of the noise sources, which decrease the accuracy of this method. However, as long as the intensity of the sound source is bigger than the intensity of the simulated noise, the method is still able to point to the direction of the sound sources. The simulation results showed the ability of the method to estimate the direction of the sound speaker when the three time-frequency transforms are used.

Simulation Results with a Real Noise

In order to simulate the real environment, a real noise of a fan was recorded and added to the sound source. A noise generated by a fan was chosen as appropriate additive noise thanks to its spectral distribution, see Figure 5.7. The SNR in this simulation was about −22 dB.

Figure 5.8 illustrates the simulation results when a real noise is used. As can be seen, the method was able to estimate the direction of arrival of the simulated sound sources. However, when STFT and zero-padding were used, additive peaks are noticed and that can lead to estimate the direction of fake sound sources that belong originally to noise sources. This effect was less noticeable when Gabor transform was used. That can be explained by the fact that Gabor transform achieves better resolution in time-frequency representation.

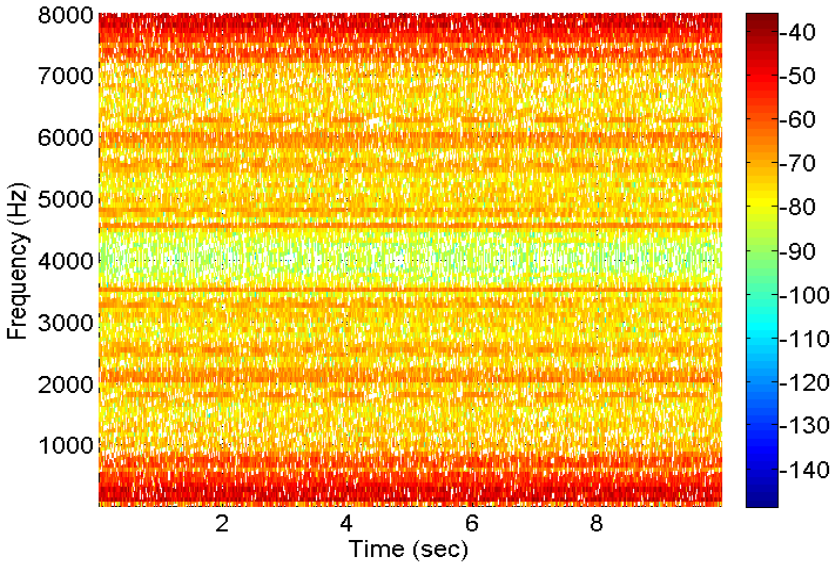


Figure 5.7: Spectral density distribution of a fan noise.

The simulation results showed the ability of this method to estimate the direction of arrival of multiple sound sources in the horizontal plane. It was shown that better resolution in time-frequency representation ensures better accuracy for this method [102].

5.2.3 Simulation Results in Three Dimensional Plane

In order to estimate the direction of arrival of multiple sound sources in three dimensional space, the four B-format signals are needed. The elevation is estimated according to eq. (5.15).

Figure 5.9 shows the simulation results for this method in three dimensional plane. In this simulation, B-format signals were generated in Matlab using eq. (5.1). A random additive noise was also generated in Matlab and added to each B-format signal. We simulated five speakers around the microphones. The positions of the speakers were assumed to be in the positions $(-150\,^\circ, -80\,^\circ, -60\,^\circ, 10\,^\circ, 110\,^\circ)$ in the horizontal plane and in the positions $(-65\,^\circ, -22\,^\circ, -15\,^\circ, 25\,^\circ, 35\,^\circ)$ in the vertical plane. STFT was used in this simulation. A random noise, which was generated in Matlab, was added to the B-format signals. The SNR in this simulation was -13 dB. As can be seen in Figure 5.9, the peaks denote the

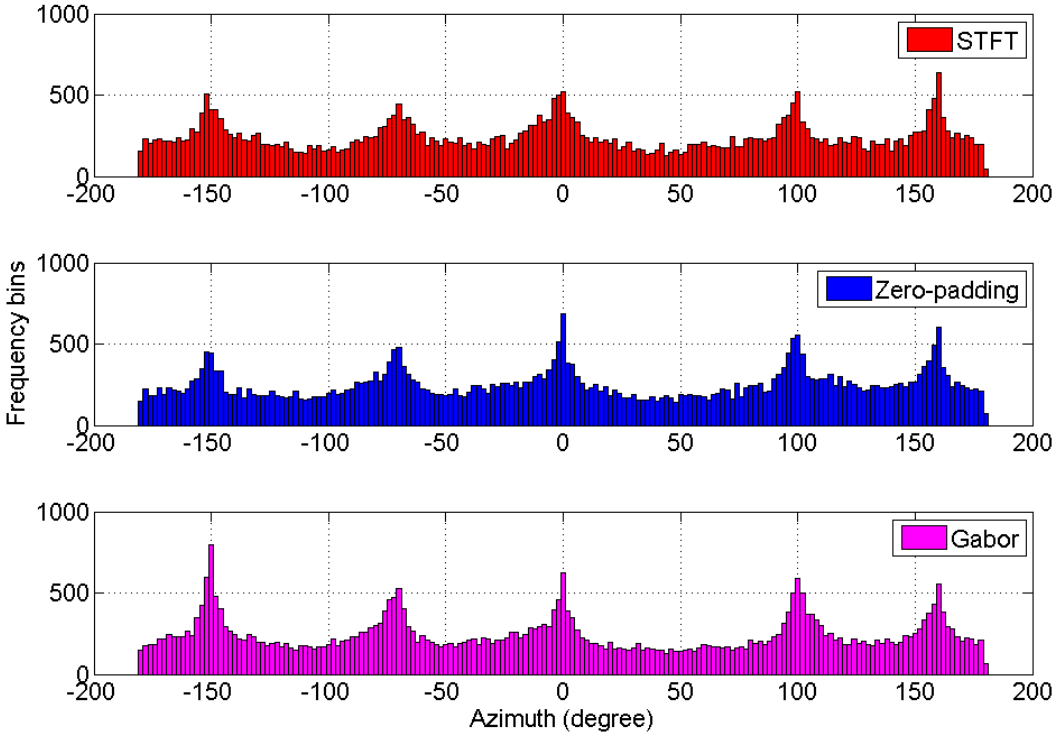


Figure 5.8: Simulation result with the presence of an additive real noise.

direction of the sound sources in both horizontal and vertical plane, and the method was able to estimate the direction of multiple speakers in 3D [103].

5.2.4 Experimental Evaluation of Energetic Analysis Method

In order to evaluate the sound direction estimation method, experiments were performed in real environments. The experiments have been carried out for both two-dimensional and three-dimensional plane. A Description of the experiments and an illustration of the experimental results are presented in the next paragraphs.

5.2.4.1 Experiment Procedure

Recording of the sound files was carried out in the acoustic laboratory, see appendix A. Three speakers (two men and one woman) were used to record the sound files. The speakers stood at different positions around the microphone. The speakers spoke simultaneously. The length of each sound file was almost 6 s. A Soundfield microphone was used to record the sound files, see Figure 5.3. This microphone records the sound as A-format signals [104].

In order to evaluate the accuracy of this method, recoding of the sound was repeated

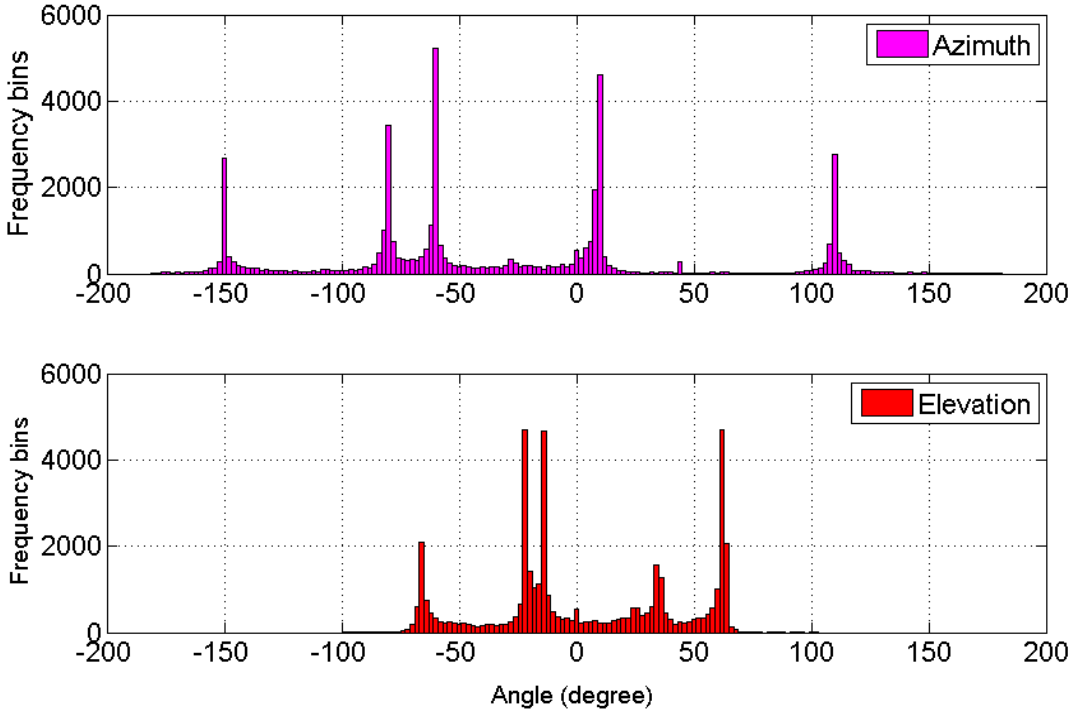


Figure 5.9: Simulation result in three dimensional plane with an additive noise.

many times, when the speakers stood at different position for each recording. The speakers stood around the microphone, where their positions have been chosen in an arbitrary way. The minimum angle between the speakers did not exceed 100 ° and it did not go below 30 °. The positions of the speakers were measured carefully to be compared with the results of the method. The positions of the speakers were measured using a leveling laser with angle scale. The absolute error of reading the scale is 2.5 ° and the horizontal accuracy of this device is 0.5 mm/m and it has the same accuracy in the vertical plane. The recording was carried out in the horizontal plane and then in the vertical plane.

5.2.4.2 Experimental Results

In following, we present the experimental results in both horizontal and vertical plane using boxplot. Each figure shows the absolute error when the method is applied for each speaker.

Experimental Results in the Horizontal Plane

The method was applied to the same sound files using different time-frequency transform, which ensures the possibility of comparison of the results depending on the resolution of the time-frequency transform in the time-frequency plane. In our experiments, we used Gabor transform, STFT and zero padding. The same parameters, which were used in the simulation, are applied here. Zero padding exceeds the number of the samples by

adding zeroes. Thus, the total number of samples in each time slot is exceeded to the next power of two. Therefore, the spectral representation accuracy is enhanced, but with no improvement of the spectral resolution.

Experimental Results when STFT is Used Figure 5.10 shows the result in the horizontal plane when STFT was applied. As can be seen from this figure, the method was able to estimate the direction of the speakers in the horizontal plane. The median absolute angle error in this case was about 4 degrees.

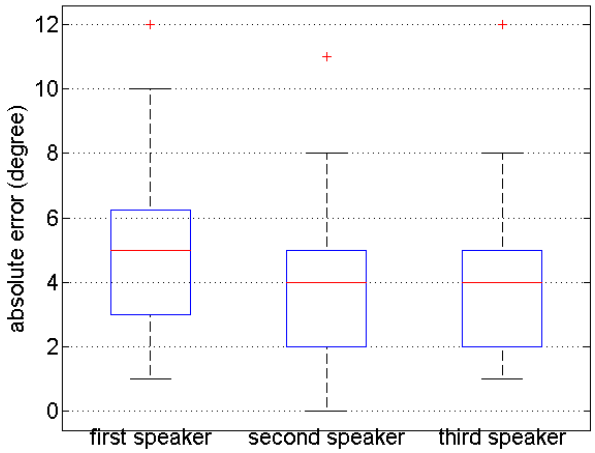


Figure 5.10: The average absolute angle error for the three speakers when STFT is used.

Experimental Results when Zero Padding is Used The average absolute angle error is illustrated in Figure 5.11. As can be clearly seen, the method was able to estimate the direction of arrival of multiple speakers who spoke simultaneously. The median absolute error was almost 3 degrees. Compared to the normal STFT, zero padding achieved better results and reduced the absolute error for each speaker.

Experimental Results when Gabor Transform is Used The results in this case are illustrated in Figure 5.12. Using Gabor transform achieved a good decrease in the absolute error. The median absolute error was gradually reduced for the speakers, and it was almost two degrees for one of them. Even more, applying Gabor transform to the method also reduced the interquartile range.

Experimental Results in the Vertical Plane

The experimental results in the vertical plane are showed in Figure 5.13. The median absolute error was almost 4 degrees. The method was able to estimate the direction of the

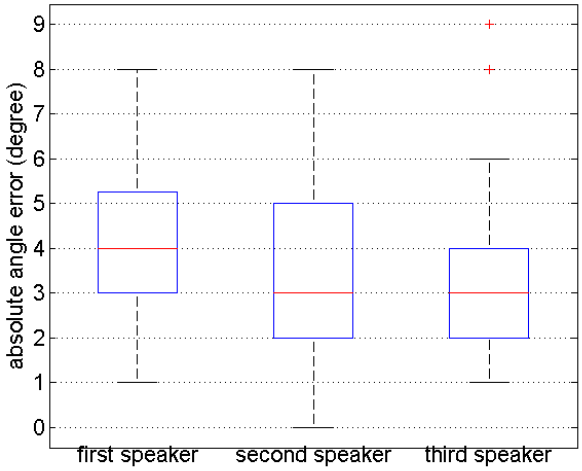


Figure 5.11: The average absolute angle error for the three speakers when zero padding is used.

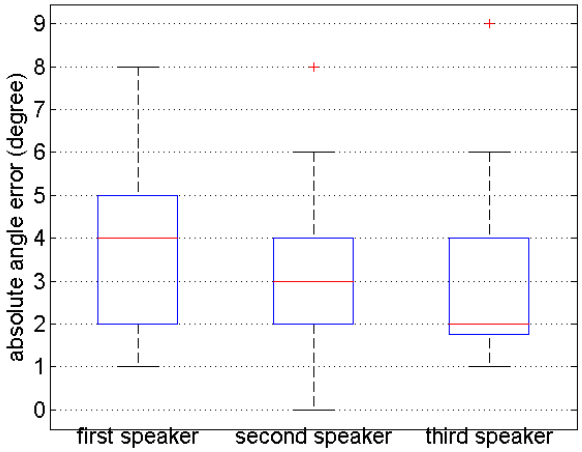


Figure 5.12: The average absolute angle error for the three speakers when Gabor transform is used.

speakers in the vertical plane. However, the median error in this case was bigger than the median error in the horizontal plane. That can be explained according to eq. (5.15), where three components of B-format signals are used to estimate the elevation, which maximizes the absolute error by involving the error that occurs when the third component of the B-format signal is recorded [103].

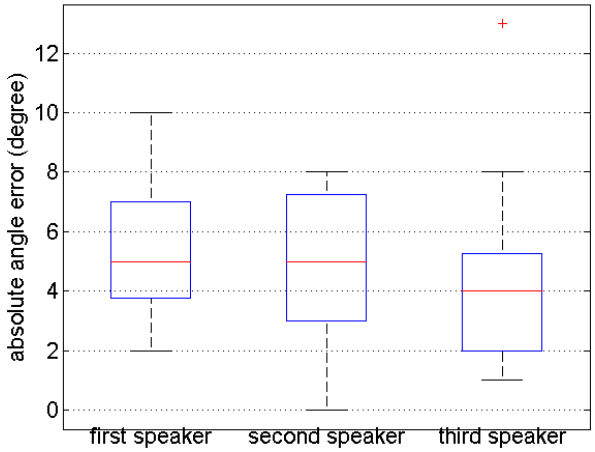


Figure 5.13: The average absolute angle error in vertical plane.

5.2.5 The Effect of SNR on the Accuracy of the Energetic Analysis Method

As for each method, the noise is the biggest problem for energetic analysis method. In order to estimate the threshold, where this method is able to estimate the direction of the speaker correctly, we simulated the method in Matlab under different signal-to-noise ratio (SNR).

SNR is calculated as the ratio between the omni-directional B-format signal $w(t)$ and the additive noise. Thereby, SNR can be written as

$$\begin{aligned} \mathrm{SNR}_{\mathrm{db}} &= 10\log(\frac{\overline{P_W}}{\overline{P_n}}) \\ &= 10\log\left(\frac{W_{\mathrm{rms}}^2}{b(n)_{\mathrm{rms}}{}^2}\right) \end{aligned} \tag{5.20}$$

where $\overline{P_W}$ is the average power of $w(t)$, $\overline{P_n}$ is the average power of the noise signal $b(n)$, $b(n)_{\mathrm{rms}}$ is the RMS value of b_n and W_{rms} is the RMS value of $w(t)$.

The simulation was carried out in case of five speakers who spoke simultaneously. The additional noise was added to the sound signals. The random noise sources were assumed to be equidistantly separated (i.e. 5 degrees from each other in the horizontal plane). The various SNR level is obtained by changing the noise power.

The simulation results under low SNR is presented in Figure 5.14. It was seen that when SNR is under a threshold, the method was not able to localize the sound sources.

The simulation results showed that the method was able to estimate the angles correctly using the three transformations when the SNR was about -20 dB. When SNR was below -23 dB, the method using STFT transform could not estimate the angles correctly, whereas

the method using zero-padding was still able to estimate the direction of sound sources when the SNR was about −24 dB. The simulation results showed that the method using Gabor transform becomes incorrect when SNR is beneath −26 dB.

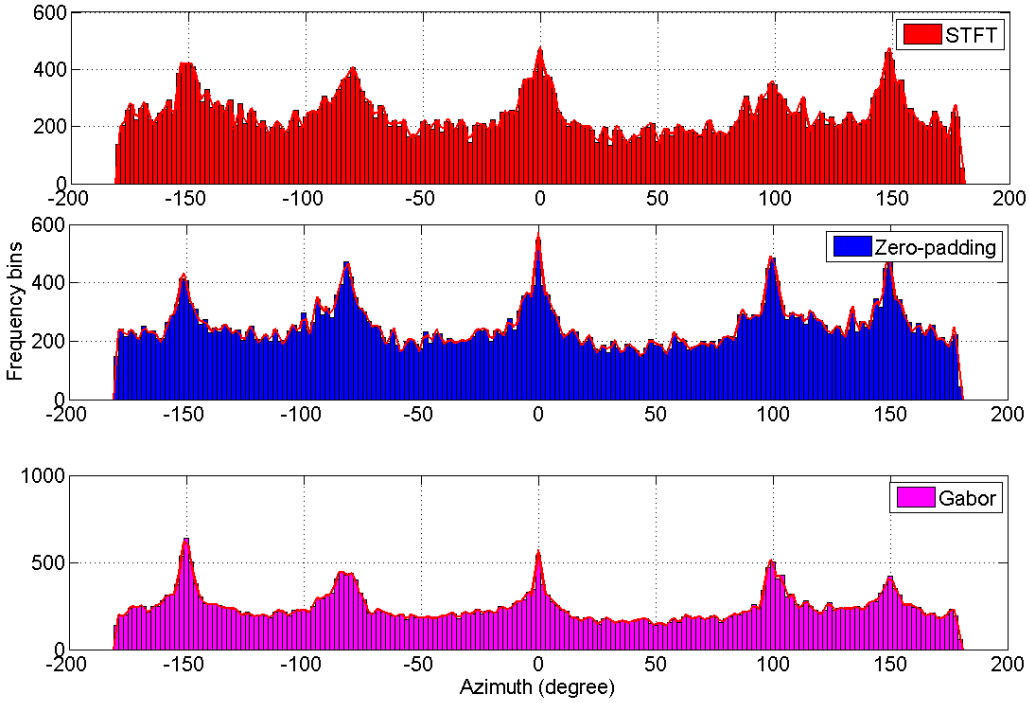


Figure 5.14: Simulation results of energetic analysis method under low SNR.

5.2.6 The Resolution of the Energetic Analysis Method

In this section, we investigate the resolution and the ability of the energetic analysis method to estimate the direction of sound sources when the sources are near to each others. In this simulation, STFT was used with Hanning window.

The first part of the simulation was performed with no noise. Figure 5.15 shows the results of this simulation, where four speakers are assumed to be around the microphones at the positions $(0\,^\circ, 1\,^\circ, 2\,^\circ, 3\,^\circ)$. The upper part of the Figure 5.15 shows the whole simulated area around the microphones in the horizontal plane i.e., from $-180\,^\circ$ to $180\,^\circ$, whereas the lowest part shows the zoomed area we are interested in.

As can be clearly seen, the peaks denote the right angles where the speakers are simulated to be. In case of no added noise, this method was able to estimate the direction of arrival of multiple speakers separated from each other with maximum of $1\,^\circ$.

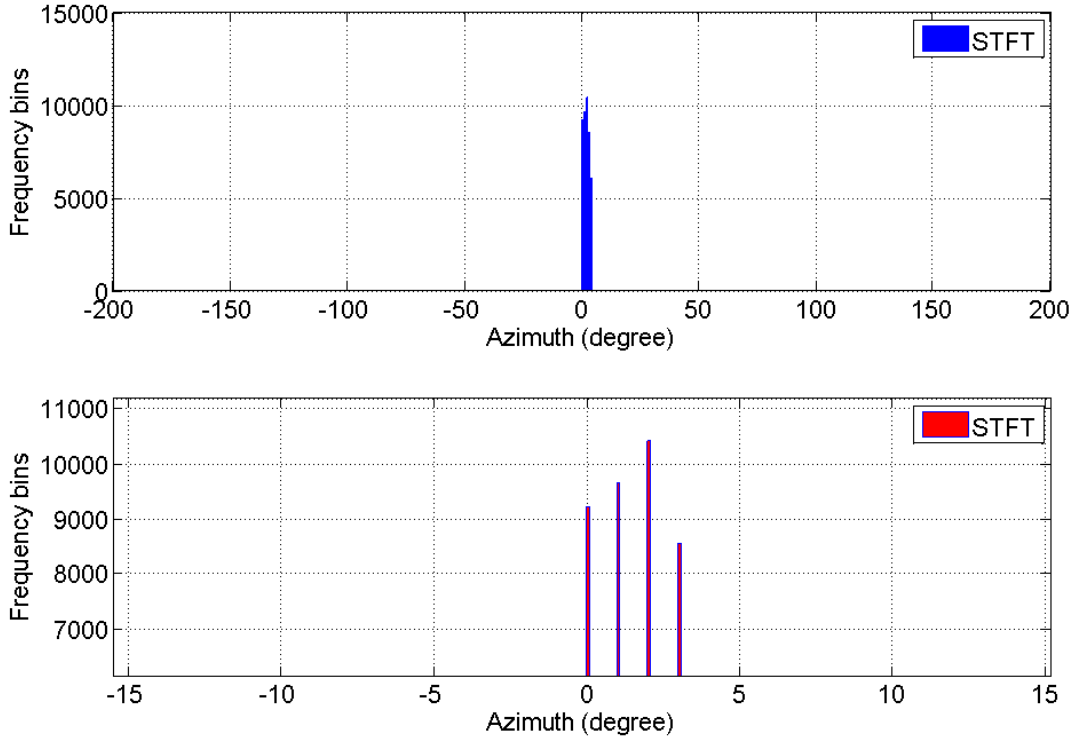


Figure 5.15: Resolution of the energetic analysis method with noise absence.

Figure 5.16 shows the simulation results of the method when a random noise was added. The noise was generated in Matlab. The SNR in this simulation was -16 dB. The same simulation scenario was applied here i.e., the speakers are separated from each other by $1\,^\circ$ only. However, five speakers were assumed to be around the microphone at the positions $(0\,^\circ, 1\,^\circ, 2\,^\circ, 3\,^\circ, 4\,^\circ)$. It can be seen on the lowest zoomed part of the simulation that the resolution of the method was not affected by adding the noise. New smaller peaks appeared, which can cause an error in this case. The new peaks are caused by the added noise, where they denote the possible positions of the noise sources.

In order to check the results in the real environment, sound files were recorded in the laboratory. In the recording part, two male speakers spoked simultaneously in front of A-format microphone. The distance between the speakers was changed in each recording. The angles between the two speakers were $(30\,^\circ, 25\,^\circ, 20\,^\circ, 15\,^\circ, 10\,^\circ)$. The method was able to distinguish between the speakers when the angle between them was $15\,^\circ$ and bigger. However, the two sound sources were seen as a one sound source when the angle was $10\,^\circ$ or less. Assuming that the speakers are standing 3 m away from the microphones, and the angle between them is $15\,^\circ$, this means that their heads are 0.7854 m far away from each other. This distance decreases of course when they are nearest to the microphone.

Figure 5.17 shows the results when two speakers spoke near to each other. The upper

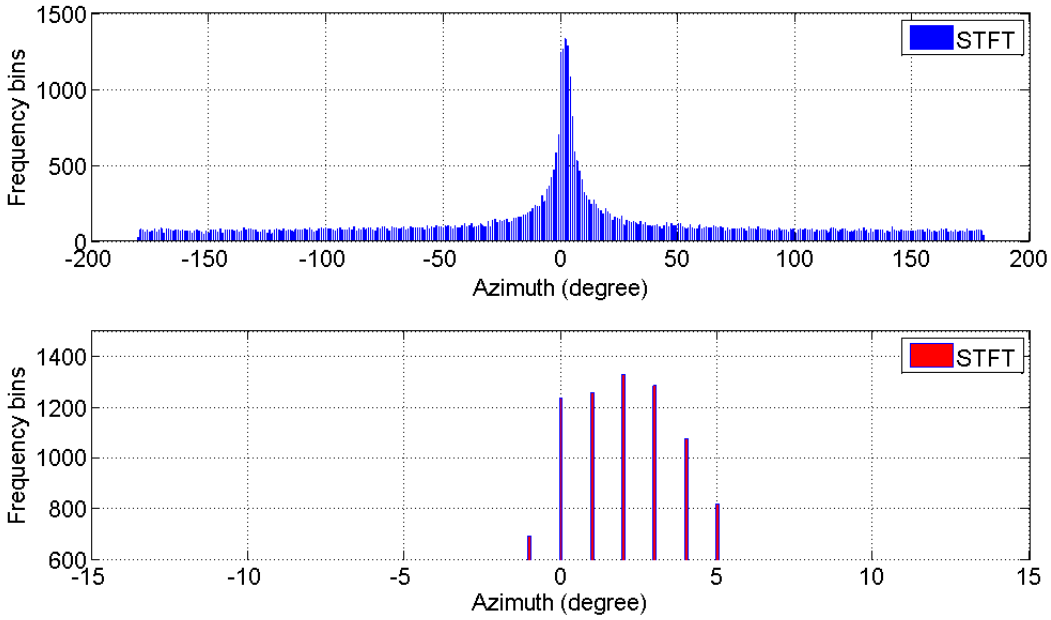


Figure 5.16: Resolution of the energetic analysis method with an additive noise.

part shows the results in the whole area around the microphone, and the lowest part shows the area we are interested in. The first speaker stood at the position 0 ° and the second one stood at the position 15 °. As can be seen, the method was able to estimate the direction of the speakers when the angle between them was about 15 °.

5.2.7 The Impact of the Used Window on the Accuracy of the Estimation Method

Different window types can be used in the time-frequency transform. Several windows have been introduced in chapter 2. As it is well known, the shape and the properties of the window function influence the properties of the time-frequency transform, though they influence the accuracy of the method. Even more, the width of the window in time domain affects the time-frequency resolution as already introduced in Heisenberg's principle in the section 2.3.1.

Figure 5.18 shows the simulation results of the energetic analysis method depending on the used window function. The simulation was carried out using STFT as a time-frequency transform with different window types; namely, the Hamming window, the Hanning window, the rectangular window, the Blackman window and the Bartlett window. The width of each window is 256 samples and the overlapping is 50%. These numbers were chosen in order to achieve a compromise between the resolution in time domain and in the frequency domain.

In this simulation, the listeners were assumed to be around the microphone at the po-

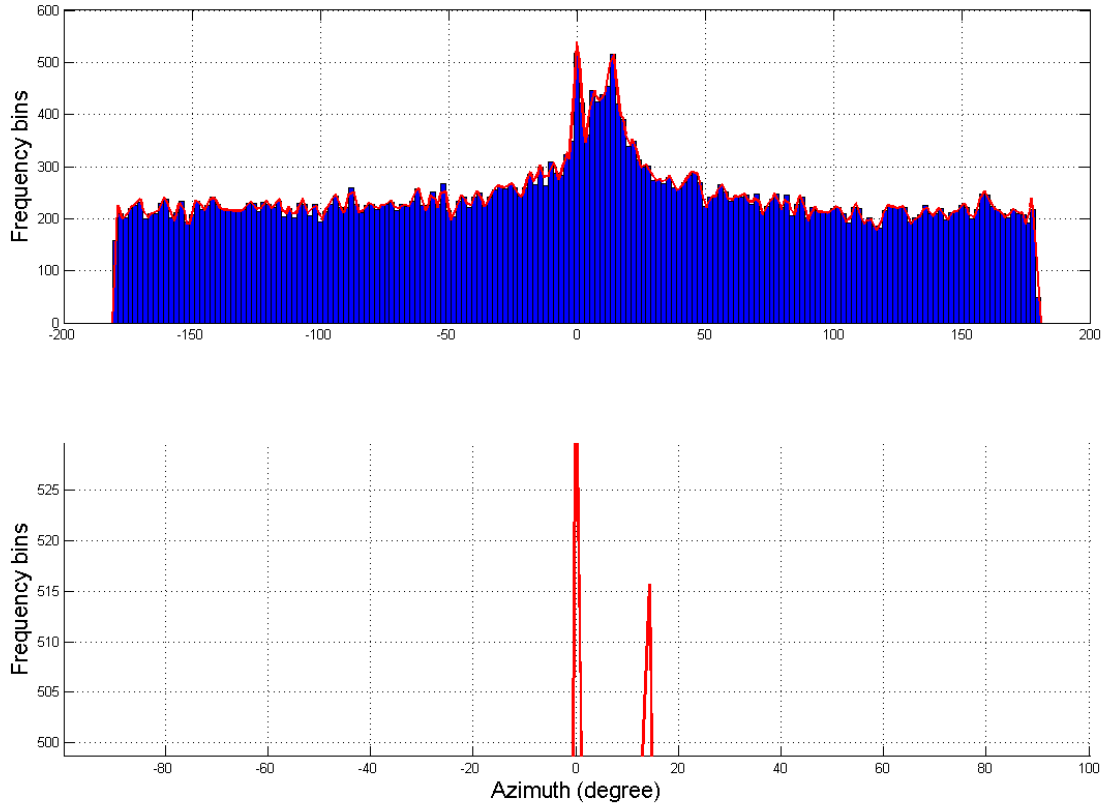


Figure 5.17: Resolution of the energetic analysis method in the real environment.

sitions $(-150\,°, -70\,°, 0\,°, 100\,°, 160\,°)$, these positions were chosen in an arbitrary way. As can bee seen in Figure 5.18, the method was able to estimate the direction of arrival of the simulated speakers, where the peaks indicate the direction of the sound sources. However, some windows achieved sharper peaks than others. The peaks achieved by the Hanning and the Hamming windows were the sharpest, whereas the Bartlett window achieved the widest. That can be explained by recalling the properties of the windows in frequency domain presented in the previous chapter, see section 2.2.

5.2.8 Tracking the Targets Using Energetic Analysis Method

The energetic analysis method provides the possibility of tracking one speaker by dividing the audio file into multiple shorter files, and estimation the direction of arrival of the speaker inside each shorter file. In our simulation, the duration of each sub-file was about 1 second. However, it should be noted that the length of the file cannot be shorter than specific variable. The minimum length of a single file, which still provides the possibility of estimation the direction of the speaker, is subject to several conditions. Thereby, we

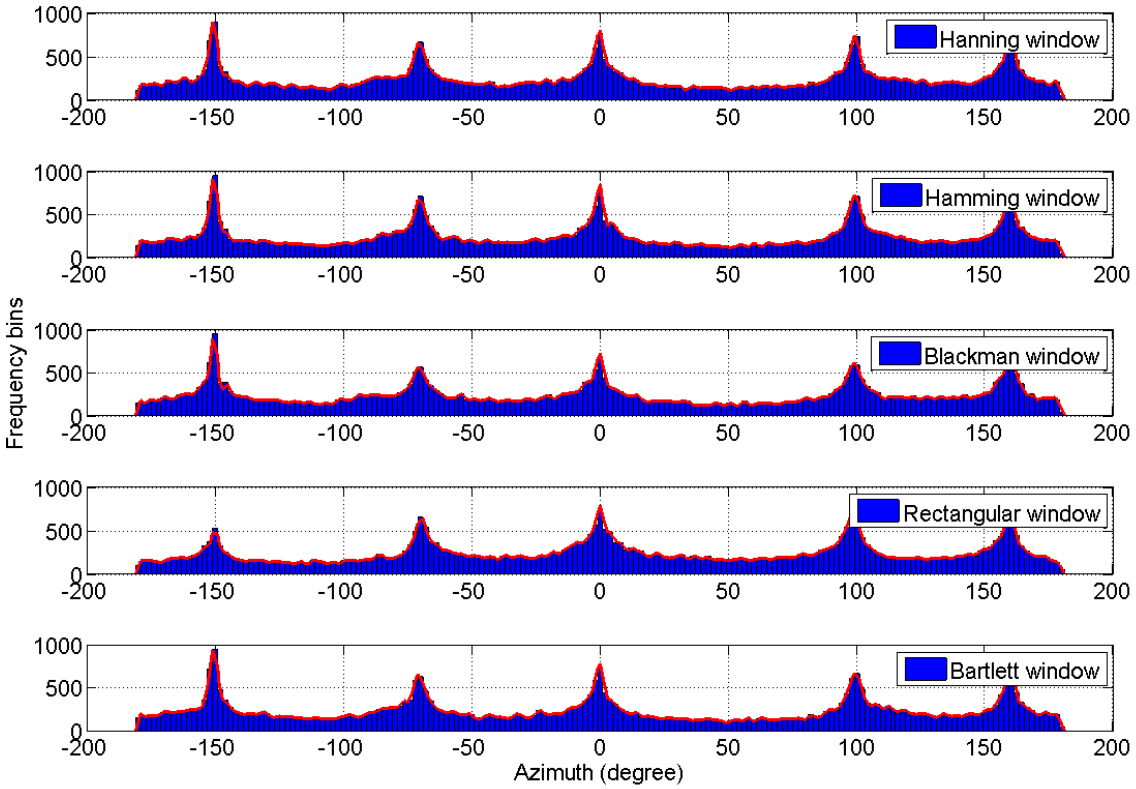


Figure 5.18: The performance of the energetic analysis method using different window functions.

should take into account the following conditions:

1. The loudness of the speaker.
2. The reverberation in the room.
3. SNR.

Figure 5.19 shows a simulation of the tracking process. The movement of one speaker was simulated using B-format signals and an additive random noise was added to the sound file with SNR about -5 dB. The mobile speaker was tracked by dividing the audio file into smaller files, where the duration of each audio file was about 1 second. As can be seen, the peaks denote the direction of the speaker. The speaker was assumed to move in a uniform circular motion. Since we simulated only the angle of the speaker regardless the distance from the microphone, the speed of the speaker differ depending on the distance of the microphone. However, the direction of the speakers changes steadily around the

microphone, where we assumed that the angular speed is 15 ° per second. The direction of the speaker differs in each segment of the figure denoting the movement of the speaker.

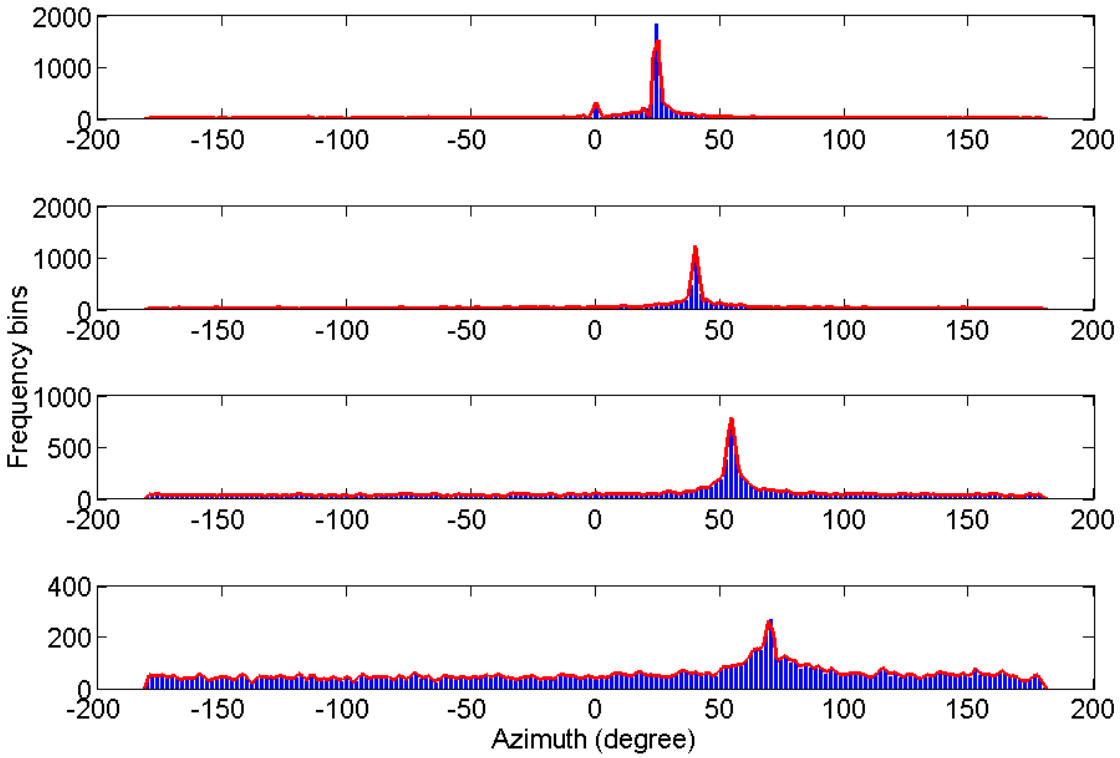


Figure 5.19: Simulation results of tracking the movement of one speaker.

5.3 Sub-Conclusion

The energetic analysis method is a suitable method for estimation the direction of arrival of multiple speakers in both horizontal and vertical plane. The method depends on analyzing and calculating the intensity of the sound. The simulation and experimental results approved the accuracy of this method. Even more, the method can be used for tracking a moving speaker.

The experiments were carried out when three speakers spoke simultaneously, the method was able to estimate the directions of the three speakers in the horizontal plane and in the vertical plane. The median of the absolute angle error was about 3 °.

The performance of this method was studied using several time-frequency transforms and using several windows function type. The results approve that better resolution in time-frequency domain ensures better performance of the method.

Chapter 6

Acoustic Zooming

In this chapter, we present a compatible system for both acoustic zooming and sound direction estimation [105]. The proposed system relies on the energetic analysis method for estimation the direction of arrival of multiple speakers, and on changing the parameters of DirAC for zooming the sound coming from the wanted direction.

The listening tests have been carried out to evaluate the reliability of this system, the tests were designed to compare the quality of the sound when several time-frequency transforms were used, and to evaluate the ability of this system to zoom the sound of one speaker and attenuate the other. The experimental results showed the ability of this system to zoom the sound. Objective measurements have been carried out as well. The results of the measurements showed the quality of the proposed system.

6.1 Description of the Proposed System

The proposed system uses the energetic analysis method to estimate the direction of multiple speakers giving the information about the speaker locations, which is needed in the zooming system [105]. Then, the system zooms the sound of one speaker by modifying the parameters of the directional audio coding. Further more, after estimation the direction of the speakers, the system can be modified to attenuate the sound of one or more speakers and keep the sound of the others at the same level as in the original sound.

The paper presented in [106] investigated an acoustic zooming method by modifying DirAC parameters to zoom the sound. However, our system provides the possibility of using two time-frequency transforms and the possibility of using sound source direction estimation method to localize the speakers.

The proposed system consists of four blocks; namely, sound source localization unit, DirAC analysis unit, zooming and synthesis unit and rendering unit, see Figure 6.1.

The input signals for this system are B-format signals presented earlier in this work, and the output signals are the modified DirAC signals.

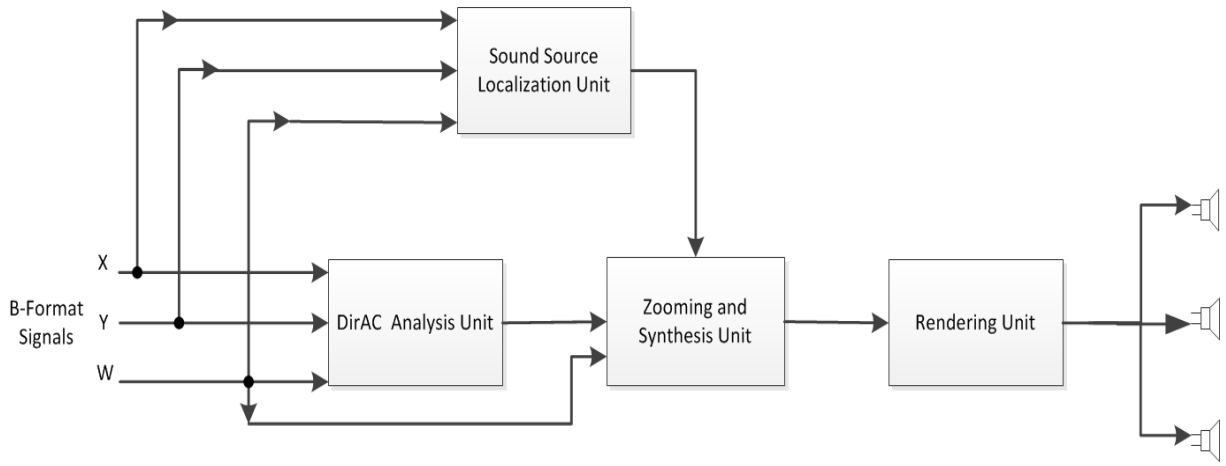


Figure 6.1: The proposed system in the two-dimensional plane.

6.1.1 The Modified Energetic Analysis Method

The energetic analysis method has been introduced in chapter 5. However, extra steps have been added to the original energetic analysis method to exploit the proprieties of the human voice and also the proprieties of the acoustic in the closed room, see Figure 6.2. These three steps are: low pass filter, DirAC analysis and estimation of the non-diffuse part.

Filters

The input signals of this unit are B-format signals. The idea of using a low pass filter comes from the fact that we want to estimate the direction of a human speech source. As it is well known, the speech spectrum can be divided into two parts, the first part is flat and it contains the frequencies up to 500 Hz, whereas the second part has a slope of -10 dB/octave, and it is applied to the frequencies higher than 500 Hz [107], [108]. Applying a low pass filter to the input signals suppresses the additional interference caused by higher frequency, which belongs to the noise signals. Therefore, we applied a low pass FIR filter with cut-off frequency equals to 3500 Hz.

We also applied a high pass filter with cut-off frequency equal to 100 Hz in order to minimize the effect of the standing waves in the laboratory. It was seen that adding these filters improves the accuracy of the energetic analysis method.

Dirac Analysis

The purpose of DirAC analysis in this stage is to estimate the diffuseness parameter, which can be used to divide the sound signal into diffuse and non-diffuse streams. The input signal of this step is the resulted filtered signals from the previous step. The signals are then divided in time and frequency, and the DirAC parameters are calculated as in the

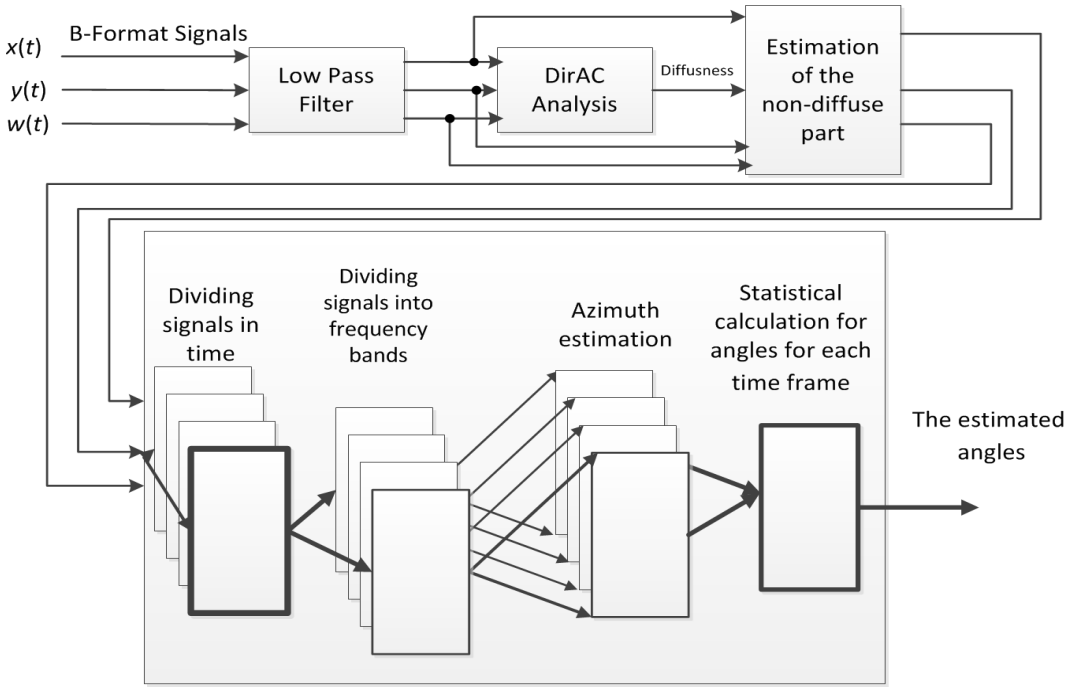


Figure 6.2: The localization unit in the two-dimensional plane.

original DirAC method. The diffuseness parameter is then estimated to be applied in the next step.

Estimation of the Non-Diffuse Part

The sound signals are first separated into diffuse and non-diffuse stream using the diffuseness parameter. The separation can be done by multiplying the signal in the frequency domain by the parameter $\sqrt{\Psi}$ and $\sqrt{1-\Psi}$ as was mentioned in section 1.2.2.4. Then the non-diffuse part can be used to improve the accuracy of this unit by eliminating the diffusing sound, which is resulted from the reverberant sound. The non-diffuse part only is transmitted to the time domain using inverse STFT or Gabor transform depending on the transform used in the transformation into frequency domain.

After processing the above mentioned steps, the original energetic analysis method is applied normally to the resulted signals. The results are in this case more accurate because of suppressing the interference caused by the diffuse sound and reverberant signals.

The absolute angle error of this method with and without the mentioned steps is illustrated in Figure 6.3 using boxplot. As can be clearly seen, the absolute angle error was reduced when these steps were applied.

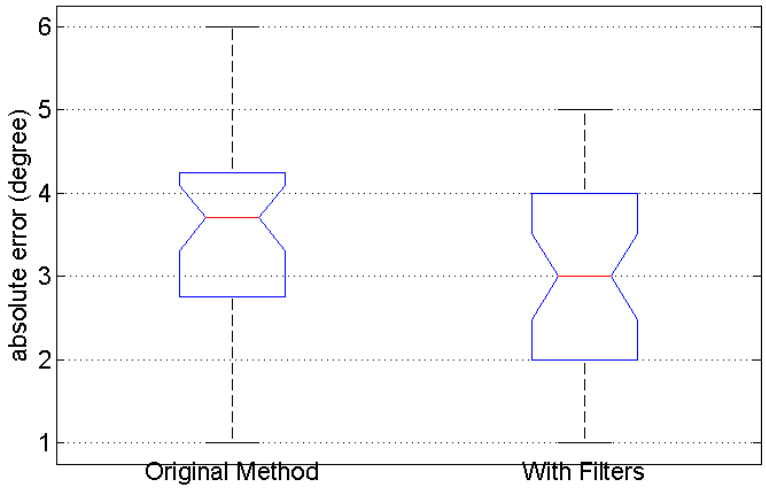


Figure 6.3: The absolute angle error of the original and the modified energetic analysis method.

6.1.2 Zooming and Synthesis Unit

The input signals for this unit are the omni-directional B-format signal ($w(t)$), the parameters estimated from the DirAC analysis unit and the information about the direction of the speakers obtained from the sound source localization unit [105].

The sound signal is first transmitted into frequency domain, and then it is divided into diffuse and non-diffuse stream depending on the diffuseness we estimated from the DirAC analysis unit. A gain factor is then applied to the non-diffuse part, and it is calculated as

$$g(m,n) = \begin{cases} g_{max} & if\ \text{DOA}(m,n) \in [\theta + \Upsilon, \theta - \Upsilon] \\ \\ g_{min} & if\ \text{DOA}(m,n) \notin [\theta + \Upsilon, \theta - \Upsilon] \end{cases} \tag{6.1}$$

where $g(m,n)$ is the gain applied to the frequency bin number m in the time sample number n, g_{max} is the maximum gain applied to the sound we want to zoom, g_{min} is the attenuation factor, $\text{DOA}(m,n)$ is the direction of arrival estimated from DirAC analysis, θ is the direction of the speaker whose sound we want to emphasis, and it is estimated from the sound source localization unit and Υ is the half of the angle in which we zoom the sound and it differs in each scenario. Υ was chosen to be 5 degrees in our experiments. It was chosen depending on the length of the arc (space) that the normal size person can occupy when he is 2 m far from the microphones.

The zooming factor impacts the quality of the sound. When a large zooming factor is used, an audible distortion occurs to the sound file, which affects the quality of the reproduced sound. Using a smoothing method improves the quality of the sound, and

minimizes the distortion of the sound.

6.1.3 Rendering Unit

When the sound is transmitted to the time domain, it can be rendered to a set of loudspeakers, or to headphones [109]. However, a prior knowledge about the distribution of the loudspeakers should be taken into account when the rendering method is applied. In our system, we chose VBAP as a suitable method for rendering the sound since it has better localization accuracy over first-order Ambisonic [91].

6.2 Description of the Experiments

The experiments were designed to evaluate the ability of zooming the sound, the resolution of the zooming technique and the precision of the mentioned system. They can be divided into three stages; namely, recording the sound, processing the sound and listening stage [105].

6.2.1 Recording the Sound

The recording was carried out in the acoustic laboratory, see appendix A. A SPS200 Soundfield microphone was used to record the sound of four speakers (three men and one woman) [104]. The listeners spoke simultaneously. A short English sentence was chosen as a test sentence. The duration of the speech was about 5 seconds. All speakers said the same sentence simultaneously, which ensures the most difficult situation for the system. The microphone was placed at the center of the laboratory, and the speakers stood at different positions around it at six different combinations. The sound signals were recorded as A-format signals, and then they were transmitted into B-format signals.

Another recording was carried out to measure the resolution of the system. In this scenario, two speakers said simultaneously the same English sentences at different positions. The speakers came closer to each other in each new recording. The purpose of this step is to measure the smallest distance between the speakers at which the system is still able to zoom the sound of one speaker.

6.2.2 Processing the Sound

The mentioned system was applied to the recorded sound files in the previous paragraph. It was built using Matlab. Two time-frequency transforms were used; namely, short-time Fourier transform (STFT) and Gabor transform. The direction of the speakers was first estimated and then the zooming method was applied to each speaker of the four speakers separately. The same zooming factors were applied when both Gabor and STFT were used.

Recalling that uncertainty principle claims that it is not possible to achieve optimal localization in time and frequency simultaneously. In order to achieve the best resolution in both time and frequency domains simultaneously, a compromise between time localization and frequency localization should be done. Therefore, we chose both Gabor and STFT as time-frequency transformations to study their effects on the quality of the resulted sound.

When STFT was used, a square-root Hanning window was applied, the length of this window was chosen to be 512 samples, the overlaps were chosen to be 256 points, the number of sampling points to calculate the discrete Fourier transform was 256 points, and the sampling frequency was 44100 Hz. A square-root Gaussian window was used when Gabor transform was applied. However, a similar window length and sampling frequency were used in both cases. The parameters were chosen depending on preliminary experiments, where the sound, processed using this parameters, was with the highest subjective quality.

6.2.3 Listening Test

In order to evaluate the zooming system, a listening test was carried out [105]. The listening test compared the original sound rendered using DirAC and the zoomed sound using both STFT and Gabor transform. The test was performed in the acoustic laboratory described in appendix A as follows: six loudspeakers were located in the vertices of a regular hexagon with distance of vertices from the sweet spot of 2.5 meters. For this test, ten listeners were used. The listeners have been chosen without any hearing impairment, at the age from 25 to 35 years. Five listeners have a good experience in the procedure of listening tests. For others, the procedures were explained carefully. The listeners included four women and six men. Each listener was seated at the position of the sweet spot of the loudspeaker setup. The listeners were asked to give an evaluation of the quality of the sound and of the loudness of the loudest speaker compared to the others. They were told to write their evaluation on a sheet of paper, which had the questions and a scale for each question. Five scales were available to describe the quality of the sound based on mean opinion score (MOS) [110]. The existed options according to MOS are presented in Table 6.1.

Quality of the speech	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 6.1: Listening-quality scale (MOS).

Another five scales were available to describe the loudness ratio of the speakers to each other. The available options that describe the ratio of the loudness of the speakers in this case are shown in Table 6.2.

The loudness ratio of the loudest speaker	Score
I cannot hear the others	5
Very higher	4
Higher	3
Slightly higher	2
The speakers have the same loudness	1

Table 6.2: Loudness ratio.

The listening tests were also devoted to measure the precision of the system. The listeners were asked to localize the sound sources. A mobile loudspeaker was used as a reference sound. The same sentence was rendered via the mobile loudspeaker and the original loudspeaker array alternately. The mobile loudspeaker was moved around the sweet spot in the same distance as the loudspeakers of the array till the listener said that the sound coming from it and the sound rendered via the original loudspeaker array have the same direction. This step was applied to each one of the four speakers in each audio file and only to the zoomed speaker in the zoomed files.

In order to study the relation between the value of the zooming factor and the degradation of the quality of the sound, a listening test was designed, where the same sound file with the same zooming area was processed with different zooming factors. In this listening test we used the degradation mean opinion score (DMOS) which was described in Annex D of ITU-T Recommendation P.800 [110]. The scales for DMOS are presented in Table 6.3

Degradation of the sound quality	Score
inaudible	5
audible but not annoying	4
slightly annoying	3
annoying	2
very annoying	1

Table 6.3: Degradation category scale (DMOS).

It should be noted that the duration of each test did not exceed 30 minutes, during which each listener evaluated three sound files.

6.3 Experimental Results

Depending on our listening test's results, the best ratio between g_{max} and g_{min} is between 13 and 15 because of the ability of zooming the sound and keeping an acceptable quality of

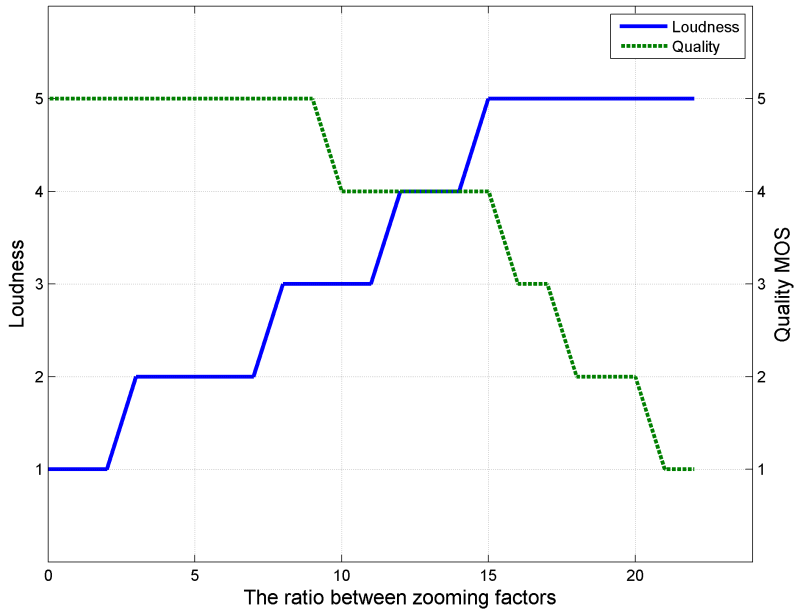


Figure 6.4: The relation between the zooming ratio and the quality of the sound.

it. Therefore, we chose the ratio 15 as a suitable value to be applied in the next listening tests. To estimate this ratio, the audio files were processed using different zooming factors as it was explained in the previous paragraph. It was seen that when small ratio between g_{max} and g_{min} is used, the zooming was not audible enough, whereas bigger ratio between g_{max} and g_{min} caused some distortion to the sound. Figure 6.4 shows the results regarding Table 6.2 and Table 6.3.

The resolution of the system was measured in a subjective way. According to our measurements, the smallest angle between the speakers at which the system was still able to zoom the sound of one speaker and attenuate the second one is 15°. When the angle between the speakers was bigger than 15°, the system worked correctly. However, when the speakers were closer to each other, the system zoomed the sound of both speakers.

A part of our experiments attended to measure the localization blur of this system, and the influence of the zooming system on this blur. In our experiments, most of the listeners explained the sound localization as (*easier*) when the zooming was applied. However, it was noticed that the listeners attended to match the sound source with the visible loudspeakers when the sound source was near them. In the original sound files i.e. without zooming, the listeners were asked to localize the four speakers, whereas they were asked to localize only the zoomed sound when the zooming was applied. The results showed that the median blur for the system was about 18°, and it was decreased a little bit when the zooming sound was applied. This little improvement in precision is mostly because of attenuating of the other sounds, which can be seen as a distraction when the

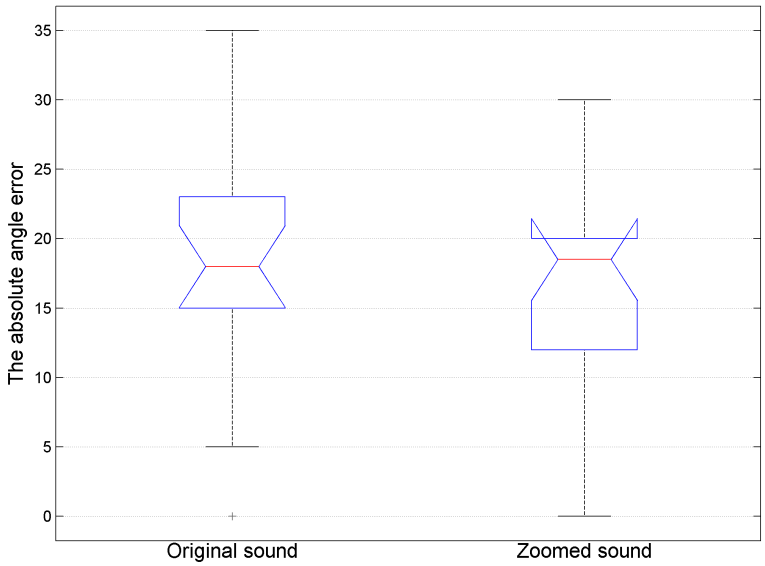


Figure 6.5: The localization blur for both original and zoomed sound.

listener focuses his attention on one speaker, see Figure 6.5.

The results of the loudness of the sound are presented in both Figure 6.6 and Figure 6.7. The results were computed for each audio file as the average score of the evaluation given by the listeners who listened to the sound file. The results are illustrated in the graphs regarding the scales presented in Table 6.2. As was seen in the previous paragraph, the zooming was applied to each speaker of the four speakers in our recordings. However, the loudness of the sound of each person differs from the others. Though, the intensity of the sound is different as well. It was seen in our experiments that zooming the sound of the loudest person achieved the best quality. When the sound of one speaker was almost inaudible in the original recording, the zoomed sound of this person achieved the worst results. The results of the sound quality are presented in both Figure 6.8 and Figure 6.9 regarding MOS score presented in Table 6.1.

Depending on our results, the experienced listeners who had taken a part in listening tests before, felt the difference between the quality of the sound files when Gabor or STFT was applied to the zooming system more than the inexperienced listeners. Figure 6.10 and Figure 6.11 compare the results when Gabor and STFT are used using boxplot. As can be clearly seen, Gabor achieved the best results. The perception of the loudness of the sound was better and it kept better sound quality.

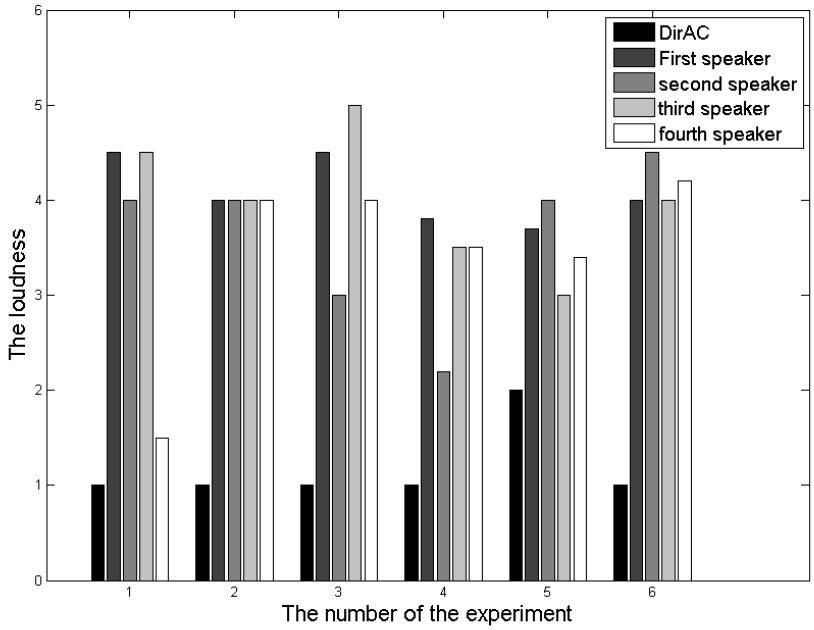


Figure 6.6: The loudness ratio between the sound of the speakers when STFT was used.

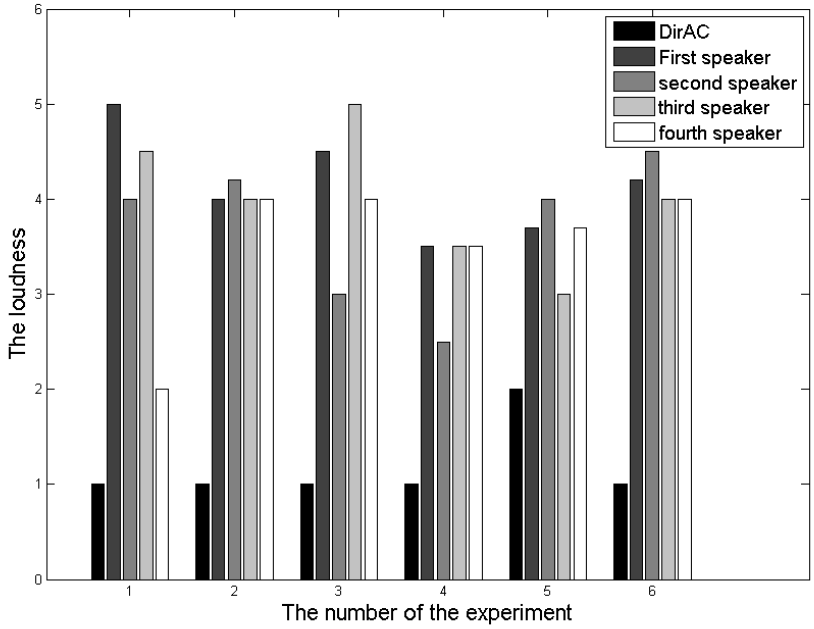


Figure 6.7: The loudness ratio between the sound of the speakers when Gabor was used.

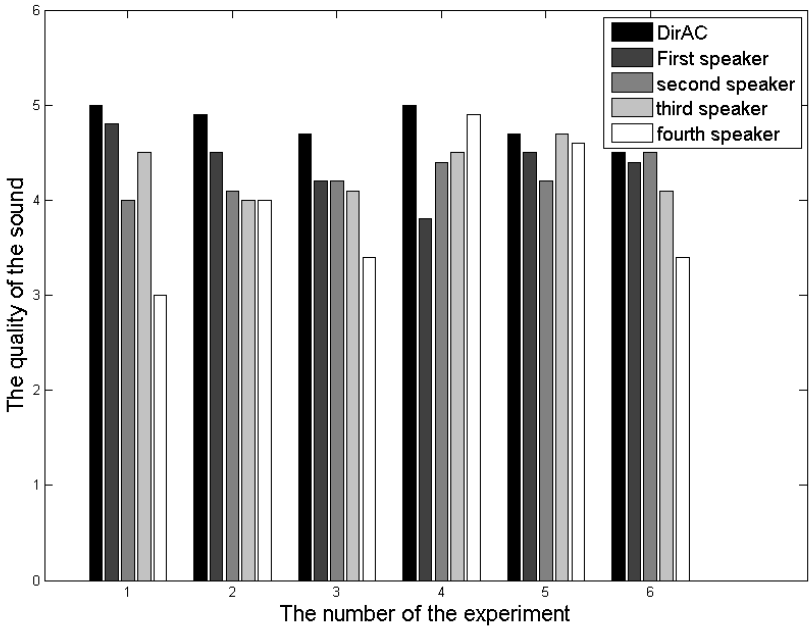


Figure 6.8: The quality of the sound files according to MOS scale when STFT was used.

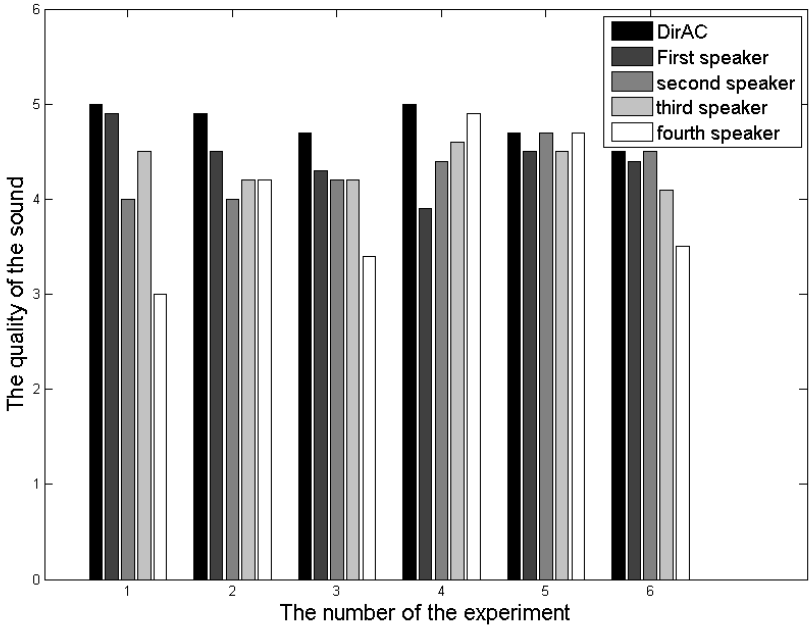


Figure 6.9: The quality of the sound files according to MOS scales when Gabor was used.

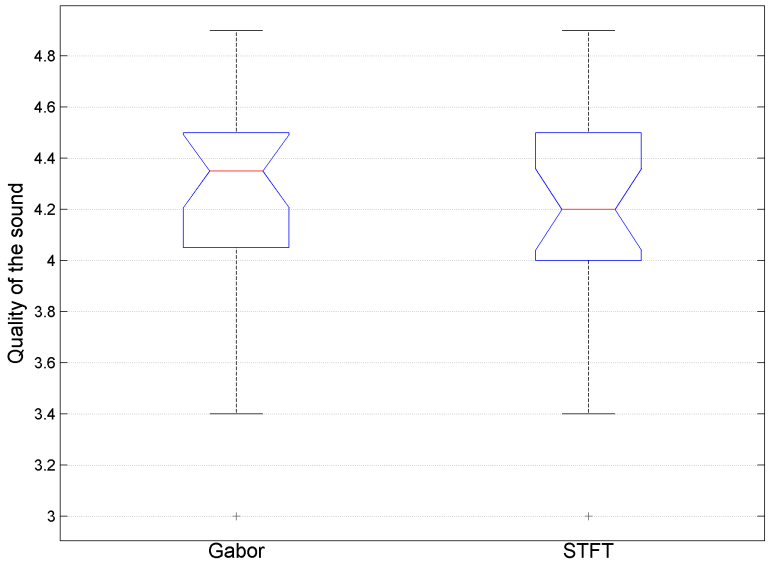


Figure 6.10: The quality of the sound when Gabor and STFT are used.

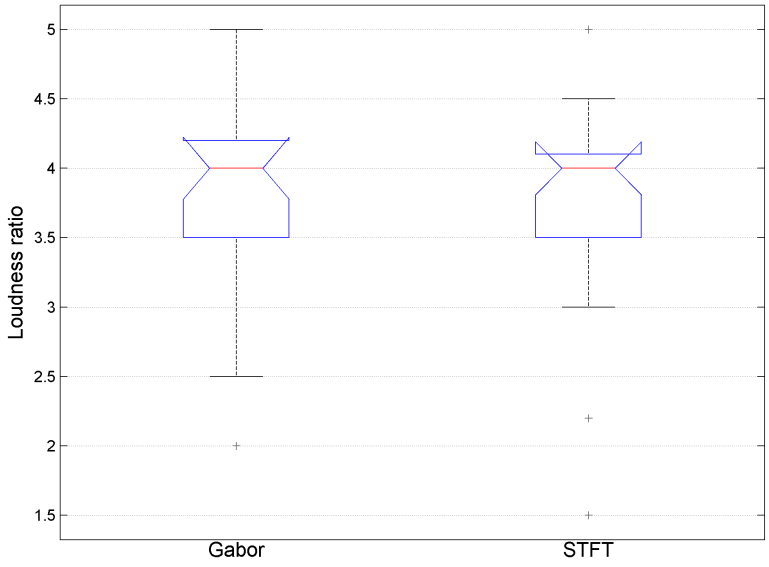


Figure 6.11: The loudness ratio between the sound of the speakers when Gabor and STFT are used.

6.4 Objective Measurement

For the objective assessment of quality of extracting the signal of the zoomed sound source from a given mixture we used the perceptual evaluation methods for audio source separation (PEASS) [111] which is designed specifically for these purposes. The algorithm [111] is based on decomposing the estimation error into three components (target distortion, interference and artifacts components), assessing the salience of each component via PEMO-Q (Perception Model for Quality assessment) [112] quality metric and combining these saliences via trained nonlinear mappings. The algorithm outputs are overall perceptual score (OPS), target-related perceptual score (TPS), interference-related perceptual score (IPS) and artifacts-related perceptual score (APS).

We used the sound of four male speakers as the source sounds of the zooming system. The sound of the speakers was recorded in the anechoic room (reverberation time 50 ± 10 ms in octave bands from 250 Hz to 8 kHz). The sampling frequency of the recordings was 44.1 kHz and the recordings were synchronized in time. In order to align the loudness of the sound sources, their level was adjusted to RMS value of -20 dBFS with maximum peak values of -3 dBFS using the Steinberg Wavelab loudness normalizer. These recordings were rendered using four loudspeakers in the same laboratory where the subjective tests were performed. The loudspeakers were placed in the same distance from the sweet spot and in the same angles as the speakers when the recordings for the subjective tests were carried out.

At first, an omnidirectional microphone was placed in the sweet spot of the loudspeaker array and the single speakers were recorded. The recordings were carried out synchronously with the playback of given speaker. The used microphone with the recording system conforms the IEC 61672 class 1. The recordings of the individual speakers were used as the reference signals for the PEASS algorithm.

In the second step, the sound of the four speakers, which was rendered using four loudspeakers simultaneously, was recorded using the SoundField microphone, where this microphone was placed at the sweet spot of the loudspeaker as well. Recording using the omnidirectional microphone was performed as well to compare the results. The sound field recorded using SoundField microphone was then processed using the DirAC without zoom and the sound of one selected speaker was zoomed in using our system with STFT and Gabor transform. The processed sound files were rendered at the same conditions used in the subjective listening tests. The same omnidirectional microphone was placed in the sweet spot of the loudspeaker array and its signal was recorded synchronously with the playback signal of the loudspeaker array. The recorded signals in the three cases (DirAC, zoomed sound using STFT and Gabor) were then used as a test signal for the PEASS algorithm.

The results of the objective assessment of the speech quality are shown in Table 6.4. As it can be seen from the PEASS results, the overall perceptual score of the zoomed speaker is definitely better than the score of all four speakers played back simultaneously (OPS=8) and also better than the score of the sound field of all four speakers rendered using DirAC without zooming (OPS=19). The results are almost the same when Gabor and STFT are

Tested signal	Original sound field of 4 speakers	Sound field rendered using DirAC without zoom	Sound field rendered with zoom	
			Gabor	STFT
OPS	8 %	19 %	38 %	38 %
TPS	81 %	38 %	44 %	44 %
IPS	1 %	15 %	55 %	52 %
APS	87 %	54 %	44 %	45 %

Table 6.4: The average results of the speech quality assessment using the PEASS algorithm.

used for the zooming. A more detailed analysis shows that a greater suppression of the other speakers (IPS) occurs using the Gabor transformation than the STFT.

Absolute values of the assessment for the zooming algorithms are relatively low, but the quality improvements compared to the situation without using the zoom is clear. For the correct interpretation of the results of the PEASS algorithm it should be noted that the OPS is only 53 when we compare the recording of one speaker captured using the omnidirectional microphone in the room where the test was performed, with the recording of the speaker in the anechoic chamber, even those two recordings differ from each other only in the natural reverberation of the room. This demonstrates high sensitivity of the PEASS algorithm to any signal change. So it is necessary to take the output values of the objective assessment algorithms as the approximate values. In this case, the results of the subjective tests are primary.

6.5 Subjective Intelligibility Test

The evaluation of the proposed zooming system included an intelligibility test. For the statistical intelligibility testing, we used a list of 50 monosyllabic phonetically balanced (PB) words from Harvard PAL PB-50 test. Phonetic balance means that the relative frequencies of the phonemes on the test list are as close as possible to the distribution of speech sounds used in English, which was in turn based on analysis of 100,000 words in newsprint [113]. Sound field of four speakers was used for the subjective intelligibility assessment. Three of them read English sentences, which were not related to the list of words, and one of the speakers was reading the PB words in random order, each spoken in the same carrier sentence. Same procedures, as in the objective assessments, were made for this test. The recorded sounds were then processed using the zoom system. The listeners sat at the sweet spot of the loudspeaker array and they listened to the sound in four cases i.e. mono-sound, rendered sound using DirAC, zoomed sound using STFT and zoomed sound using Gabor transform. The listeners wrote down the recognized words.

The Chance-Adjusted Correct Response rate (CACR) for one listener is calculated

using the following formula [114] to compensate for the chance level:

$$S = \frac{N_C - N_I}{N} 100 \tag{6.2}$$

where S is the response rate adjusted for chance ("true" correct response rate), N is the total number of PB words used in test (50), N_C is number of correctly recognized PB words, and N_I is the number of incorrectly recognized PB words. A completely random response will result in half of the responses to be correct. With this equation, random response will give average response rate of 0 %, not 50 % [114].

The results of the subjective intelligibility test are shown in Table 6.5. The resulting response rate for the intelligibility was computed as the average value for 10 listeners. It is clear from the results, that the Acoustic Zoom significantly improves the intelligibility. As can be seen, the same trend as in the objective results is noticed. However, the results can be compared only qualitatively.

The Tested Signal	Intelligibility Score
Acoustic zoom using Gabor	92.4
Acoustic zoom using STFT	89.6
DirAC	24
All	16.4

Table 6.5: Intelligibility Score.

6.6 Sub-Conclusion

A new system for estimation the direction of the speakers and zooming the sound of one of them was introduced. This system depends on the energetic analysis method for estimation the direction of the speakers, and on modifying the DirAC parameters for zooming the sound. Two time-frequency transforms are used namely, STFT and Gabor transform. Several listening tests have been carried out to evaluate the effect of the zooming ratio on the quality of the sound, the precision (localization blur), the quality of the sound, and the performance of the acoustic zooming system. The listening tests were mostly designed depending on ITU recommendations. The subjective experiments showed that Gabor transform achieved better results than STFT. It also showed that the resolution of this system is about 15°, and the precision (localization blur) is almost 18°.

Objective tests were done as well. The objective tests were in conformity with the subjective tests. PEASS algorithm evaluated the attenuation of other speakers (IPS) to be over 50%, whereas the ratio of the loudness of the zoomed speaker is about 3.5 till 4 from 5 point on MOS scale according to the results of the subjective tests. The comparison of the results of quality assessment is more complicated because the listeners were not told if they have to evaluate the quality degradation of the zoomed sound (equivalent to the

TPS) or the quality degradation of the sound due to artifacts (equivalent to the APS). A closer analysis of each evaluation of PEASS algorithm shows that the zoomed speaker is more separated from other speakers when Gabor transform is used than when STFT is used, but other artifacts occur.

Chapter 7

Conclusions

This dissertation aimed at designing an acoustic zooming system, which enables zooming the sound of one speaker among the other speakers. Throughout this work, two major disciplines have been investigated; namely, sound source localization and surround sound rendering. The proposed system uses the results of this work in the both mentioned fields, creating the possibility of further investigation in this area in order to work in real time.

The thesis presented an overview and state of the art of the current sound source localization methods, together with an introduction and state of the art of the sound rendering techniques in Chapter 1.

Chapter 2 investigated different time-frequency transforms, recalled their formula and explained the conception of optimal localization in the time-frequency domain or so called Heisenberg's principle, whose effect on sound direction estimation method was studied in the next chapters.

Chapter 4 presented the simulation results of sound localization methods; namely, CC, PHAT and ML methods, and experimented them in the laboratory. The experimental results showed the ability of these methods to estimate the direction of one speaker in closed rooms. It was seen that PHAT method has the advantage over CC and ML methods against the noise in the simulation environment, and it also achieved the best experimental results.

The accuracy of some sound rendering techniques was also experimented in order to choose the best technique which can be used in the acoustic zooming system. The experimental results showed that VBAP has better accuracy compared to the Ambisonic first-order decoders.

Chapter 5 presented a new method for multiple sound source direction estimation, showed the simulation and the experimental results of this method, and studied the effect of the used transform on the accuracy of this method. The simulation results showed the ability of this method to estimate the direction of multiple speakers in both two-dimensional and three-dimensional plane, which was approved by the experimental results as well. The effect of the used time-frequency transform on the accuracy of this method has been studied as well, showing that better accuracy in time-frequency domain ensures better accuracy of this method. Even more, simulation has been performed to investigate

the ability of this method to track a moving target.

A combined system of sound source localization and acoustic zooming was presented in Chapter 6. The system combines the effort of the previous chapters providing the possibility of estimation the direction of multiple speakers and zooming the sound of one of them. The listening tests have been performed as well to evaluate this system. The evaluation approved the ability of this system and also showed the effect of time-frequency transform on the quality of the sound.

7.1 Contributions of the Thesis

The main contributions of this work can be summarized as follows:

- The thesis started by investigating the existed sound source localization methods. Some experiments have been implemented in order to compare the accuracy of these methods.
- A new sound source direction estimation method, which was inspired by DirAC, has been introduced in this work. The simulation and the experimental results of this method have been shown in Chapter 5.
- A system for acoustic zooming has been tested using several time-frequency transforms. This system depends on the sound direction estimation method and on modifying the parameters of DirAC.
- The listening tests and objective measurements have been performed as well. The results of these listening tests and measurements approved the accuracy of this acoustic zooming system.

The author of the thesis suggests exploring the possibility of working the proposed system in the real time. Then, the system can be used in video-conferencing system.

Bibliography

[1] V. Algazi and R. Duda, "Headphone-based spatial sound," *Signal Processing Magazine, IEEE*, vol. 28, no. 1, pp. 33–42, 2011.

[2] F. L. Wightman and D. J. Kistler, "Monaural sound localization revisited," *The Journal of the Acoustical Society of America*, vol. 101, p. 1050, 1997.

[3] J. W. Schnupp and C. E. Carr, "On hearing with more than one ear: lessons from evolution," *Nature neuroscience*, vol. 12, no. 6, pp. 692–697, 2009.

[4] R. S. Woodworth and H. Schlosberg, *Experimental psychology*. Holt, 1954.

[5] G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane," *The Journal of the Acoustical Society of America*, vol. 62, p. 157, 1977.

[6] J. Zwislocki and R. Feldman, "Just noticeable differences in dichotic phase," *The Journal of the Acoustical Society of America*, vol. 28, p. 860, 1956.

[7] R. Klumpp and H. Eady, "Some measurements of interaural time difference thresholds," *The Journal of the Acoustical Society of America*, vol. 28, p. 859, 1956.

[8] B. Grothe, M. Pecka, and D. McAlpine, "Mechanisms of sound localization in mammals," *Physiological Reviews*, vol. 90, no. 3, pp. 983–1012, 2010.

[9] L. Rayleigh, "Xii. on our perception of sound direction," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214–232, 1907.

[10] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: the duplex theory of sound localization revisited," *The Journal of the Acoustical Society of America*, vol. 111, p. 2219, 2002.

[11] B. C. Moore and B. C. Moore, *An introduction to the psychology of hearing*. Academic press San Diego, 2003, vol. 4.

[12] D. Begault, *3-D Sound for Virtual Reality and Multimedia*. AP Professional, 1994.

[13] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Mit Press, 1997.

[14] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space," in *Audio Engineering Society Convention 107*. Audio Engineering Society, 1999.

[15] G. zheng Yu, B. sun Xie, and X. xu Chen, "Analysis on minimum-phase characteristics of measured head-related transfer functions affected by sound source responses," *Computers and Electrical Engineering*, vol. 38, no. 1, pp. 45 – 51, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0045790611000772

[16] N. N. de Moura, J. Seixas, A. V. Greco *et al.*, "Independent component analysis for optimal passive sonar signal detection," in *Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on*. IEEE, 2007, pp. 671–678.

[17] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1692–1715, 2010.

[18] L. Brayda, C. Wellekens, M. Matassoni, and M. Omologo, "Speech recognition in reverberant environments using remote microphones," in *Multimedia, 2006. ISM'06. Eighth IEEE International Symposium on*. IEEE, 2006, pp. 584–591.

[19] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, ser. Digital Signal Processing. Springer, 2010.

[20] A. D. Waite and A. Waite, *Sonar for practising engineers*. Wiley, 2002, vol. 3.

[21] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer, 2008, vol. 1.

[22] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, vol. 1. IEEE, 2004, pp. 1033–1038.

[23] R. Duraiswami, D. Zotkin, and L. S. Davis, "Active speech source localization by a dual coarse-to-fine search," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 5. IEEE, 2001, pp. 3309–3312.

[24] J.-M. Valin, F. Michaud, and J. Rouat, "Robust 3d localization and tracking of sound sources using beamforming and particle filtering," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 4. IEEE, 2006, pp. IV–IV.

[25] Y.-H. Hu and D. Li, "Energy based collaborative source localization using acoustic micro-sensor array," in *Multimedia Signal Processing, 2002 IEEE Workshop on*, 2002, pp. 371–375.

[26] X. Sheng and Y.-H. Hu, "Energy based acoustic source localization," in *Information Processing in Sensor Networks*. Springer, 2003, pp. 285–300.

[27] R. Niu and P. Varshney, "Target location estimation in sensor networks with quantized data," *Signal Processing, IEEE Transactions on*, vol. 54, no. 12, pp. 4519–4528, 2006.

[28] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, p. 384, 2000.

[29] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, 1976.

[30] B. Champagne, S. Bédard, and A. Stéphenne, "Performance of time-delay estimation in the presence of room reverberation," *Speech and Audio Processing, IEEE Transactions on*, vol. 4, no. 2, pp. 148–152, 1996.

[31] G. C. Carter, "Coherence and time delay estimation," *Proceedings of the IEEE*, vol. 75, no. 2, pp. 236–255, 1987.

[32] F. R. Moore, *Elements of computer music*. Prentice-Hall, Inc., 1990.

[33] H. Olson, *Modern Sound Reproduction*. R. E. Krieger Publishing Company, 1978.

[34] F. Rumsey, *Spatial Audio*. Taylor & Francis, 2001.

[35] F. Rumsey and T. McCormick, *Sound and Recording*. Elsevier/Focal, 2009.

[36] A. D. Blumlein, "British patent 394,325," *Directional effect in sound systems*, 1931.

[37] V. Pulkki and M. Karjalainen, "Localization of amplitude-panned virtual sources i: stereophonic panning," *Journal of the Audio Engineering Society*, vol. 49, no. 9, pp. 739–752, 2001.

[38] B. B. Bauer, "Phasor analysis of some stereophonic phenomena," *The Journal of the Acoustical Society of America*, vol. 33, no. 11, pp. 1536–1539, 2005.

[39] X. Bosun, "Signal mixing for a 5.1-channel surround sound system'analysis and experiment," *Journal of the Audio Engineering Society*, vol. 49, no. 4, pp. 263–274, 2001.

[40] V. Pulkki *et al.*, *Spatial sound generation and perception by amplitude panning techniques*. Helsinki University of Technology, 2001.

[41] (2010) Recommendation itu-r bs.775-2 multichannel stereophonic sound system with and without accompanying picture. Accessed: 2014-03-15. [Online]. Available: http://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.775-2-200607-S!!PDF-E.pdf

[42] Why ambisonics offers "the best sounds surround". Accessed: 2014-03-5. [Online]. Available: http://www.ambisonic.net/

[43] M. A. Gerzon, "Periphony: With-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1973.

[44] D. H. Cooper and T. Shiga, "Discrete-matrix multichannel stereo," *Journal of the Audio Engineering Society*, vol. 20, no. 5, pp. 346–360, 1972.

[45] M. A. Gerzon, "Ambisonics in multichannel broadcasting and video," *Journal of the Audio Engineering Society*, vol. 33, no. 11, pp. 859–871, 1985.

[46] J. S. Bamford and J. Vanderkooy, "Ambisonic sound for us," in *Audio Engineering Society Convention 99*. Audio Engineering Society, 1995.

[47] A. Heller, R. Lee, and E. Benjamin, "Is my decoder ambisonic?" in *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

[48] F. Hollerweger. (2005) An introduction to higher order ambisonic. Accessed: 2014-03-5. [Online]. Available: http://flo.mur.at/writings/HOA-intro.pdf

[49] Hoa technical notes - b-format. Accessed: 2014-03-5. [Online]. Available: http://www.blueripplesound.com/b-format

[50] M. A. Gerzon and G. J. Barton, "Ambisonic decoders for hdtv," in *Audio Engineering Society Convention 92*. Audio Engineering Society, 1992.

[51] D. Moore and J. Wakefield, "Exploiting human spatial resolution in surround sound decoder design," in *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

[52] ——, "The design of ambisonic decoders for the itu 5.1 layout with even performance characteristics," in *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

[53] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[54] ——, "Compensating displacement of amplitude-panned virtual sources," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society, 2002.

[55] W. Snow, "Basic principles of stereophonic sound," *Audio, IRE Transactions on*, no. 2, pp. 42–53, 1955.

[56] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2778, 1993.

[57] B. B. Baker and E. T. Copson, *The mathematical theory of Huygens' principle*. American Mathematical Soc., 2003, vol. 329.

[58] U. Horbach, A. Karamustafaoglu, and M. M. Boone, "Practical implementation of a data-based wave field reproduction system," in *Audio Engineering Society Convention 108*. Audio Engineering Society, 2000.

[59] M. Naoe, T. Kimura, Y. Yamakata, and M. Katsumoto, "Performance evaluation of 3d sound field reproduction system using a few loudspeakers and wave field synthesis," in *Universal Communication, 2008. ISUC'08. Second International Symposium on*. IEEE, 2008, pp. 36–41.

[60] E. Corteel, L. Rohr, X. Falourd, K.-V. Nguyen, H. Lissek *et al.*, "Practical 3 dimensional sound reproduction using wave field synthesis, theory and perceptual validation," *Acoustics 2012 Nantes*, 2012.

[61] S. Spors, R. Rabenstein, and J. Ahrens, "The theory of wave field synthesis revisited," in *124th AES Convention*, 2008, pp. 17–20.

[62] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.

[63] J. Merimaa and V. Pulkki, "Spatial impulse response rendering," in *Proc. of the 7th Intl. Conf. on Digital Audio Effects (DAFX'04), Naples, Italy*, 2004.

[64] ——, "Spatial impulse response rendering i: Analysis and synthesis," *Journal of the Audio Engineering Society*, vol. 53, no. 12, pp. 1115–1127, 2005.

[65] J. Vilkamo, T. Lokki, and V. Pulkki, "Directional audio coding: Virtual microphone-based synthesis and subjective evaluation," *Journal of the Audio Engineering Society*, vol. 57, no. 9, pp. 709–724, 2009.

[66] V. Pulkki, M. Laitinen, J. Vilkamo, J. Ahonen, T. Lokki, and T. Pihlajamäki, "Directional audio coding-perception-based reproduction of spatial sound," in *Int. Workshop on the Principles and Applications of Spatial Hearing, Miyagi, Japan*, 2009.

[67] R. E. A. C. Paley and N. Wiener, *Fourier transforms in the complex domain*. American Mathematical Soc., 1934, vol. 19.

[68] Y. Wang and M. Vilermo, "Modified discrete cosine transform: its implications for audio coding and error concealment," *Journal of the Audio Engineering Society*, vol. 51, no. 1/2, pp. 52–61, 2003.

[69] B. Ricaud, G. Stempfel, B. Torrésani, C. Wiesmeyr, H. Lachambre, and D. Onchis, "An optimally concentrated gabor transform for localized time-frequency components," *Advances in Computational Mathematics*, pp. 1–20, 2013.

[70] A. Kumar, G. Singh, and S. Anurag, "Design of nearly perfect reconstructed non-uniform filter bank by constrained equiripple fir technique," *Applied Soft Computing*, vol. 13, no. 1, pp. 353–360, 2013.

[71] O. Thiergart, G. Del Galdo, M. Taseska, J. A. Pineda Pardo, and F. Kuech, "In situ microphone array calibration for parameter estimation in directional audio coding," in *Audio Engineering Society Convention 128*. Audio Engineering Society, 2010.

[72] V. Pulkki, C. Faller *et al.*, "Directional audio coding: Filterbank and stft-based design," in *Preprint 120th Conv. Aud. Eng. Soc.*, no. LCAV-CONF-2006-022, 2006.

[73] A. Papoulis, *Signal analysis*. McGraw-Hill, 1978, vol. 191.

[74] R. G. Lyons, *Understanding digital signal processing*. Pearson Education, 2010.

[75] J. T. Machado, B. Paatkai, and I. J. Rudas, *Intelligent engineering systems and computational cybernetics*. Springer, 2009.

[76] R. L. Allen and D. Mills, *Signal Analysis: time, frequency, scale, and structure*. John Wiley & Sons, 2004.

[77] D. Gabor, "Theory of communication. part 1: The analysis of information," *Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of*, vol. 93, no. 26, pp. 429–441, 1946.

[78] P. Balazs, "Regular and irregular gabor multipliers with application to psychoacoustic masking," *PhD thesis, University of Vienna*, 2005.

[79] R. Carmona, W.-L. Hwang, and B. Torresani, *Practical Time-Frequency Analysis: Gabor and Wavelet Transforms, with an Implementation in S*. Academic Press, 1998, vol. 9.

[80] P. L. Søndergaard *et al.*, "An efficient algorithm for the discrete gabor transform using full length windows," in *SAMPTA'09, International Conference on Sampling Theory and Applications*, 2009.

[81] W. Heisenberg, "Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik," *Zeitschrift fur Physik*, vol. 43, pp. 172–198, Mar. 1927.

[82] G. B. Folland and A. Sitaram, "The uncertainty principle: a mathematical survey," *Journal of Fourier Analysis and Applications*, vol. 3, no. 3, pp. 207–238, 1997.

[83] F. Hlawatsch and G. F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations," *IEEE Signal Processing Magazine*, vol. 9, no. 2, pp. 21–67, 1992.

[84] J. R. Deller, J. G. Proakis, and J. H. Hansen, *Discrete-time processing of speech signals*. Ieee New York, NY, USA:, 2000.

[85] R. Martin and R. V. Cox, "New speech enhancement techniques for low bit rate speech coding," in *Speech Coding Proceedings, 1999 IEEE Workshop on*. IEEE, 1999, pp. 165–167.

[86] "Idgt - inverse discrete gabor transform," in *LTFAT The Large Time-Frequency Analysis Toolbox*, accessed: 2014-3-5. [Online]. Available: http://ltfat.sourceforge.net/doc/gabor/idgt.php

[87] H. Khaddour, "A comparison of algorithms of sound source localization based on time delay estimation," *Elektrorevue,Internetovy casopis*, vol. 2, no. 1, pp. 31–37, 2011.

[88] H. Khaddour and A. Warda, "Sound source localization using time delay estimation." in *6th International Conference on Teleinformatics ICT.*, 2011, pp. 179–182.

[89] Minispl measurement microphone. Accessed: 2014-2-5. [Online]. Available: http://www.nti-audio.com/Portals/0/data/en/MiniSPL-Measurement-Microphone-Product-Data.pdf

[90] Ø. Hjelle and M. Dæhlen, *Triangulations and Applications*, ser. Mathematics and Visualization. Physica-Verlag, 2006.

[91] H. Khaddour and M. Trzos, "Representation of sound field using ambisonic," *Elektrorevue,Internetovy casopis*, vol. 1, no. 2, pp. 1–7, 2010.

[92] M. Frank, F. Zotter, and A. Sontacchi, "Localization experiments using different 2d ambisonics decoders," in *25th TONMEISTERTAGUNG – VDT INTERNATIONAL CONVENTION*, 2008, accessed: 2014-5-5. [Online]. Available: http://iem.kug.ac.at/fileadmin/media/iem/projects/2008/tmt08.pdf

[93] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, ser. Springer series in information sciences. Springer, 2007.

[94] S. Bertet, J. Daniel, L. Gros, E. Parizet, and O. Warusfel, "Investigation of the perceived spatial resolution of higher order ambisonics sound fields: A subjective evaluation involving virtual and real 3d microphones," in *Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments*. Audio Engineering Society, 2007.

[95] M. Trzos and H. Khaddour, "Localization blur of 2d ambisonic reproduction in small rooms." in *Telecommunications and Signal Processing (TSP), 2010 33th International Conference on*, August 2010, pp. 56–59.

[96] E. Benjamin, A. Heller, and R. Lee, "Localization in horizontal-only ambisonic systems," in *Audio Engineering Society Convention 121*. Audio Engineering Society, 2006.

[97] M. A. Gerzon, "The design of precisely coincident microphone arrays for stereo and surround sound," in *Audio Engineering Society Convention 50*. Audio Engineering Society, 1975.

[98] H. Khaddour, "Sound source localization based on b-format signals," in *Telecommunications and Signal Processing (TSP), 2011 34th International Conference on*. IEEE, 2011, pp. 335–338.

[99] H. Khaddour and J. Schimmel, "Multiple sound sources localization using energetic analysis method," *Elektrorevue,Internetovy casopis*, vol. 3, no. 4, pp. 5–9, 2012.

[100] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. Access Online via Elsevier, 1999.

[101] J. Ahonen, M. Kallinger, F. Küch, V. Pulkki, and R. Schultz-Amling, "Directional analysis of sound field with linear microphone array and applications in sound reproduction," in *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

[102] H. Khaddour and D. Kurc, "Impact of applied transform on accuracy of energetic analysis method," in *Telecommunications and Signal Processing (TSP), 2013 36th International Conference on*, July 2013, pp. 464–468.

[103] H. Khaddour, J. Schimmel, and M. Trzos, "Estimation of direction of arrival of multiple sound sources in 3d space using b-format," *International Journal of Advances in Telecommunications, Electrotechnics, Signals and Systems*, vol. 2, no. 2, pp. 63–67, 2013.

[104] Sound field technology, how was it work. Accessed: 2013-12-5. [Online]. Available: http://www.soundfield.com/soundfield/soundfield.php

[105] H. Khaddour, J. Schimmel, and F. Rund, "A novel combined system of direction estimation and sound zooming of multiple speakers," *Radioengineering*, vol. 24, no. 2, 2015, in print.

[106] R. Schultz-Amling, F. Kuech, O. Thiergart, and M. Kallinger, "Acoustical zooming based on a parametric sound field representation," in *Audio Engineering Society Convention 128*. Audio Engineering Society, 2010.

[107] N. Wittenberg, *Understanding Voice Over IP Technology.* Cengage Learning, 2009.

[108] S. Furui, *Digital speech processing: synthesis, and recognition.* CRC Press, 2000.

[109] V. Pulkki, "Applications of directional audio coding in audio," in *Proceedings of the 19th International Congress of Acoustics*, 2007.

[110] Methods for subjective determination of transmission quality. Accessed: 2013-12-5. [Online]. Available: http://www.itu.int/rec/T-REC-P.800-199608-I/en

[111] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *Latent Variable Analysis and Signal Separation.* Springer, 2012, pp. 430–437.

[112] R. Huber and B. Kollmeier, "Pemo-q ; a new method for objective audio quality assessment using a model of auditory perception," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 1902–1911, 2006.

[113] S. A. Gelfand, *Essentials of audiology.* Thieme, 2009.

[114] K. Kondo, *Subjective quality measurement of speech: its evaluation, estimation and applications.* Springer Science & Business Media, 2012.

Author's Publications

[1] H. Khaddour and M. Trzos, "Representation of sound field using ambisonic," *Elektrorevue,Internetovy casopis*, vol. 1, no. 2, pp. 1–7, 2010.

[2] M. Trzos and H. Khaddour, "Localization blur of 2d ambisonic reproduction in small rooms." in *Telecommunications and Signal Processing (TSP), 2010 33th International Conference on*, August 2010, pp. 56–59.

[3] H. Khaddour, "A comparison of algorithms of sound source localization based on time delay estimation," *Elektrorevue,Internetovy casopis*, vol. 2, no. 1, pp. 31–37, 2011.

[4] H. Khaddour and A. Warda, "Sound source localization using time delay estimation." in *6th International Conference on Teleinformatics ICT.*, 2011, pp. 179–182.

[5] H. Khaddour, "Sound source localization based on b-format signals," in *Telecommunications and Signal Processing (TSP), 2011 34th International Conference on.* IEEE, 2011, pp. 335–338.

[6] M. Trzos, and H. Khaddour "Efficient Spectral Estimation of Non- Stationary Harmonic Signals Using Harmonic Transform, " *International Journal of Advances in Telecommunications, Electrotechnics, Signals and Systems*, vol. 1, no. 2- 3, pp. 1–3, 2012.

[7] H. Khaddour and J. Schimmel, "Multiple sound sources localization using energetic analysis method," *Elektrorevue,Internetovy casopis*, vol. 3, no. 4, pp. 5–9, 2012.

[8] H. Khaddour and D. Kurc, "Impact of applied transform on accuracy of energetic analysis method," in *Telecommunications and Signal Processing (TSP), 2013 36th International Conference on*, July 2013, pp. 464–468.

[9] H. Khaddour, J. Schimmel, and M. Trzos, "Estimation of direction of arrival of multiple sound sources in 3d space using b-format," *International Journal of Advances in Telecommunications, Electrotechnics, Signals and Systems*, vol. 2, no. 2, pp. 63–67, 2013.

[10] D. Kurc, V. Mach, K. Orlovsky, and H. Khaddour, "Sound Source Localization with DAS Beamforming Method Using Small Number of Microphones," in *Telecommunications and Signal Processing (TSP), 2013 36th International Conference on*, July 2013, pp. 526–532.

[11] H. Khaddour, J. Schimmel, and F. Rund, "A novel combined system of direction estimation and sound zooming of multiple speakers," *Radioengineering*, vol. 24, no. 2, 2015, in print.

Appendix A

Laboratory

Sound recordings and listening tests were carried out in the acoustic laboratory at Department of Telecommu- nications FEEC, Brno University of Technology that meets the ITU-R BS.1116-1 requirements for the listening conditions and reproduction devices; the laboratory provides semi-diffuse field with reverberation time $RT60$ around 0.3 s for one-third octave bands from 125 Hz, see Figure A.1.

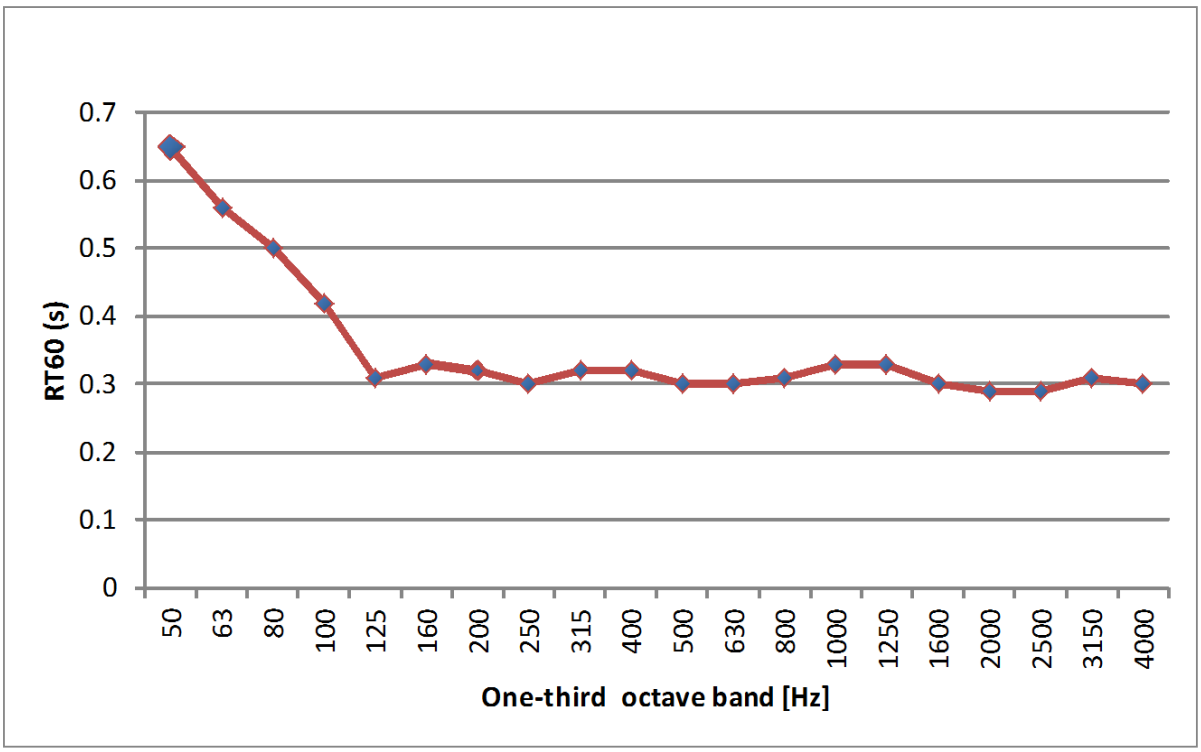


Figure A.1: $RT60$ measured in the laboratory.

Curriculum Vitae

Personal Information

Name: Hasan Khaddour.
Title: Ing.
Place of Birth: Latakia, Syria.
Nationality: Syrian.
Telephone: +420 776 670 485.
Email : xkhadd00@stud.feec.vutbr.cz

Education

2009-2015: Ph.D. student at Brno University of Technology, Czech Republic.
2002-2007: Engineering student at Tishreen University, Syria.
1999-2002: Upper Secondary Education, Latakia, Syria.

Employment History

2007 - 2008: Assistant at Faculty of Mechanical and Electrical Engineering, Department of Communication and Electronic Engineering.

Languages

Arabic - native speaker.
English - excellent command.
Czech - very good command.

Honors and Awards

Top Students Award for five years in a row at Tishreen University, Syria.

Participation in Projects

- FEKT-S-11-17 – Research of Sophisticated Methods for Digital Audio and Image Signal Processing. Holder: Prof. Z. Smékal. 2011.
- MSM21630513 – Electronic Communication Systems and Technologies of Novel Generations (ELKOM). Holders: Prof. Z. Raida, Prof. K. Vrba, Prof. J. Jan. 2008–2011.
- 1595/F1/2012 Innovation of Studio Engineering subject. Holder: Ing. J. Schimmel, 2012.