Optimizing Static Degrees of Freedom in Sound Field Reproduction

by

Hanieh Khalilian

M.Sc., Sharif University of Technology, 2008
B.Sc., Iran University of Science and Technology, 2006

Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Engineering Science
Faculty of Applied Science

SIMON FRASER UNIVERSITY
Summer 2016

Approval

Name:	**Hanieh Khalilian**
Degree:	**Doctor of Philosophy (Electrical Engineering)**
Title:	***Optimizing Static Degrees of Freedom in Sound Field Reproduction***
Examining Committee:	**Chair:** Dr. Bonnie L. Gray Professor

Dr. Ivan V. Bajić
Senior Supervisor
Associate Professor _____

Dr. Rodney G. Vaughan
Co-Supervisor
Professor _____

Dr. Jie Liang
Internal Examiner
Professor _____

Dr. Thushara D. Abhayapala
External Examiner
Professor
College of Engineering and Computer Science
Australian National University _____

Date Defended: 31 August 2016

Abstract

Sound Field Reproduction (SFR) is for creating a desired sound field from a primary source by using multiple loudspeakers. Its research calls for numerical experiments by simulation. SFR performance can be improved by optimizing the static Degrees of Freedom, i.e., the locations and patterns of loudspeakers. The approach is to decrease the sound field reproduction error without increasing the operational complexity, and this requires that the possible locations and frequencies of the primary source are known a priori.

To optimize the loudspeaker locations, two placement methods are developed. In the first method, an idealized Acoustic Transfer Function (ATF) matrix that minimizes the reproduction error, but which may not be realizable, is derived for a fixed number of uniformly placed, omnidirectional loudspeakers. The loudspeakers are then re-positioned within their aperture so that their realizable ATF matrix best approximates the idealized ATF matrix. In the second method, a new algorithm is called âĂŞ Constrained Matching Pursuit (CMP), which optimizes loudspeaker location while constraining the total loudspeaker power to avoid acoustic hotspots. CMP is also used to jointly optimize the radiation patterns and locations of the loudspeakers. These methods optimize for a single frequency, and a method is presented which extends the case to multiple frequencies for the primary source. The multi-frequency method is deployed in the audio layer of an immersive communications system. An existing model for the Head Related Transfer Function (receiving pattern) is adapted for the computation of the loudspeaker excitation functions, called the dynamic Degrees of Freedom.

Subjective and objective tests are applied and concur that the quality of speech of the SFR in a reverberant room is significantly improved compared to a system with the same number of loudspeakers that are uniformly spaced and omni-directional, and which have the same total power constraint and computation complexity.

Keywords: Sound Field Reproduction, Placement Optimization, Pattern Optimization, Immersive Communication.

Dedication

I dedicate this thesis to my husband, Mohammad. None of these were possible without him.

I would like to dedicate this thesis to my Mom, my aunt, my Dad, Bahar and Ali. I love you all from the bottom of my heart.

Acknowledgements

I would like to thank my senior supervisor Dr. Ivan Bajić for his help, support and patience during my PhD. He is a great help and always available for his students. I really enjoyed working with him, and I learned a lot from him. I am grateful to my co-supervisor, Dr. Rodney Vaughan. He always encouraged me during my time here at SFU and he is a source of energy and passion for everyone who is working with him. For me, he is more than an academic supervisor. he taught me how to live beautifully.

I would like to thank my committee members: Dr. Jie Liang, Dr. Bonnie Gray, and Dr Thushara Abhayapala for their valuable questions and comments during the defense and after.

I would like to express my deepest gratitude to my mother, Zahra Rafiean. Her unconditional love and support always gives me strength to follow my heart. She devoted her live to her children for which I can never thank her enough. I always feel her close to myself even if I live far from her. I like to thank my dad, Mohammadreza Khalilian, for all that he has selflessly done for me throughout my life. I really like to thank my aunt, Zohreh Rafieian. She has been and continues to be the epitome of encouragement and inspiration in my life. Without her, I have not been where I am today. She has not been anything less than my mother for me, and I feel very fortunate to have her in my life. I also thank my lovely sister, Bahar Khalilian, and my lovely brother, Ali Khalilian, for being there always for me and for supporting me.

This dissertation would not have been possible without My husband, Mohammad Shahid Zadeh Mahani. He supported me in every single moment of my PhD study. His unconditional love and support were my biggest source of comfort in toughest times. He is my best friend and mentor during the difficult days of my life and my study. I learned from him to stand up and continue even in difficult situations. Here is the best place to thank my mother and my father in law.

I like to thank those people who helped me to start this path: Dr. Shahrokh Ghaemmaghami, Dr. Iman Gholampour, Shaghayegh Norastefar, Salma Mirhadi, Mona Omidyeganeh, Soodeh Ahani, and Mansooreh Shakeri. I also like to thank those people to make the life easier for me here in Canad where I was far from my family: Golnoosh Saeedi Nejad, Ali Hoseini, Mahsa Najibi, Amir Valizadeh, Parastoo Geranmayeh, Sara Namazi, Hosein Kha-

toonabadi, Hamid Mirzaei, Ehsan Asadi, Mehdi Seyfi, Reza Mohamadnia, Zamzam Kordi, and Behnam Molavi.

I would like to thank my colleagues at Kongsberg Mesotech Ltd.: Rob Huxtable, Steve Pearce, and Jinyun Ren. I had a great time with them for the last 18 months at this company.

I like to thank Renée McCallum, Golnoosh Saeedi Nejad, Steve Pearce, Maral Dehghani, and Mohamad Shahidzadeh Mahani for reading and editing this dissertation and my presentation, and helped me to improve them.

Table of Contents

List of Tables

List of Figures

List of Acronyms

2-D	Two-Dimensional
3-D	Three-Dimensional
ANC	Active Noise Cancellation
ATF	Acoustic Transfer Function
CMP	Constrained Matching Pursuit
DoF	Degree of Freedom
FSFR	Focused Sound Field Reproduction
HOA	Higher Order Ambisonics
HRTF	Head Related Transfer Function
ILD	Interaural Level Difference
IPD	Interaural Phase Difference
ITD	Interaural Time Difference
LASSO	Least Absolute Shrinkage and Selection Operator
LR	Loudspeaker Region
LS	Least Square
MIMO	Multiple inputs Multiple outputs
MP	Matching Pursuit
SFR	Sound Field Reproduction
SNR	Signal to Noise Ratio
SVD	Singular Value Decomposition
WFS	Wave Field Synthesis

List of Symbols

p	The pressure
A	The amplitude of primary source
$\mathbf{G}$	The ATF matrix
$\mathbf{G}^{\text{ideal}}$	The ideal ATF matrix
M	The number of sampling points
m	The index of sampling point
N	The number of loudspeakers
n	The index of loudspeakers
$G_{m,n}$	The Green's function of n-th loudspeaker at m-th sampling point
$G'_{m,n}$	The ATF of n-th loudspeaker at m-th sampling point
$g'_{m,n}$	The Acoustic Impulse Response of n-th loudspeaker at m-th sampling point
$\mathbf{x}_n$	The location of n-th loudspeaker
$\mathbf{y}_m$	The location of m-th sampling point
$\mathbf{s}$	The complex amplitudes of the loudspeakers
$\mathbf{s}^{\text{opt}}$	The optimum value of $\mathbf{s}$
$\mathbf{p}^{\text{des}}$	The desired vector
$\mathbf{p}$	The reproduced vector
$\mathbf{e}$	The error vector
f	The frequency
ω	The angular frequency
k	The wave number
C	The speed of sound
N_v	The number of candidate locations
N_c	The parameter of SVD-based placement method
$\mathbf{h}$	The ATF of the candidate locations
$p_{\max}$	The maximum normalized power
$\mathfrak{D}$	The dictionary
$\mathbf{b}$	The dictionary members
λ	The wavelength
ρ_0	The air density

ζ	The upper limit of the ratio of the
	ℓ_2-norm of the reproduced vector to that of the desired one
γ	The regularization parameter in Least Square algorithm
γ^{lasso}	The regularization parameter in Lasso algorithm
$\|.\|_\eta$	The ℓ_η-norm of a vector
σ^g	The singular values of matrix $\mathbf{G}$
$\mathbf{U}^g$	The left singular matrix of $\mathbf{G}$
$\mathbf{V}^g$	The right singular matrix of $\mathbf{G}$
$\mathbf{\Sigma}^g$	The singular matrix of $\mathbf{G}$
$R^n\mathbf{p}^{\text{des}}$	The residual vector of $\mathbf{p}^{\text{des}}$ at n-th iteration
α_n	The coefficients of CMP algorithm
p_n	The assigned power to n-th iteration by CMP algorithm
$p_{\min}$	The minimum assigned power to the iterations in CMP algorithm
J	The Lagrangian cost function
$\mathcal{L}$	The radiation pattern of a loudspeaker
$\mathcal{M}$	The receiving pattern of a microphone
$\mathcal{T}$	The radiation pattern of a talker
$\mathcal{H}$	The receiving pattern of human head
$Y_{m_d,l}(.)$	The spherical harmonic function of order l and degree m_d
$h_l(.)$	The spherical Hankel function of order l
$H_l(.)$	The cylindrical Hankel function of order l
w_l	The harmonic coefficients of a 2-D higher order loudspeakers
L	The order of loudspeaker
$r_{\min}$	The minimum distance between the LR and listening area
cond(.)	The condition number of a matrix
$C_{m_d,l}$	The harmonic coefficients of a higher order loudspeaker
$C'_{m_d,l}$	The expansion coefficients of a higher order loudspeaker
SNR_{mic}	The signal to noise ratio of microphones
L_t	The order of truncation in HOA method
$\mathbf{v}$	The microphone noise vector
T	The threshold of active talker detection

Chapter 1

Introduction

1.1 Sound field reproduction

"Sound Field Reproduction" (SFR), also known as sound field synthesis, sound field rendering, and audio holography, is the recreation of a desired sound field in a region of interest by using an array of loudspeakers. While the goal of SFR is a physical system for improving the real-world human acoustic experience, the current state-of-the-art of SFR is simulated systems. There are several reasons for this: real-world systems and experiments are too expensive for most research institutions; progress on the fundamentals, including performance, feasibility and limits, requires new theory, techniques and algorithms; simulation allows sophisticated systems to be modelled and assessed with mathematical metrics of performance. Currently, real-world SFR systems that are feasible for all acoustic situations do not exist, and this is why SFR is topical research. In short, it is emphasized that the SFR systems treated in this thesis comprise mathematical models and simulations, rather than real-world systems. Similarly, the references cited are for models and simulations unless stated. The following terminology is often used in SFR, and they are illustrated in Fig. 1.1.

Primary source: The source(s) of the desired sound field is called the primary source. In SFR literature, it is assumed that the desired sound field is either originating from a point source or being modeled by a plane wave. The reason is that any desired sound field can be represented in terms of plane waves or spherical waves.

The spherical waves originate from point sources. The frequency domain representation of the pressure created by a point source located at point $\mathbf{x}_0$ is:

$$p(\mathbf{x}) = A\frac{e^{ik\|\mathbf{x}-\mathbf{x}_0\|_2}}{4\pi\|\mathbf{x}-\mathbf{x}_0\|_2} \tag{1.1}$$

where $i = \sqrt{-1}$, A is the complex amplitude of the point source, $k = 2\pi f/C$ is the wave number, f is the frequency, and C is the speed of sound.

The desired field from a point source is characterized by a plane wave when the source is located far enough from the listening area. The dynamic pressure caused by a plane wave at point $\mathbf{x}$ is:

$$p(\mathbf{x}) = Ae^{i\mathbf{k}\cdot\mathbf{x}} \tag{1.2}$$

where $\mathbf{k}\cdot\mathbf{x}$ is the inner product of vectors $\mathbf{k}$ and $\mathbf{x}$, $\mathbf{k} = k\cdot\vec{\mathbf{k}}$, and $\vec{\mathbf{k}}$ is the unit vector in the direction of the plane wave.

Secondary sources: The loudspeakers which create the sound field are referred to as secondary sources. The sound field in the listening area depends on the radiation patterns, locations, and the complex amplitudes (amplitude and phase) of the loudspeakers. The complex amplitudes are also referred to as "driving functions" or "excitations" in the SFR literature. Given the locations and radiation patterns of the loudspeakers, different SFR methods try to optimize the complex amplitudes. Other methods also optimize the loudspeakers' locations and patterns in order to further improve the SFR performance. Optimization of the loudspeakers' placement and pattern is the main focus of this thesis. Another important factor in SFR system is the power of the secondary source which impacts the system performance and will be treated in this thesis.

Listening area: The region in which the sound field is to be recreated with the smallest possible error is called the "listening area" or "region of interest". In conventional SFR systems, the listening area is composed of two points around the ear canals. These points are called "sweet points" [1]. In modern acoustics, the listening area is a 2-D surface or a 3-D volume. Some SFR methods consider the listening area as a continuous region and recreate the sound field analytically in this region. In a sampled listening area, the effort is to reproduce at discrete points only. This thesis focuses on the second approach, and the sampled locations of the listening area are referred to as "sampling points". The number and arrangement of the sampling points also determine the accuracy of modeled SFR systems. The sound field reproduced inside of the listening area is referred to as "interior sound field", and the sound field which is reproduced outside of the listening area is called " exterior sound field".

Loudspeaker region: The region in which loudspeakers (secondary sources) are located is called the "Loudspeaker Region" (LR). The LR is also called the source aperture. The majority of the SFR literature consider the LR to be around the listening area in order to have the best coverage. For instance, for 2-D SFR, the loudspeakers are located on the perimeter of a circle, and the listening area is a disk at the center of that circle. For 3-D SFR systems, the loudspeakers are on the surface of a sphere while the listening area is a smaller sphere inside the LR. However, when the possible locations of the primary sources are limited, the LR does not need to cover the whole listening area. In this case, a linear or planar LR can be placed at one side of the listening area which covers only a part of the listening area.

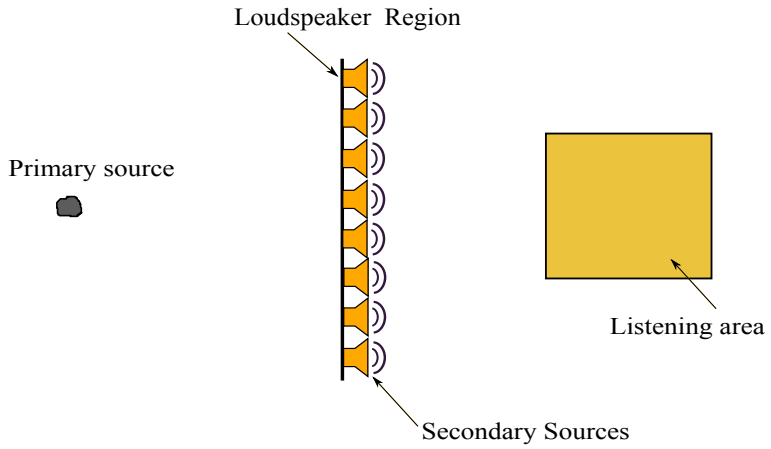


Figure 1.1: Illustration of a Sound Field Reproduction system

The LR is usually located between the primary source and the listening area, as shown in Fig. 1.1, which is the case in this thesis. The SFR system in which the primary source is located between the LR and the listening area is called "Focused Sound Field Reproduction" (FSFR) system (not shown here).

Reproduction error: In a modeled SFR system, the difference between the desired sound field and the one reproduced by the loudspeaker array is defined as the error field. The energy of the error field is known as the "reproduction error". In an SFR system, different parameters are selected such that the reproduction error in the listening area is minimized. In addition to the reproduction error, other kinds of perceptual error metrics also exist for SFR systems. In this thesis, some of these perceptual metrics are measured through the numerical system models.

1.2 Applications of sound field reproduction

The goal of an SFR system is to recreate a desired field such that the listeners feel they are in the original environment where the sound field is created. SFR has many applications in different industries. The motivating applications are reviewed in this section.

Active Noise Cancellation: The process of reducing the total noise sound field in a zone is called noise cancellation. This process can be achieved by isolating the zone using sound absorbent materials, or it can be done by generating an anti-noise field in the zone using loudspeakers. The former is called passive noise cancellation while the latter is referred to as active noise cancellation. Active noise cancellation can be decomposed into two tasks: (1) estimating the parameters of the unwanted sound and (2) generating the inverse sound field such that the superposition of the generated field and the unwanted one leads to cancellation. In this context, the anti noise is the desired sound field, and the listening area becomes the silent zone. The goal of SFR systems in this application is to

recreate a silent region while the loudness of the sound field does not change considerably outside the silent zone. Limiting the power of loudspeakers is one way to limit the power of the sound field outside of the listening area. A simple way forward is to split the listening area to two separate regions: inside and outside of the silent zone. In this method, the sound field inside of the silent zone is the anti noise while the desired field outside of the silent zone, the second listening area, is minimized.

Echo cancellation: An echo is an attenuated version of the original sound signal which arrives to the listener after being reflected from the surrounding medium. Therefore, the echo always has a delay. The reflections which arrive just after the direct sound are early echos, and the reflections which arrive after a long time are referred to as late echos which cause multiple hearing. This multiple hearing of the same sound often degrades the sound quality, for example for speech signal. In this case, an SFR system can be deployed to remove the late echos and improves the sound quality. In this application, by selecting the desired field to be the opposite of the late echos, the superposition of the received sound and the recreated one will be equal to summation of the direct signal and early echos.

3-D cinema: In this application an array of loudspeakers are located around the cinema hall in order to recreate the sounds of the movie such that the audiences feel that they are in the movie scene. Changing the direction of the sound based on what is happening in the movie at each moment is one way to create this feeling for the listeners.

Multi-zone SFR system: In the multi-zone SFR, different sound fields are regenerated in different listening areas. In this application, different listeners in the same environment, such as a room or a car, are able to listen to different music or radio channels depending on their locations.

Immersive communication: An immersive communication system aims at removing the perception of physical barriers in a communication system. Users of such systems feel that they are sharing the same environment. To implement such systems, visual, auditory, and other senses (such as touch and smell) should be recreated in a way that users feel there is no physical boundary. Therefore, a 3-D sound field reproduction system is an inextricable part of an immersive communication system. This system regenerates the auditory and localization senses for each user based on the information received from the other users. This thesis studies this application of the SFR system in Chapter 5, and this application is elaborated in the next section.

1.3 Immersive communications

Communication systems greatly advanced over the past years, including radio, television, mobile phone, internet-based voice and video calling [2]. However these types of communications are not as natural as meeting face-to-face. For this purpose, immersive communications aims at exchanging the natural social signals between remote people as if

they are in the same environment. From the information point of view, the human visual system can receive information at equivalent rate about 10 Mb/sec, the haptic system 1 Mb/sec, the auditory and olfactory system about 10 Kb/sec, and the gustatory system about 1 Kb/sec [2,3]. Since the links in the internet backbone can currently transfer 1Tb/sec data, how to communicate the data is not so important [2]. The main question is how to design a system in order to communicate in a way that users feel the same naturalness as in face-to-face communication. To answer this question, visual, audio, and other information (related to other senses such as smell and touch) should be captured at the transmitting end and be reproduced at the receiving end. The most important components of an immersive communication system are shown in Fig. 1.2.

Precise recreation of visual information plays an important role in the naturalness of the immersive communication system. The requirements of a visual immersive system are stereoscopy, motion parallax, and adaptive focus. The stereoscopy means that each eye of all viewers must see an appropriate and different view. The motion parallax means that the views in two eyes must change when the viewers move. The adaptive focus implies that based on the focal planes of the viewers the objects should move in and out of focus. To satisfy these requirements, ideally a display with a light field across the pupil of each eye is needed. Head-mounted displays which are close enough to the viewers can satisfy the above-mentioned requirements, but distant displays should be large enough and be wrapped around the viewers to have these requirements [2]. The requirements of an ideal display along with different kinds of displays are explained in the following paragraph.

An ideal light field should show the light ray passing through a 3-D location in space at each angle individually for all frequencies and times. Conventional displays show the light ray for 2-D locations which is fixed for all angles. Stereoscopic 3-D displays provide two views for each eye separately for 2-D locations. To see such a display, special glasses are required to separate the received signal properly such that each eye can see its corresponding view. In these displays, viewers receive a fixed scene independent of their locations, which does not satisfy the requirements of a visual immersive system. Multi-view displays contain N lights in N directions at each 2-D location, and they can be used without glasses. Improving the sense of immersiveness is possible in these displays by increasing the size of the display which leads to high cost in the existing LCD, or plasma, etc., displays. Nowadays, large displays are built using multiple projectors or mosaicing the LCD displays. Most recently, roll-to-roll (R2R) techniques are being employed in the design of video screens for immersive communications. Therefore, designing a display which has all the requirements of the visual immersive system is a challenge and is beyond the scope of this thesis.

The second important aspect of an immersive communication system is the audio layer in these systems. An audio immersive system should work in a way that listeners can distinguish who is talking, what is the relative location and direction of the talker with respect to each listener, and how many talkers are active at each time. For this purpose,

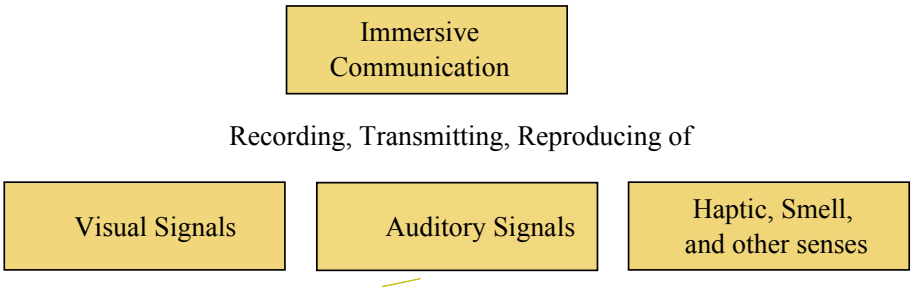


Figure 1.2: Components of an immersive communication system

the sound field should be recorded at one end, transmitted to the other end, and reproduced as if the listeners feel that they are in the same room with the talkers.

Two distinct methods are used for sound field capturing. In the first method, a clean version of an audio signal corresponding to each talker is recorded and transmitted. In this method, the audio files of the talkers should be separated and processed to be free of echo and ambient noise. For this purpose, a microphone can be put close to each talker, or the sound field can be recorded by an array of microphones and processed algorithmically for source separation, echo cancellation, and etc. In the second approach, the sound field is recorded around the heads of virtual listeners without any processing and conveyed for sound field reproduction. This technique is faster and easier because it does not need to process the audio signal. However, this method needs larger bandwidth to transmit the related signals of all listeners. In addition, the exact geometry at the transmitting end may not be similar to the one at the receiving end.

For sound rendering at the receiving room, the sound field can either be reproduced in each ear canal using a headset, or it can be reproduced using an array of loudspeakers around the ear canals acting as listening areas. The disadvantage of using a headset is that the listeners can feel detached from their spatial environment. The second method is more natural, but it encounters challenges such as detecting the listeners' locations, accurately finding the transfer function which depends on each listener, employing a real-time system with low latency, and optimizing the system parameters for SFR.

To solve some of these problems the audio immersive system can be used along with the video screening system. Interaction between audio and video systems can be used in finding the locations of the listeners and talkers, or even in approximation of the listener transfer function which depends on the listener. In this thesis, the acoustic layer of an immersive communication system is designed with the help of a video-based monitoring system, in Chapter 5.

Having ideal auditory and visual systems is a big stride toward a complete immersive system because they simplify the recreation of the other senses. For example, in haptic communication, in order to create the interaction between humans and objects, the shape and size of the objects can be felt through the visual system used for immersive communi-

cation. However, other aspects of this kind of communication are challenging. For example, sense of touch is regarded as a multi-dimensional signal, and various types of sensors are required to recreate this sense. References [4–7] have studied communication of the other senses.

As mentioned earlier, an immersive communications system captures different signals and recreates them such that the user senses that they are receiving natural signals. Therefore, the quality of this systems should be judged by the users, and finding an objective metric which reflects the user experience may be impossible. However, for sound field reproduction there are several metrics that can assess the quality and the sense of direction. In a sound field rendering system, reproduction error measures the accuracy of the sound field rendering while Interaural Level Difference and Interaural Time Difference evaluate the sense of direction.

In the literature, SFR methods are employed in the audio layer of an immersive communication system in different configurations. The key papers are reviewed here. An overview of the requirements and technical challenges related to audio immersive systems can be found in [2, 8, 9]. In [10], a 3-D sound field reproduction system is designed and implemented based on the boundary surface control principle, or Wave Field Synthesis method. The loudspeaker and microphone arrays are constructed in the shape of a dome. In this method the sound field of an environment (such as a jungle or orchestra) is recorded by the microphone array, and then reproduced in a room. The results of the subjective listening experiments show that most of the subjects rated the reproduced sound field as "very good", which was the highest among the rankings.

In [11], a sound field reproduction system for immersive audio is suggested in which the loudspeakers are placed all around the audience in the reproduction room. The loudspeakers' excitations (driving functions) are determined using Wave Field Synthesis. Here, again, subjective tests show that the quality of the reproduced field is high. In the immersive audio system described in [12], loudspeakers are located around the desktop and the sound field originating from a point source located in the middle of the screen is recreated. While the above approaches focused on headphone-free systems, the authors of [13] proposed a low cost sound field reproduction system with the aid of headsets.

In this thesis, an optimized SFR system is used in the audio layer of an immersive communication system, and the perceptual error metrics are measured numerically in order to show how SFR system optimization enhances the quality of an immersive communication system.

1.4 Sound field reproduction techniques

Sound field reproduction (SFR) is a re-emerging field of study that originated more than 50 years ago. One of the earliest references is [1], where a system consisting of two loudspeakers

and two microphones is designed to reproduce a sound field at two points around human ears. This method is called "stereophonic reproduction" and allows the sound field to be recreated in a very small region (two points, "sweet points"). In [14], a multi-channel sound reproduction system is presented which records a sound field on an enclosed surface by an array of microphones, and reproduces the sound field in that enclosed volume in a reverberant room by an array of loudspeakers.

The goal of all SFR methods is to optimize the system parameters such that the reproduction error is minimized in the listening area. Some of these parameters are determined ahead of time, and they are fixed during the system operation. These parameters are called static parameters or static Degrees of Freedom (DoFs), and they are optimized or determined intuitively. Some other parameters are found during the system operation, and they change by changing the desired field. These parameters are called dynamic parameters or dynamic DoFs. In most of the SFR literature, all parameters except the complex amplitudes of loudspeakers are considered as static DoFs. In these systems, the complex amplitudes of the loudspeakers change in real time to minimize the reproduction error. In modern acoustics, there are three prominent classes of SFR methods [15] which find the complex amplitudes of the loudspeakers provided that all other parameters are given:

1. Wave Field Synthesis (WFS) [16–28],
2. Higher Order Ambisonics (HOA) [29–37], and
3. Direct approximation (pressure matching) [15, 38–41].

Optimization of the loudspeakers' locations and radiation patterns as static or dynamic DoFs are two other techniques of sound field reproduction. These five techniques in the literature are reviewed in the following subsections.

1.4.1 Wave field synthesis (WFS)

This method is based on Huygens' principle which states that every sound field can be represented by a superposition of spherical waves. According to the Kirchhoff-Helmholtz (KH) integral, the pressure at each point in a source-free volume can be determined completely if the pressure and velocity on the enclosing surface of that volume are known. Therefore, a desired sound field is recreated in a source-free volume if the pressure and velocity (which is the gradient of pressure) of the desired field are completely regenerated on its surrounding surface [16].

The governing equation of WFS, the KH integral, is:

$$p(\mathbf{x}) = -\int_{\partial V} \left(G(\mathbf{x}|\mathbf{x}_0) \frac{\partial}{\partial \mathbf{n}} S(\mathbf{x}_0) - S(\mathbf{x}_0) \frac{\partial}{\partial \mathbf{n}} G(\mathbf{x}|\mathbf{x}_0) \right) dS_0. \tag{1.3}$$

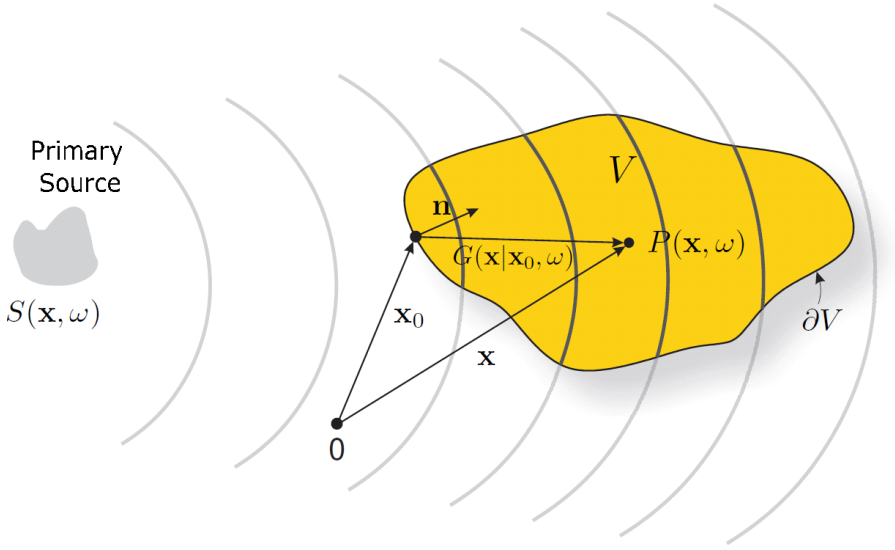


Figure 1.3: Wave Field Synthesis [16].

In this equation, $p(\mathbf{x})$ is the pressure at point $\mathbf{x}$ in volume V, G is Green's function, S is the pressure caused by the primary source on the enclosing surface of V, and $\mathbf{n}$ is the unit vector inward perpendicular to the enclosing surface of V . This integral can be visualized with the help of Fig. 1.3, where $\mathbf{x}$ is an arbitrary point in volume V, and $\mathbf{x}_0$ is a point on the enclosing surface ∂V. If the pressure $S(\mathbf{x}_0)$ and velocity $\frac{\partial}{\partial \mathbf{n}} S(\mathbf{x}_0)$ caused by the primary source, S, are known at every point on the enclosing surface, the pressure caused by S at $\mathbf{x}$ can be calculated from Eq. (1.3).

Green's function in three dimensions is:

$$G(\mathbf{x}|\mathbf{x}_0) = \frac{1}{4\pi} \frac{e^{j\omega/c\|\mathbf{x}-\mathbf{x}_0\|_2}}{\|\mathbf{x}-\mathbf{x}_0\|_2} + F, \tag{1.4}$$

where ω is the angular frequency of the primary source and F can be any function which satisfies the wave equation. In the conventional form of the KH integral, F is equal to zero. This Green's function and its differential can be realized using monopole and dipole loudspeaker patterns respectively. Therefore, the desired field can be recreated in volume V by placing co-located monopole and dipole loudspeakers at all points on its surrounding surface. Based on Eq. (1.3), the amplitude of the monopole loudspeaker at $\mathbf{x}_0$ is equal to the velocity of the primary source at this point $\frac{\partial}{\partial \mathbf{n}} S(\mathbf{x}_0)$, while the amplitude of the dipole is equal to the pressure at this point, $S(\mathbf{x}_0)$.

The second term in the integral, $S(\mathbf{x}_0) \frac{\partial}{\partial \mathbf{n}} G(\mathbf{x}|\mathbf{x}_0)$, can be eliminated if the gradient of Green's function is zero on the enclosing surface of the volume. This type of Green's

function is called Neumann Green's function [16], denoted by G_N, and is characterized by:

$$\frac{\partial G_N(\mathbf{x}|\mathbf{x}_0)}{\partial \mathbf{n}}\Big|_{\mathbf{x}_0 \in \partial V} = 0. \tag{1.5}$$

Finding the Neumann Green's function is hard in general, especially for complex geometries. A large source-free volume can be imagined as a half space. In this case, the surrounding surface of that volume is a plane as shown in Fig. 1.4. For this planar contour, the Neumann Green's function is given by:

$$G_N(\mathbf{x}|\mathbf{x}_0) = 2 \cdot G(\mathbf{x}|\mathbf{x}_0) = 2 \cdot \frac{1}{4\pi} \frac{e^{i\omega/c\|\mathbf{x}-\mathbf{x}_0\|_2}}{\|\mathbf{x}-\mathbf{x}_0\|_2}. \tag{1.6}$$

In Fig. 1.4, $\mathbf{x}$ is an arbitrary point in volume V, and $\mathbf{x}'$ is the mirrored point of $\mathbf{x}$ with respect to the dividing plane. The pressure caused by G is equal at $\mathbf{x}$ and $\mathbf{x}'$, so $\frac{\partial}{\partial \mathbf{n}} G(\mathbf{x}|\mathbf{x}_0)$ is equal to zero in this structure. For the planar contour, the KH integral converts to a Rayleigh I integral:

$$P(\mathbf{x}) = -\int_{\partial V} (G(\mathbf{x}|\mathbf{x}_0) \frac{\partial}{\partial \mathbf{n}} S(\mathbf{x}_0) dS_0. \tag{1.7}$$

Hence, in this configuration, the sound pressure at each point of the half space is determined if the velocity of all points on the dividing plane is known. For example, considering the Cartesian coordinates, given the velocity of all points of $x-y$ plane, the pressure of all points with $z > 0$ is calculated by Eq. (1.7). This implies that using a continuous planar distribution of loudspeakers, the sound field can be reconstructed at every point of the half space correctly, without any error [16].

On the other hand, if in this geometry, G is selected such that it is equal to zero on the surface of the dividing plane, then the first term of Eq. (1.3) vanishes, and the pressure of each point with $z > 0$ is calculated by a Rayleigh II integral as follow:

$$P(\mathbf{x}) = -\int_{\partial V} -S(\mathbf{x}_0) \frac{\partial}{\partial \mathbf{n}} G(\mathbf{x}|\mathbf{x}_0) dS_0. \tag{1.8}$$

To find G such that its value is zero on $x-y$ plane, function F should be equal to:

$$F = -\frac{e^{i\omega/c\|\mathbf{x}'-\mathbf{x}_0\|_2}}{\|\mathbf{x}'-\mathbf{x}_0\|_2}, \tag{1.9}$$

where $\mathbf{x}'$ is obtained by mirroring $\mathbf{x}$ with respect to the $x-y$ plane.

This integral states that any desired sound field can be reproduced without error for points $z > 0$ using a continuous distribution of dipoles on the $x-y$ plane. The driving function of each dipole is equal to the pressure of the primary source at the dipole location.

Sound field reproduction with the WFS method faces two major practical problems:

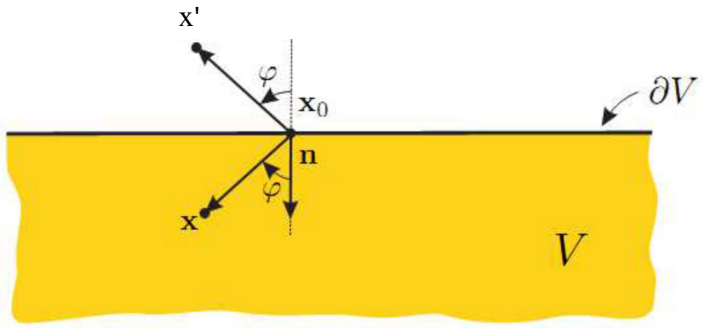


Figure 1.4: Wave Field Synthesis in half space [16].

- To generate a practical discrete loudspeaker distribution, the continuous distribution should be sampled with a rate above the Nyquist rate [17]. That is, the spacing between adjacent loudspeakers should be less than a half a wavelength for a plane wave reproduction in free space.

- Idealized WFS requires that the loudspeakers be placed on the entire dividing plane, so the required number of loudspeakers, even if discretized, is infinite. Of course, this is also not practical, and for any real application one must truncate the loudspeaker array.

Truncation and discretization (with spatial undersampling) lead to reproduction error in the practical versions of WFS.

The main advantage of WFS relative to other SFR methods (discussed below) is that, in principle, it allows the dimensions of the listening area to be very large [17]. The WFS literature is now reviewed.

The concept of wave field synthesis is introduced in [19]. In [42], based on the concept of WFS, the driving function of a continuous linear array of loudspeakers, with any arbitrary shape, is calculated for SFR on a linear listening area, with any arbitrary shape, both for non-focused and focused primary sources. The effect of truncation and discritization are also investigated in [42]. In addition, the sound field is reproduced by the WFS method on a planar listening area using two orthogonal linear arrays of loudspeakers [42].

In [25], the WFS method is used for SFR in a reverberant room. A new method is proposed based on wave domain adaptive filtering for room compensation.

Another method is also proposed in [20] for SFR by employing WFS method in a reverberant room. This method is called Adaptive WFS (AWFS) which is a combination of WFS and active noise cancellation techniques in order to treat the reverberation effect in sound field reproduction.

In [21], it is shown that the performance of the WFS method is improved only if the loudspeakers which have the most significant effect on the sound field reproduction are

active. In other words, this method selects the loudspeakers which should be active in the WFS method for SFR.

The authors of [26] have presented a 3-D SFR system which works based on the WFS method with a small number of loudspeakers. The theory of the WFS method is revisited [16] for 2-D and 3-D SFR system with any arbitrary shape of the LR.

Ahrens and Spors have presented a new SFR method in [17, 35, 37, 43, 44] which works in the wave-domain. In this method, the reproduced sound field is written as a spatial convolution of the transfer function of the loudspeakers and their driving function. The convolution is then transformed into a multiplication in the wave domain, i.e., the desired (reproduced) field is equal to driving function of the loudspeakers multiplied by their transfer function. Expressed as an inverse filter, the driving function of the loudspeakers is the desired field divided by the transfer function in the wave domain. A closed form for the driving function is derived by taking the inverse spatial Fourier transform of the driving function in the wave domain. This SFR technique is called Spectrum Division Method (SDM).

The challenges of the SDM method are similar to those of the WFS method. In this method in contrast to the WFS method, the patterns can be selected arbitrarily, and they are not necessarily monopole or dipole. Furthermore, in SDM, the LR can be of any shape, and it is not necessarily to be the enclosing surface of the listening area. However, the configurations for which SDM is applicable are limited, and these configurations face the same issues as in the WFS method. The reasons are two-fold: first, calculating the wave domain representation of the transfer function is not easy for any arbitrary loudspeaker. Second, having a closed form of the inverse Fourier transform of the obtained driving function is not easy for any arbitrary configuration. Therefore, to have a closed form solution, as in the WFS method, the loudspeakers are assumed to be continuous, and they are placed either on an infinite plane or line in the Cartesian coordinate, or on a sphere or around a circle in the spherical coordinate. This implies that descritization and truncation are also necessary in this method, and they are sources of error,

In [43], a closed form for the excitation function of a linear array of dipole loudspeakers is obtained by SDM assuming that the listening area is another line.

SDM is investigated in [17] for the monopole distribution of loudspeakers for 2-D and 3-D plane wave reproduction. The effect of truncation and discretization is investigated, and the maximum distance between loudspeakers is calculated [17]. Based on this analysis, it is revealed that a large number of loudspeakers should be employed for 3-D SFR using a planar array of loudspeakers. As an example, to reproduce a plane wave with a frequency of 800 Hz at broadside, the required number of loudspeakers on a 3 m $\times$ 3 m planar LR is 196. For this reason, the authors in [17] have employed a linear array of loudspeakers to reproduce a desired sound field along another line as the listening area. The number

of loudspeakers is set to 20, and the primary source is assumed to be a plane wave with a specific direction at 1000 Hz.

In [45], the SDM method is employed for 2-D SFR using a circular loudspeaker array with a non-omnidirectional radiation pattern. The 3-D version of this configuration is presented in [46].

1.4.2 Higher Order Ambisonics

In the Higher Order Ambisonics (HOA) method, the desired sound field is first decomposed in terms of the spherical harmonic functions [29] (following the notation of [29]):

$$p^{\text{des}}(\mathbf{x}) = \sum_{l=0}^{\infty} \sum_{m_d=-l}^{l} B_{m_d l}(x;k) Y_{m_d}^{l}(\vec{\mathbf{x}}), \tag{1.10}$$

In this equation, $x = \|\mathbf{x}\|_2$, $\vec{\mathbf{x}} = \mathbf{x}/x$, $Y_{m_d}^{l}$ is the spherical harmonic function of order l and degree m_d, and $B_{m_d l}(x;k)$ is the harmonic coefficient which is independent of the angular information of point $\mathbf{x}$. The spherical harmonic functions are:

$$Y_{m_d}^{l}(\vec{\mathbf{x}}) = \sqrt{\frac{(2l+1)(l-|m_d|)!}{4\pi(l+|m_d|)!}} P_{l,|m_d|} cos(\theta) e^{i m_d \phi} \tag{1.11}$$

where P is the Legendre function, and θ and ϕ are the elevation and azimuth angles to point $\mathbf{x}$. The harmonic coefficients corresponding to a plane wave in Eq. (1.2), and a point source in Eq. (1.1) are given by [47]:

$$B_{m_d l}^{pl}(x;k) = 4\pi i^l j_l(kx) (Y_{m_d}^{l}(\vec{\mathbf{k}}))^* \tag{1.12}$$

$$B_{m_d l}^{ps}(x;k) = -4\pi i k j_l(kx) x_0 e^{-ikx_0} h_l(kx_0) (Y_{m_d}^{l}(\vec{\mathbf{x_0}}))^* \tag{1.13}$$

where j_l and h_l are spherical Bessel and Hankel functions of order l respectively, $x_0 = \|\mathbf{x_0}\|_2$, and $\vec{\mathbf{x}_0}$ is the unit vector in the direction of point $\mathbf{x}_0$.

This first step in HOA, i.e. decomposing the sound field into spherical harmonics, is sometimes referred to as "spatial encoding" [29] in acoustics. After this step, the sound field is approximated by a limited number of spherical harmonics as:

$$p^{\text{des}}(\mathbf{x}) \cong \sum_{l=0}^{L_t} \sum_{m_d=-l}^{m_d=l} B_{m_d l}(x;k) Y_{m_d}^{l}(\vec{\mathbf{x}}), \tag{1.14}$$

where L_t is called the order of truncation. Then, the reproduced field by the loudspeakers is expanded in terms of the spherical harmonic functions, and this expansion in truncated up to the same order of truncation. Let N be the number of loudspeakers, the recreated

field by the loudspeaker array is:

$$\begin{aligned} p(\mathbf{x}) = \sum_{n=0}^{N} s_n G'_n(\mathbf{x}|\mathbf{x}_n) &= \sum_{n=0}^{N} s_n \left(\sum_{l=0}^{\infty} \sum_{m_d=-l}^{m_d=l} C^n_{m_d l}(x;k) Y^l_{m_d}(\vec{\mathbf{x}}) \right) \\ &\cong \sum_{l=0}^{L_t} \sum_{m_d=-l}^{l} \left(\sum_{n=0}^{N} s_n C^n_{m_d l}(x;k) \right) Y^l_{m_d}(\vec{\mathbf{x}}), \end{aligned} \tag{1.15}$$

where s_n is the complex amplitude (excitation) of the n-th loudspeaker, $G'_n(\mathbf{x}|\mathbf{x}_n)$ is the transfer function of n-th loudspeakers to point $\mathbf{x}$ which is located at point $\mathbf{x}_n$, and $C^n_{m_d l}(x;k)$ is the harmonic coefficient of G'_n. Comparing Eqs. (1.14) and (1.15) results in the following system of equations with N unknown parameters and $(L_t+1)^2$ equations:

$$\sum_{n=0}^{N} s_n C^n_{m_d l}(x;k) = B_{m_d l}(x;k) \qquad \text{for} \quad 0 \le l \le L \quad -l \le m_d \le l \tag{1.16}$$

Therefore, the complex amplitudes of loudspeakers are derived by solving this system of equations which is called"spatial decoding" [29]. The above system can be solved exactly for:

$$N > (L_t+1)^2. \tag{1.17}$$

Therefore, for correct reproduction, the number of loudspeakers should be equal to or greater than $(L_t+1)^2$ [29].

The reproduction error in the HOA method comes from the approximation of the desired field in Eq. (1.14) by a limited number of spherical harmonics:

$$\text{Error} = \sum_{l=L_t+1}^{\infty} \sum_{m_d=-l}^{m_d=l} B_{m_d l}(x;k) Y^l_{m_d}(\vec{\mathbf{x}}) \tag{1.18}$$

This implies that the reproduction error decreases by increasing the order of truncation, i.e., by increasing the number of loudspeakers. Depending on the desired field, the order of truncation to achieve a specific reproduction error is different. For example, to achieve less than 4% reproduction error for reproducing a plane wave with a frequency of f inside a sphere with a radius of r, the truncation order is found by [29]:

$$L_t > \lceil kr \rceil, \tag{1.19}$$

According to Eqs. (1.17) and (1.19) by increasing the frequency and size of the listening area, the number of loudspeakers increases exponentially. In other words, for a fixed number of loudspeakers and order of truncation, the size of the listening area decreases at higher frequencies for plane wave reconstruction.

Generally, with a reasonable number of loudspeakers, the HOA method works accurately for small listening areas [29] in contrast to the WFS method which is applicable for large listening areas with larger reproduction error.

The SFR literature which has employed the HOA method for SFR is reviewed in the following.

In [29], HOA method is used for plane wave reproduction. Here, 25 loudspeakers are placed on a sphere with a radius of 1 m surrounding the listening area, which is a smaller sphere of radius 0.2 m. The placement of loudspeakers is determined by sphere packing. The results are presented only for reconstruction of a plane wave with a specific direction of propagation and a frequency of 1 kHz. The number of loudspeakers is selected such that the reproduction error is less than 4% for the selected frequency and listening area.

In [32], the HOA method is re-explained and compared with the WFS method. In this paper, 32 loudspeakers are located on the perimeter of a circle, and the listening area is inside of a circle of radius 1.5 m. The reproduced fields resulting from the HOA and WFS methods, along with the regions from the listening area for which the error is less than 5% and 20%, are depicted graphically. According to the presented results, the HOA method is more accurate than WFS. The same authors presented different aspects of the HOA method in more detail for 2-D and 3-D reconstruction by employing circular and spherical loudspeaker arrays in [31].

In [36], a closed form for the excitation function of loudspeakers is derived analytically via the HOA method for a circular array of loudspeakers. In this method, 56 loudspeakers are located around a circle with a radius of 1.5 m, and the desired field is a plane wave with a specific direction of propagation and a frequency of 1000 Hz.

The theory of recording and reproducing a sound field by the HOA method in 3-D free space and a reverberant room is explained in [48]. Here, loudspeakers are arranged on a surface of a sphere and the desired field originates from a virtual point source.

In [49], a continuous distribution of loudspeakers is considered on the perimeter of a circle, and a closed form for the excitation function of the continuous loudspeakers is derived analytically through the HOA method. Then, the calculated continuous driving function is sampled to obtain the complex amplitudes of the discrete loudspeakers.

The method in [30] presents a structure for plane wave reproduction in a sphere using three circular arrays (rings) with a total of 18 loudspeakers by the HOA method. The radius of the largest ring is 2 m. The results are presented for reproduction of the various plane waves with different frequencies and directions of propagation. In this method, the error changes between zero percent to 200 percent with the size of the listening area and the direction of propagation of the desired plane wave. For example, when the radius of the listening area (sphere) is $r = 0.5$ and the direction of propagation is $(\theta = \pi, \phi = \pi/4)$ the error is 10%, while for $(\theta = \pi/2, \phi = \pi/4)$, the error is 100% and it is 70% for $(\theta = 3\pi/4, \phi = \pi/4)$.

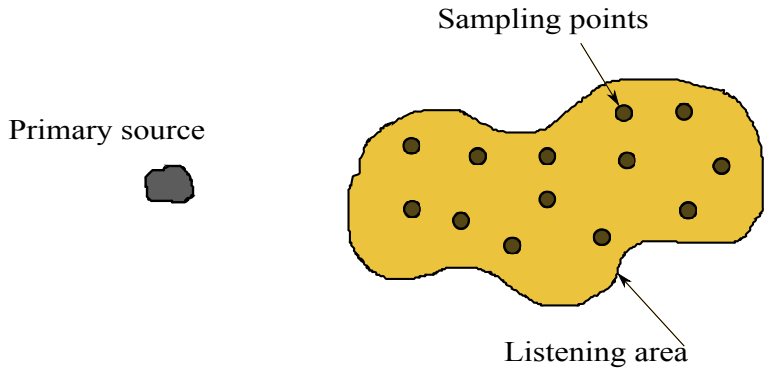


Figure 1.5: Sampling points in the direct optimization method.

Multi-zone SFR in 2-D by using the HOA method is presented in [50]. In this method, loudspeakers are placed on the perimeter of a circle which contains all zones.

1.4.3 Direct Approximation (pressure matching)

This method is called direct approximation because the error metrics can be treated directly. Rather than attempting to exactly create the desired field, which is usually not possible because of practical constraints, direct approximation methods attempt to minimize the reproduction error at some discrete points of the listening area. These points are called sampling points and they are shown in Fig. 1.5. The Nyquist distance between the sampling points depends on the frequency and the distance between the listening area and primary source. For instance, for plane wave reproduction, the distance between the sampling points should be equal or less than half of the wavelength $\lambda = C/f$ [51]. However, for spherical wave reproduction, the Nyquist distance in meter between the sampling points decreases by decreasing the distance between the primary source and listening area [51].

The closed-form solutions will not always be available in direct approximation-based SFR, so numerical methods are deployed for SFR. The main advantage of this method is that it is applicable to any arbitrary system configuration.

This method tries to drive the error signal at sampling points down to zero. In contrast to the previously described methods, different parameters and conditions can be explicitly considered here, and this method can be formulated as a so-called multi-objective optimization problem. Let N be the number of loudspeakers, and vector $\mathbf{s} = [s_1, ..., s_N]^T$ be the complex amplitudes of loudspeakers. In general the optimum value for vector $\mathbf{s}$ is found by solving the optimization problem:

$$\mathbf{s}^{\text{opt}} = \begin{cases} \underset{\mathbf{s};h(\mathbf{s})\leq h_0}{\arg\min} & F_1(\mathbf{s}) \\ \vdots & \\ \underset{\mathbf{s};h(\mathbf{s})\leq h_0}{\arg\min} & F_{M_c}(\mathbf{s}) \end{cases} \tag{1.20}$$

where h and F_j's can be any arbitrary function of $\mathbf{s}$, and M_c is the number of functions which are considered in this method. In the SFR literature, the complex amplitudes of loudspeakers are determined so that the energy of the reproduction error is minimized while keeping the ℓ_η-norm of the loudspeakers less than a predetermined value. Therefore, $M_c = 1$, $F_1 = \|\mathbf{e}\|_2^2$, where $\mathbf{e}$ is the error vector at sampling points, and $h(.)$ is equal to the ℓ_η-norm of the loudspeakers.

Let the desired pressure (phase and amplitude of a sinusoidal wave) at M sampling points be arranged in an $M \times 1$ vector $\mathbf{p}^{\text{des}}$, and the reproduced pressures at the sampling points by the loudspeaker array be arranged in another $M \times 1$ vector $\mathbf{p}(\mathbf{s})$, which is a function of the complex amplitudes of the loudspeakers. The direct approach attempts to solve the following problem of finding the optimum values for the complex amplitudes of the loudspeakers:

$$\mathbf{s}^{\text{opt}} = \underset{\|\mathbf{s}\|_\eta^2 \leq p_{\eta_{\max}}}{\arg\min} \quad \|\mathbf{p}_2^{\text{des}} - \mathbf{p}(\mathbf{s})\|_2^2, \tag{1.21}$$

where $\|\mathbf{s}\|_\eta$ is the ℓ_η-norm of the loudspeakers, and $p_{\eta_{\max}}$ is the maximum value of ℓ_η-norm. This problem is also known as "Least Square" (LS) optimization when $\eta = 2$. The ℓ_η-norm of the loudspeakers for $\eta \geq 1$ is calculated by:

$$\|\mathbf{s}\|_\eta = \left(\sum_{n=1}^{N} |s_n|^\eta \right)^{\frac{1}{\eta}}. \tag{1.22}$$

The ℓ_0-norm of a vector is defined as the number of non-zero elements of that vector.

Eq. (1.21) can be solved by the Lagrangian method with the following cost function:

$$J = \left\| \mathbf{p}^{\text{des}} - \mathbf{p}(\mathbf{s}) \right\|_2^2 + \gamma \|\mathbf{s}\|_\eta^2 \tag{1.23}$$

where γ is the regularization parameter. The solution of this minimization problem leads to sparse representation of $\mathbf{s}$ for $\eta = 1$. With this condition, the error is minimized while the complex amplitudes of the specified number of loudspeakers are close to zero.

For $\eta = 2$, $\|\mathbf{s}\|_2^2$ is proportional to the total power of the loudspeaker array. The maximum value of ℓ_2-norm is denoted by $p_{\max}$ and it is referred to as "maximum normalized power". To find the complex amplitudes of the loudspeakers, the direct approximation method is employed by limiting the ℓ_2-norm of the loudspeakers in this thesis.

In the direct approximation approach, the normalized reproduction error at the sampling point is calculated as:

$$\text{Error} = \frac{\left\| \mathbf{p}^{\text{des}} - \mathbf{p}(\mathbf{s}^{\text{opt}}) \right\|_2^2}{\|\mathbf{p}^{\text{des}}\|_2^2}. \tag{1.24}$$

The normalized reproduction error depends on the desired field, SFR configuration, operating frequency, the number of loudspeakers, the number of sampling points, and the maximum normalized power. As mentioned earlier, by increasing the frequency, the Nyquist

distance in meter between the sampling points decreases, and the number of sampling points increases. Therefore, with a reasonable number of loudspeakers, by increasing frequency or increasing the size of the listening area, the system equation in Eq. (1.21) becomes overdetermined which means that the error cannot be zero in general. In summary, as in the HOA method, the direct approximation methods work precisely for small listening areas while, in contrast to the HOA method, they can be applied to any arbitrary geometry, and other conditions can be considered during the optimization process. In the following paragraphs, the literature which use the direct approximation method for sound field reproduction is reviewed.

The Direct approximation method is used in [52] for SFR in three different structures. Here, the cost function is only the ℓ_2 norm of the reproduction error. In all structures, a desired plane wave is reproduced on a 0.5 m$\times$0.5 m square, and the number of sampling points is 64. In the first structure, two loudspeakers are placed in front of the listening area, and the angle between the two-loudspeaker array and the listening area is 60 degrees. In the second structure, four loudspeaker are placed on the vertices of a square, and the listening area is at the center of that square. In a third structure, four loudspeakers are placed on a sector with angle of 90 degrees, and the listening area is at the center of that sector. The normalized reproduction error versus the frequency of the plane wave and the direction of the plane wave is illustrated in this paper for frequencies less than 1000 Hz.

Gauthier and Berry [38] employed the direct approximation approach to recreate a plane wave in a square region. In this work, 51 loudspeakers are placed on perimeter of a 7.5 m $\times$ 6.4 m rectangle, and the listening area is a square with side length of 1 m in the center of this rectangle. The operating frequency of a single tone primary source is less than 1500 Hz, and the results are presented in the context of recreating plane waves with different directions of propagation. The energy of the reproduction error is minimized using the least squares (LS) solution while the total normalized power of the loudspeakers (more specifically, $\|\mathbf{s}\|_2^2$) is constrained.

Arrangement and number of sampling points are crucial parameters in the SFR using the direct approximation methods. This problem is addressed in [53] by analyzing the desired field, which is also called the planacoustic function(PAF), on the listening area.

In [15], Lilis *et al.* employed the "Least absolute shrinkage and selection operator" (Lasso) to recreate a desired field. In contrast to the LS approach from [38], in Lasso, the number of loudspeakers (ℓ_1-norm) is constrained in order to promote sparsity. The operating frequency in this paper is less than 1400 Hz, and the results are presented for a 2-D listening area. The desired field in this paper is a spherical wave, which originates from a specific point. Two different configurations were studied. In the first one, 49 loudspeakers are placed around a circle, and the region of interest is a square with side lengths of 3 m, located at the center of the circle. The second configuration has an amphitheater topology and the number of loudspeakers is 90. Dimensions of the listening area is about 4 m $\times$ 9 m.

The results show that Lasso outperforms LS in these two scenarios in terms of reproduction error for the same maximum normalized power.

The optimization problem formulated in Eq. (1.21) can be solved using a Lagrangian approach [54–57] by converting the constrained problem in Eq. (1.21) to an unconstrained problem of minimizing Eq. (1.23) . Coefficient γ is known as a Lagrange multiplier or a regularization factor. In [39], two new approaches are proposed to find the best regularization factor for the LS-based sound field reproduction. The first approach uses SVD, while the second one is based on an iterative Newton method. In the scenario studied in [39], the desired field is a spherical wave. Four circular arrays of loudspeakers are located around a circle with a radius of 1.5 m. The radius of each circular array is 0.2 m and each contained 8 monopole loudspeakers, for a total of 32 loudspeakers. The listening area is a 2-D circle whose radius is 0.5 m. The operating frequency in this paper is 750 Hz.

1.4.4 Optimization of the loudspeaker placement

Depending on the SFR formulation and cost function, the reproduction error can be made a convex function of the excitation vector, which simplifies the solution. However, the reproduction error cannot be a convex function of loudspeaker locations [15], which makes globally optimal loudspeaker placement challenging. Perhaps for this reason, loudspeaker placement for SFR has been less studied, and the research has mostly focused on finding the excitation vector after loudspeaker locations are chosen ad-hoc, e.g. uniform, and spaced using a free space sampling criteria.

To optimize the loudspeaker locations, the reproduced field is considered as a function of the locations and complex amplitudes of the loudspeakers. Using the formulation in the direct approximation method, the following non-convex optimization problem should be solved for finding the loudspeaker locations constrained by their LR

$$(\mathbf{s}^{\text{opt}}, \mathbf{x}_1^{\text{opt}}, \mathbf{x}_2^{\text{opt}}, .., \mathbf{x}_N^{\text{opt}}) = \underset{\mathbf{s},\mathbf{x}_n}{\arg\min} \ \ \|\mathbf{p}^{\text{des}} - \mathbf{p}(\mathbf{s}, \mathbf{x}_1, .., \mathbf{x}_N)\|_2^2 \quad s.t. \quad \mathbf{x}_n \in \text{LR}, \tag{1.25}$$

where N is the number of loudspeakers and LR is the loudspeaker region. In the above formulation, the reproduced field is considered as a continuous function of the loudspeaker locations. The solution of this problem may result in an impractically small distance between the loudspeakers. To account for this, the loudspeaker region is uniformly sampled such that the distance between samples is in accordance with the practical size of loudspeakers. Then, the above optimization problem selects the optimum locations from these samples from the LR which are referred to as candidate locations.

In practice, the loudspeaker locations should be optimized as static DoFs before the system operation because it is infeasible to change these parameters during the system operation. Therefore, in this thesis and other literature, the loudspeaker locations are

determined provided that some parameters of the primary source(s) such as its frequency range and its possible locations are given.

In [58], the number and locations of loudspeakers are optimized for SFR without any information regarding the primary source except for the operating frequency. As explained above, optimum locations are selected from the candidate locations form the LR. This method optimizes the loudspeaker locations iteratively only based on the system configuration (loudspeaker region and listening area). In this method, either the location of the first loudspeaker is given, or it can be selected from the candidate locations if the desired field is known. With the location of the first loudspeaker as an initial condition, this algorithm finds the locations of the other loudspeakers from the candidate locations one by one, such that the selected acoustic impedance vectors are linearly independent.

Given the desired field, loudspeaker placement for SFR with a view towards reducing the reproduction error is studied in [15]. In this method, densely spaced candidate loudspeakers are placed on the LR, and the excitation vector is computed to minimize the reproduction error while limiting its ℓ_1 norm. This method results in a sparse linear approximation of the desired field in terms of the Acoustic Transfer Functions (ATFs) of the loudspeakers. This minimization problem is solved in [15] by Lasso. Since this technique results in a sparse solution, the excitations of a considerable number of loudspeakers are zero, therefore, they can be removed from the array. In [59], the same strategy is employed to optimize the locations of loudspeakers for multi-zone SFR.

1.4.5 Optimization of the loudspeaker pattern

Optimizing the loudspeaker patterns as a static or dynamic DoFs is another SFR technique. The pressure at an arbitrary point is given by the complex amplitudes and the associated transfer functions of the loudspeakers. The transfer function of a loudspeaker is the multiplication of the Green's function by the radiation pattern of that loudspeaker. Again, using the formulation in the direct approximation method, the radiation patterns of the loudspeakers are found through the following optimization problem:

$$(\mathbf{s}^{\text{opt}}, \mathcal{L}_1^{\text{opt}}, \mathcal{L}_2^{\text{opt}}, .., \mathcal{L}_N^{\text{opt}}) = \underset{\mathbf{s}, \mathcal{L}_n}{\arg\min} \ \ \|\mathbf{p}^{\text{des}} - \mathbf{p}(\mathbf{s}, \mathcal{L}_1, .., \mathcal{L}_N)\|_2^2 \tag{1.26}$$

where $\mathcal{L}$ is the radiation pattern of loudspeakers.

The radiation patterns can be optimized through the HOA method by replacing the transfer function G' in Eq. (1.15) with $G_n(\mathbf{x}|\mathbf{x}_n)\mathcal{L}_n(\mathbf{x}|\mathbf{x}_n)$, and considering $\mathcal{L}$ as unknown functions.

Any radiation pattern can be expanded in terms of the spherical harmonic function. This expansion can be approximated by a limited number of spherical harmonics. Therefore, the problem of optimizing the loudspeaker patterns can be simplified to optimizing the harmonic coefficients of a specific number of spherical harmonic functions. With this simplification,

in both formulations, HOA and direct approximation, the optimization problem is convex in terms of the harmonic coefficients of the loudspeakers.

Furthermore, the harmonic coefficients of the loudspeakers can be optimized either as dynamic DoFs or as static DoFs. Considering these harmonic coefficients as dynamic DoFs is equivalent to replacing one loudspeaker by a number loudspeakers whose radiation patterns are equal to the spherical harmonic functions. It means that the complex amplitudes of all loudspeakers (spherical harmonic functions) should be updated during the system operation. It increases the system accuracy at the expense of increasing the system operational complexity. Considering these parameters as static DoFs improves the system performance, not as much as the previous case, while it does not increase the dynamic system complexity. In this case, the radiation patterns of the loudspeakers are designed before the system operation, and only their complex amplitudes update during the system operation. In the following paragraphs, the literature that optimizes the loudspeaker radiation patterns is reviewed.

In [60], the radiation patterns of the loudspeakers are a fixed linear combination of a monopole and a radially-directed dipole, while the same author presented another method in [61] which determines the harmonic coefficients of monopole and radially-directed dipole based on the desired field in real time and during system operation. Therefore, in the first case [60], the radiation patterns of the loudspeakers are regarded as static DoFs while they are dynamic in [61].

In [62] and [63], the radiation patterns of higher order loudspeakers are optimized for 2-D and 3-D SFR based on the HOA method. In [62], the higher order loudspeakers are on the perimeter of a circle while in [63] they are on the surface of a sphere. In both methods the obtained radiation patterns vary during system operation, s o they are regarded as dynamic DoFs.

1.5 Contribution and thesis outline

As mentioned earlier, some system parameters in SFR are determined before the system operation as static DoFs. These parameters are found intuitively, or they are optimized if some parameters of the primary source(s) are given. The focus of this thesis is on optimization of the system parameters, specifically the locations and radiation patterns of the loudspeakers, before the system operation.

To optimize the loudspeaker locations, we will introduce two methods in Chapter 2. In the first method, the ATF matrix of a benchmark array of uniformly placed loudspeakers undergoes Singular Value Decomposition (SVD). This creates an “ideal” ATF matrix for minimizing reproduction error and under the power constraint. The “ideal” ATF matrix is then best approximated by selecting the appropriate locations for the loudspeakers. This method will be referred to as *SVD-based placement* algorithm.

The second method utilizes constrained matching pursuit (CMP), an extension of matching pursuit [64], to optimize the loudspeaker locations under a power constraint. At each iteration, the goal is to select the locations whose ATF is the most correlated with the current reproduction error while constraining power. This method will be referred to as *CMP-based placement* in this thesis. Chapter 3 compares the performance of these two algorithms against the existing methods of [58] and [15] for 3-D SFR using a planar array of loudspeakers in a cubic listening area.

To optimize the radiation patterns of the loudspeakers, in Chapter 4, a new algorithm is presented for finding the harmonic coefficients of higher order loudspeakers. In this algorithm, the higher order radiation patterns are optimized before the system operation. It uses Constrained Matching Pursuit for the given frequency range and possible locations of the primary source. Therefore, in contrast to the methods in [62, 63], the performance of the proposed SFR method improves without increasing the dynamic system complexity. In addition, a new method which jointly optimizes both radiation patterns and locations of the loudspeakers is introduced in this chapter. The performance of this method is compared against different systems in which only pattern or placement is optimized.

This optimized SFR system is employed in the acoustic layer of an immersive communication system in Chapter 5. In the system configuration, the speech sound field is captured at one end by a microphone array working with a video monitoring system. At the other end the captured speech is recreated with the pattern selection algorithm. The joint optimization algorithm is deployed before the system operation to design the radiation pattern and location of the loudspeakers.

During the system operation, the Head Related Transfer Function (HRTF)-based reproduction error is considered in calculating the complex amplitudes of the loudspeakers. With this strategy the localization cues are preserved after sound field reproduction which is to improves the sense of immersion. To measure the sense of immersion, in addition to HRTF-based reproduction error, two important parameters in localizations, the Interaural Time Difference (ITD) and Interaural Level Difference (ILD), are logged.

Finally, Chapter 6 concludes the thesis and presents the future steps in order to improve the performance of the proposed SFR algorithms, and consequently the quality of the immersive communication system.

This PhD research study resulted in the following publications:

Conference Papers:

1. H. Khalilian, I. V. Bajić, and R. G. Vaughan, Towards optimal loudspeaker placement for sound field reproduction, In Proc. IEEE ICASSP'13, Vancouver, BC., May 2013, pp 321-325.

2. H. Khalilian, I. V. Bajić, and R. G. Vaughan. Loudspeaker placement for sound field reproduction by constrained matching pursuit. In Proc. IEEE Workshop on

Applications of Signal Processing to Audio and Acoustics (WASPAAâĂŹ13), New Paltz, NY, Oct. 2013, pp 1-4.

3. H. Khalilian, I. V. Bajić, and R. G. Vaughan. 3-D sound field reproduction using diverse loudspeaker patterns. In IEEE International Conference on Multimedia and Expo Workshops, San Jose, CA, Jul. 2013, pp 1-4.

4. H. Khalilian, I. V. Bajić, and R. G. Vaughan. Joint Joint optimization of loudspeaker placement and radiation patterns for Sound Field Reproduction. In Proc. IEEE ICASSP'15, South Brisbane, QLD, April 2015, pp. 519-523.

5. H. Khalilian, I. V. Bajić, and R. G. Vaughan. A Glimpse of 3-D Acoustics for Immersive Communication. In The 29th Annual IEEE Canadian conference on electrical and computer engineering, Vancouver, BC., May 2016.

6. H. Khalilian, I. V. Bajić, and R. G. Vaughan. Learning to Reproduce a Sound Field, Accepted for presentation at IEEE International Workshop on Machine Learning for Signal Processing, Salerno, Italy, Sep. 2016.

Journal Papers:

1. H. Khalilian, I. V. Bajić and R. G. Vaughan, Comparison of Loudspeaker Placement Methods for Sound Field Reproduction, in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 8, pp. 1364-1379, Aug. 2016.

2. H. Khalilian, I. V. Bajić and R. G. Vaughan, Sound Field Reproduction for Immersive Communication , Submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing.

In addition to the above thesis papers, the following contributions are also part of my PhD study:

Conference Paper:

- H. Khalilian and I. V. Bajić, Multiplicative video watermarking with semi-blind maximum likelihood decoding for copyright protection, In IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PacRim), Victoria, BC., Aug. 2011, pp 125-130.

Journal Paper:

- H. Khalilian and I. V. Bajić, Video Watermarking With Empirical PCA-Based Decoding, in IEEE Transactions on Image Processing, vol. 22, no. 12, pp. 4825-4840, Dec. 2013.

Chapter 2

Loudspeaker Placement

2.1 Introduction

In this chapter, we will present two algorithms to optimize the loudspeaker locations on a given LR. In the first algorithm, we first compute the Acoustic Transfer Function (ATF) matrix of the benchmark configuration- uniform array of omni-directional loudspeakers on the LR. The Singular Value Decomposition (SVD) of this ATF matrix is then calculated, and the reproduction error and the total normalized power of loudspeakers are expressed in terms of the SVD matrices. The SVD matrices are then manipulated to reduce the reproduction error and the total normalized power. A new ("ideal") ATF matrix is formed based on the manipulated SVD matrices, which may be unrealizable. After this, the loudspeaker locations are adjusted in order to best approximate the "ideal" ATF matrix with realizable forms. The SVD approach relates to MIMO wireless communications where the parallel, independent, singular channels maximize the mutual information (*c.f.*, minimal transmission error for a fixed rate) between transmitting and receiving arrays. When all the available power is directed through only the channel with the highest singular value, it is called dominant mode MIMO. The analogy here is that maximizing the mutual information between the loudspeakers and the sampling points offers a strong starting point for being able to minimize the reproduction error, and the ideal ATF matrix is found based on this approach.

For the second algorithm, we develop the Constrained Matching Pursuit (CMP) algorithm to find the location of loudspeakers. This algorithm is an extension of matching pursuit [64], and here, it optimizes loudspeaker locations under a power constraint. Power constraints exist in all practical SFR systems, since no real system can draw infinite power. Both CMP and SVD-based placements take power constraint into account in the process of selecting loudspeaker locations. At each iteration, the goal is to select the location of each loudspeaker whose ATF is the most correlated with the current reproduction error, while taking the power constraint into account.

This chapter is organized as follows. Section 2.2 lays out the preliminaries and terminology of SFR based on direct approximation. The GS-based [58] and Lasso-based [15] placement methods, which are the existing placement methods in the SFR literature, are explained in Sections 2.3 and 2.4 respectively. SVD-based loudspeaker placement is presented in Section 2.5. First, the procedure of finding the "ideal" ATF matrix is explained in Section 2.5.1. Sections 2.5.2 and 2.5.3 present the ideal ATF matrix from the mathematical and SVD points of view respectively. The realizability of the ideal ATF matrix is investigated in Section 2.5.4 followed by the placement algorithm in Section 2.5.5. CMP-based placement method is then presented in Section 2.6. The matching pursuit algorithm and its constrained version are explained in Sections 2.6.1 and 2.6.2. The CMP algorithm is mathematically analyzed in Section 2.6.3, and the CMP-based placement algorithm is proposed in Section 2.6.4. Section 2.7 concludes this chapter. The performance of these four placement algorithms will be compared in Chapter 3 for SFR by a planar loudspeaker array in a cubic listening area and by a circular array of loudspeakers in a 2-D square listening area.

The contributions of this chapter are:

- Introducing a new placement method based on the SVD analysis of the ATF matrix.
- Analyzing the ideal ATF matrix.
- Developing a Constrained Matching Pursuit (CMP) algorithm by modifying the matching pursuit algorithm to work under power limitation.
- Introducing a new placement algorithm based on a Constrained Matching Pursuit algorithm which considers power constraints in finding the loudspeakers' locations.

2.2 Preliminaries

Let N be the number of loudspeakers on the LR and $M > N$ be the the number of sampling points in the region of interest (listening area). The pressure created by the n-th loudspeaker at the m-th sampling point at time t is given by

$$p(\mathbf{y}_m|\mathbf{x}_n, t) = s_n(t) * g'(\mathbf{y}_m|\mathbf{x}_n, t), \tag{2.1}$$

where $\mathbf{x}_n$ and $\mathbf{y}_m$ are the Cartesian coordinates of the n-th loudspeaker and the m-th sampling point, $*$ is the time-domain convolution, $s_n(t)$ is the excitation of the n-th loudspeaker, and $g'(\mathbf{x}_n|\mathbf{y}_m, t)$ is the acoustic impulse response of the n-th loudspeaker at the m-th sampling point. The corresponding relationship in the frequency domain is

$$P(\mathbf{y}_m|\mathbf{x}_n, f) = s_n(f) \cdot G'(\mathbf{y}_m|\mathbf{x}_n, f) \tag{2.2}$$

where $s_n(f)$ (it is not capitalized to fit the existing notation in the literature) is the Fourier transform of $s_n(t)$ and $G'(\mathbf{y}_m|\mathbf{x}_n, f)$ is the Fourier transform of $g'(\mathbf{y}_m|\mathbf{x}_n, t)$, that is, the Acoustic Transfer Function (ATF) of the n-th loudspeaker at the m-th sampling point. The outward traveling wave of an omni-directional point-source loudspeaker in free space can be expressed as:

$$G'(\mathbf{y}_m|\mathbf{x}_n, f) = \frac{e^{ik\|\mathbf{x}_n - \mathbf{y}_m\|_2}}{4\pi\|\mathbf{x}_n - \mathbf{y}_m\|_2} \cdot \mathcal{L}(\mathbf{y}_m|\mathbf{x}_n, f) = G(\mathbf{y}_m|\mathbf{x}_n, f) \cdot 1. \tag{2.3}$$

where $i = \sqrt{-1}$, $k = 2\pi/\lambda$ is the wave number, $\lambda = C/f$ is the wavelength, C is the speed of sound, $G(\mathbf{y}_m|\mathbf{x}_n, f)$ is the Green's function, $\mathcal{L}(\mathbf{y}_m|\mathbf{x}_n, f)$ is the complex gain of the radiation pattern of the loudspeaker at sampling point. $\mathcal{L}(\mathbf{y}_m|\mathbf{x}_n, f)$ is equal to 1 for an omni-directional source. This convention is adopted in this thesis, and the link equation is explained in Appendix C.

Consider a single frequency f and let $\mathbf{s}(f) = [s_1(f), s_2(f), .., s_N(f)]^T$ be the vector of loudspeakers' complex excitations at that frequency. Let $\mathbf{p}^{\text{des}}(f)$ and $\mathbf{p}(f)$ be the M-dimensional vectors containing samples of the desired and reproduced sound field, respectively, at the M sampling points. Following the convention in the SFR literature [15, 38], the frequency f is dropped from the notation, and the vectors are written $\mathbf{s}$, $\mathbf{p}^{\text{des}}$ and $\mathbf{p}$. The acoustic transfer functions of N omni-directional loudspeakers at M sampling points are arranged in ATF matrix $\mathbf{G}$:

$$\mathbf{G} = \begin{bmatrix} G'(\mathbf{y}_1|\mathbf{x}_1) & G'(\mathbf{y}_1|\mathbf{x}_2) & \cdots & G'(\mathbf{y}_1|\mathbf{x}_N) \\ G'(\mathbf{y}_2|\mathbf{x}_1) & G'(\mathbf{y}_2|\mathbf{x}_2) & \cdots & G'(\mathbf{y}_2|\mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ G'(\mathbf{y}_M|\mathbf{x}_1) & G'(\mathbf{y}_M|\mathbf{x}_2) & \cdots & G'(\mathbf{y}_M|\mathbf{x}_N) \end{bmatrix}.$$

The field reproduced at the M sampling points by N loudspeakers is given by the ATF matrix multiplied by the complex amplitude vectors $\mathbf{s}$:

$$\mathbf{p} = \begin{bmatrix} G'(\mathbf{y}_1|\mathbf{x}_1) & G'(\mathbf{y}_1|\mathbf{x}_2) & \cdots & G'(\mathbf{y}_1|\mathbf{x}_N) \\ G'(\mathbf{y}_2|\mathbf{x}_1) & G'(\mathbf{y}_2|\mathbf{x}_2) & \cdots & G'(\mathbf{y}_2|\mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ G'(\mathbf{y}_M|\mathbf{x}_1) & G'(\mathbf{y}_M|\mathbf{x}_2) & \cdots & G'(\mathbf{y}_M|\mathbf{x}_N) \end{bmatrix} \cdot \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{bmatrix} = \mathbf{G}\mathbf{s}, \tag{2.4}$$

The power of the n-th omni-directional loudspeaker is given by [65]:

$$P_n = \frac{|s_n|^2}{8\pi\rho_0 C}, \tag{2.5}$$

where ρ_0 is the air density. The total power of N loudspeakers is

$$\sum_{n=1}^{N} P_n = \frac{1}{8\pi\rho_0 C} \cdot \sum_{n=1}^{N} |s_n|^2 = \frac{1}{8\pi\rho_0 C} \cdot \|\mathbf{s}\|_2^2 \quad \text{watts.} \tag{2.6}$$

Hence, since the total power is proportional to $\|\mathbf{s}\|_2^2$, the quantity $\|\mathbf{s}\|_2^2$ is referred to as the "normalized power" in this thesis.

Given the loudspeaker locations, the Least Squares (LS) method, which is a direct approximation method, tries to optimize the complex amplitudes of the loudspeakers by solving the following optimization problem:

$$\mathbf{s}^{\text{opt}} = \underset{\|\mathbf{s}\|_2^2 \leq p_{\max}}{\arg\min} \quad \|\mathbf{p}^{\text{des}} - \mathbf{G}\mathbf{s}\|_2^2, \tag{2.7}$$

where $p_{\max}$ is the maximum normalized power. This problem can be solved by the Lagrangian multiplier as:

$$J = \|\mathbf{p}^{\text{des}} - \mathbf{G} \cdot \mathbf{s}\|_2^2 + \gamma \|\mathbf{s}\|_2^2. \tag{2.8}$$

where γ is the regularization parameter, and it is determined such that the normalized power of the loudspeaker is limited to $p_{\max}$ [39]. The solution of Eq. (2.8) is [38]:

$$\mathbf{s}^{\text{opt}} = (\mathbf{G}^H \cdot \mathbf{G} + \gamma \mathbf{I})^{-1} \cdot \mathbf{G}^H \cdot \mathbf{p}^{\text{des}}. \tag{2.9}$$

In order to optimize the loudspeakers' locations, first the loudspeaker region is densely sampled at N_v candidate locations. Then, the loudspeakers' locations are selected from these candidate locations by different placement methods. Let $\mathbf{G}_v$ be the ATF matrix from the candidate locations to the sampling points, and $\mathbf{s}_v$ be the corresponding complex amplitude vector to the candidate loudspeakers. The problem being addressed for loudspeaker placement is striving for the optimum loudspeaker locations, $\mathbf{x}_n$'s, from the non-convex optimization:

$$\min_{\mathbf{x}_n, \mathbf{s}_v} \|\mathbf{p}^{\text{des}} - \mathbf{G}_v \mathbf{s}_v\|_2^2 \quad s.t. \quad \|\mathbf{s}_v\|_0 = N \quad , \quad \|\mathbf{s}_v\|_2^2 \leq p_{\max}. \tag{2.10}$$

This problem is solved in [15] without considering the power constraint (the second condition in Eq. (2.10)) and by replacing the ℓ_0-norm in Eq. (2.10) with the ℓ_1-norm. This replacement makes the problem convex, which is the basis of the Lasso-based placement from [15]. Therefore, although Lasso is a convenient suboptimal solution in the absence of the power constraint, it does not address the power constraint. The CMP- and SVD- based methods are greedy algorithms to account for the ℓ_0-norm. Power constraint is taken into account in the CMP-based method directly, while it is considered indirectly in the SVD-based method within the computation of one of its parameters. The placement method from [58] finds the loudspeakers' locations from the candidate locations only based on the

system configuration and without considering power limitation and solving Eq. (2.10) by Gram-Schmidt orthogolization. These algorithms are explained in the following sections.

2.3 GS-based placement

This placement algorithm works based on the Gram-Schmidt orthogonalization of the ATF vectors. Hence, the inputs of this algorithm are only the ATF vectors, and the desired vector is not taken into account for finding the loudspeakers' locations. In this algorithm, the first ATF is selected arbitrarily, and other ATFs are selected iteratively such that the selected vectors form an orthogonal basis. This method is summarized in Algorithm 1. Since the desired vector does not play any role in optimizing the loudspeakers' locations, it is expected that this algorithm has worse performance when compared with other placement algorithms in which the desired field is considered. However, since the ATFs vectors are selected to form an orthogonal space, it is expected that the selected ATF matrix is a well conditioned matrix, and this algorithm results in a robust solution for SFR.

In this algorithm, after that the loudspeakers' locations are found by the GS-based method, the LS method, from Eq. (2.9), is employed to find the complex amplitudes of the loudspeakers.

Algorithm 1 GS-based placement

Input: N_v ▷ number of candidate locations

Input: $\{\mathbf{x}_n\}_{n=1}^{N_v}$ ▷ candidate loudspeaker locations

Input: N ▷ allowed number of loudspeakers

Input: M ▷ number of sampling points

Input: $\{\mathbf{y}_m\}_{m=1}^{M}$ ▷ locations of sampling points

Output: A^{gs} ▷ selected locations of loudspeakers

1: Set $A^{\text{gs}} = \emptyset$.

2: Using Eq. (2.3), calculate the ATF of each candidate source ($\mathbf{x}_n$) at M sampling points ($\mathbf{y}_m$). Denote these ATF vectors $\mathbf{h}_1$, $\mathbf{h}_2$,..., $\mathbf{h}_{N_v}$, where $\mathbf{h}_l$ is the ATF of the l-th candidate location $\mathbf{x}_l$, at M sampling points.

3: Initialize $n = 1$, $\mathbf{h}^{(1)} = \mathbf{h}_1$, $\widehat{\mathbf{h}^{(1)}} = \mathbf{h}_1/\|\mathbf{h}_1\|_2$, $A^{\text{gs}} = A^{\text{gs}} \cup \mathbf{x}_1$ where $\mathbf{h}^{(1)}$ is the initial selected ATF vector.

4: **for** $n = 2$ to N **do**

5: For all candidate locations, calculate $\mathbf{r}_j = \mathbf{h}_j - \mathbf{p}$, where $\mathbf{p} = \sum_{l=1}^{n-1} \mathbf{p}_l$, and $\mathbf{p}_l =< \left(\widehat{\mathbf{h}^{(l)}}\right)^H \mathbf{h}_j > \widehat{\mathbf{h}^{(l)}}$ is the projection of the j-th candidate ATF on the previously selected ATFs.

6: Find index j^* that maximizes $\|\mathbf{r}_j\|_2$.

7: Remove the selected location from the search list.

8: Let $\mathbf{h}^{(n)} = \mathbf{h}_{j^*}$, $\widehat{\mathbf{h}^{(n)}} = \mathbf{r}_{j^*}/\|\mathbf{r}_{j^*}\|_2$, $A^{\text{gs}} = A^{\text{gs}} \cup \{\mathbf{x}_{j^*}\}$.

9: **end for**

10: **return** A^{gs}

2.4 Lasso-based placement

In the SFR approach presented in [15], the loudspeaker excitation vector is calculated through the following optimization problem:

$$\mathbf{s}^{\text{sparse}} = \underset{\mathbf{s}_v \in \mathbb{C}^{N_v}}{\arg\min} \left[\|\mathbf{p}^{\text{des}} - \mathbf{G}_v \mathbf{s}_v\|_2^2 + \gamma^{\text{lasso}} \|\mathbf{s}_v\|_1 \right]. \tag{2.11}$$

Comparing the cost function in Eq. (2.11) to the least squares cost function in Eq. (2.8), it is seen that the ℓ_1-norm of the excitation vector is involved, whereas the LS in Eq. (2.8) involves ℓ_2 norm of the excitation vector. The result is that the solution of Eq. (2.11) is sparser than the LS solution in Eq. (2.9). The optimization problem in Eq. (2.11) can be solved by Lasso, shown in Algorithm 2. In [15], it is shown experimentally that in a specific, under-sampled sound field, the Lasso-based complex excitation vector resulting from Eq. (2.11) leads to better SFR performance in comparison with those obtained from Eq. (2.9). Since the number of required sampling points increases with frequency [53], for a fixed size of listening area, the Lasso-based solution should be better in general than the

LS-based solution as the frequency increases, if the number of sampling points M remains fixed.

Algorithm 2 Lasso Solver [15]

1: Initialize $\mathbf{s}^{n_{it}} = \mathbf{0}_N$, where $\mathbf{s}^{n_{it}}$ is the optimized vector at n_{it}-th iteration.
2: **for** $n_{it} = 1$ to N_{it} **do**
3: **for** $j = 1$ to N **do**
4: Calculate error vector corresponding to n_{it}-th iteration and n-th entry as:

$$\mathbf{e}_j^{n_{it}} = \mathbf{p}^{\text{des}} - \sum_{n=1}^{j-1} \mathbf{g}_n s_n^{n_{it}} - \sum_{n=j+1}^{N} \mathbf{g}_n s_n^{n_{it}-1}, \tag{2.12}$$

where $\mathbf{g}_n$ is the n-th column of $\mathbf{G}$.
5: Find each entry of $\mathbf{s}^{n_{it}}$ through:

$$s_j^{n_{it}} = \left(\|\mathbf{g}_j\|_2^2\right)^{-1} \left(\mathbf{g}_j^H \mathbf{e}_j^{n_{it}} - \gamma^{\text{lasso}}\right)_+ e^{i\angle(\mathbf{g}_j^H \mathbf{e}_j^{n_{it}})} \tag{2.13}$$

where $(x)_+$ is equal to x if $x > 0$ otherwise it is zero, and γ^{lasso} is the regularization parameter.
6: **end for**
7: **end for**
8: **return** $\mathbf{s}^{\text{sparse}} = \mathbf{s}^{N_{it}}$

The focus of [15] is on finding the complex excitation vector from Eq. (2.11) and comparing the results with the LS solution from Eq. (2.9). However, the authors also proposed an algorithm for judicious loudspeaker placement by Lasso. For a single-tone primary source, this procedure is summarized in Algorithm 3. The regularization factor γ^{Lasso} in Eq. (2.11) is determined by the cross validation method from [15].

Algorithm 3 Lasso-based placement

Input: N_v ▷ number of candidate locations

Input: $\{\mathbf{x}_n\}_{n=1}^{N_v}$ ▷ candidate loudspeaker locations

Input: N ▷ allowed number of loudspeakers

Input: M ▷ number of sampling points

Input: $\{\mathbf{y}_m\}_{m=1}^{M}$ ▷ locations of sampling points

Output: A^{lasso} ▷ selected locations of loudspeakers

1: Set $A^{\text{lasso}} = \emptyset$.

2: Using Eq. (2.3), calculate the ATF of each candidate source ($\mathbf{x}_n$) at M sampling points ($\mathbf{y}_m$). Denote these ATF vectors $\mathbf{h}_1$, $\mathbf{h}_2$,..., $\mathbf{h}_{N_v}$, where $\mathbf{h}_l$ is the ATF of the l-th candidate location $\mathbf{x}_l$, at M sampling points.

3: Using the Lasso solver in Algorithm 2 [15,66], find the complex excitation vector $\mathbf{s}^{\text{sparse}}$ from Eq. (2.11) with exactly N non-zero coefficients. This is an iterative procedure with N_{it} iterations.

4: $A^{\text{lasso}} = \{\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, ..., \mathbf{x}_{(N)}\}$, where $\mathbf{x}_{(j)}$ is the location corresponding to the j-th non-zero element in $\mathbf{s}^{\text{sparse}}$.

5: **return** A^{lasso}

In this thesis, to ensure that the number of non-zero coefficients is exactly equal to N, the regularization factor γ^{lasso} in Eq. (2.11) is determined by a systematic search in the range $[0, \|\mathbf{G}^H\mathbf{p}^{\text{des}}\|_\infty)$ with a step size of $10^{-3} \cdot \|\mathbf{G}^H\mathbf{p}^{\text{des}}\|_\infty$. If the number of non-zero entries is equal to N, the corresponding candidate γ^{lasso} is selected as the final regularization parameter. Otherwise, if the number of non-zero entries is not equal to N for any candidate γ^{lasso}, then the candidate γ^{lasso} that results in the closest number (from above) of non-zero entries to N is selected as the final regularization parameter, and the N largest entries are selected as the non-zero entries, the others being set to zero. Algorithm 3 is only used to find suitable locations for the loudspeakers. Once these locations are obtained, the ATF matrix $\mathbf{G}$ is constructed, and the excitation vector is obtained using Eq. (2.9).

2.5 SVD-based loudspeaker placement

2.5.1 The ideal ATF matrix

Let $\mathbf{G}$ be the ATF matrix of the benchmark configuration. The LS solution from Eq. (2.9) is expressed in terms of the Singular Value Decomposition (SVD) matrices of the ATF matrix in [39]. The SVD of $\mathbf{G}$ is denoted

$$\mathbf{G} = \mathbf{U}^g \cdot \mathbf{\Sigma}^g \cdot \mathbf{V}^{gH}, \tag{2.14}$$

where $\mathbf{U}^g$ and $\mathbf{V}^g$ are, respectively, $M \times M$ and $N \times N$ unitary matrices, and $\mathbf{\Sigma}^g$ is an $M \times N$ diagonal matrix containing singular values of $\mathbf{G}$ arranged in a decreasing order. In Eq. (2.14), the superscript $(\cdot)^H$ denotes the conjugate transpose. By substituting Eq. (2.14) into Eq. (2.9), the solution $\mathbf{s}^{\text{opt}}$ from Eq. (2.9) can be expressed in terms of the elements of SVD matrices as follows:

$$\mathbf{s}^{\text{opt}} = \sum_{n=1}^{N} \frac{\sigma_n^g}{(\sigma_n^g)^2 + \gamma} c_n^g \mathbf{v}_n^g. \tag{2.15}$$

In this equation, σ_n^g is the n-th singular value of $\mathbf{G}$, i.e., also the n-th diagonal element of $\mathbf{\Sigma}^g$, $c_n^g = \mathbf{u}_n^{g\,H} \mathbf{p}^{\text{des}}$ is the projection of the desired field vector onto the n-th column of $\mathbf{U}^g$, denoted by $\mathbf{u}_n^g$, and $\mathbf{v}_n^g$ is the n-th column of $\mathbf{V}^g$.

Further, the total normalized power of the loudspeaker array can also be expressed in terms of the elements of SVD matrices [39]:

$$\|\mathbf{s}^{\text{opt}}\|_2^2 = \sum_{n=1}^{N} \frac{(\sigma_n^g)^2}{((\sigma_n^g)^2 + \gamma)^2} |c_n^g|^2. \tag{2.16}$$

Therefore, the total normalized loudspeaker power is proportional to $|c_n^g|^2$ and decreasing with γ. In addition, the squared ℓ_2-norm of the reproduction error can be expressed as:

$$\|\mathbf{p}^{\text{des}} - \mathbf{G} \cdot \mathbf{s}^{\text{opt}}\|_2^2 = \sum_{n=1}^{N} \frac{\gamma^2}{((\sigma_n^g)^2 + \gamma)^2} |c_n^g|^2 + \sum_{m=N+1}^{M} |c_m^g|^2, \tag{2.17}$$

which means that the ℓ_2-norm of the error is increasing with $|c_n^g|$ and γ. Hence, by changing the regularization factor γ there is a trade-off between the total normalized power and the error norm. However, both of these parameters, power and error, are monotonic with $|c_n^g|$, meaning that if $|c_n^g|$ increases, so do the power and error, and vice versa. Therefore, by reducing magnitudes of c_n^g's, both the power and the reproduction error can be reduced.

In the SVD-based loudspeaker placement method, the magnitudes of c_n^g's are decreased in order to improve the performance of the SFR system. As mentioned before, $|c_n^g|$'s are the projections of the desired field onto the columns of $\mathbf{U}^g$. Since we have no control over the desired field, in order to change c_n^g's, we have to change the elements of $\mathbf{U}$. After manipulation, the new matrix will be called $\mathbf{U}^{\text{ideal}}$. Since $\mathbf{U}^g$ is a unitary matrix, its columns form an orthonormal basis for an M-dimensional vector space. In addition, we have the following inequalities:

$$0 \leq |c_n^g|^2 = |(\mathbf{u}_n^g)^H \mathbf{p}^{\text{des}}|^2 \leq \|\mathbf{u}_n^g\|_2^2 \cdot \|\mathbf{p}^{\text{des}}\|_2^2 = \|\mathbf{p}^{\text{des}}\|_2^2. \tag{2.18}$$

The minimum value of $|c_n^g|^2$ is 0, and it is achieved when $\mathbf{u}_n^g$ is orthogonal to the desired vector $\mathbf{p}^{\text{des}}$. The maximum value of this parameter is $\|\mathbf{p}^{\text{des}}\|_2^2$, and it is achieved when $\mathbf{u}_n^g$ is parallel with $\mathbf{p}^{\text{des}}$.

Suppose we select one of the columns of $\mathbf{U}^{\text{ideal}}$, say the r-th column, as a unit vector parallel with $\mathbf{p}^{\text{des}}$:

$$\mathbf{u}_r^{\text{ideal}} = \frac{\mathbf{p}^{\text{des}}}{\|\mathbf{p}^{\text{des}}\|_2}. \tag{2.19}$$

The other $M-1$ columns of $\mathbf{U}^{\text{ideal}}$ need to be orthogonal to $\mathbf{u}_r^{\text{ideal}}$, so they will lie in the $(M-1)$-dimensional null space of $\mathbf{u}_r^{\text{ideal}}$:

$$\mathbf{u}_q^{\text{ideal}} \in \text{Null}\{\mathbf{u}_r^{\text{ideal}}\} \quad \text{for } q \neq r. \tag{2.20}$$

With such a choice of $\mathbf{U}^{\text{ideal}}$, all terms, except $n = r$, in the summations in Eqs. (2.16) and (2.17) reduce to zero, because the projections of $\mathbf{p}^{\text{des}}$ onto the corresponding columns of $\mathbf{U}^{\text{ideal}}$ are zero. The total normalized power and reproduction error become:

$$\|\mathbf{s}\|_2^2 = \begin{cases} \frac{(\sigma_r^g)^2}{((\sigma_r^g)^2+\gamma)^2}\|\mathbf{p}^{\text{des}}\|_2^2 & \text{if } r \leq N, \\ 0 & \text{if } r > N. \end{cases} \tag{2.21}$$

$$\|\mathbf{p}^{des} - \mathbf{p}\|_2^2 = \begin{cases} \frac{\gamma^2}{((\sigma_r^g)^2+\gamma)^2}\|\mathbf{p}^{\text{des}}\|_2^2 & \text{if } r \leq N, \\ \|\mathbf{p}^{\text{des}}\|_2^2 & \text{if } r > N. \end{cases} \tag{2.22}$$

Now the question is – what is the best value for r? According to Eqs. (2.21) and (2.22), if $r > N$, then the total normalized power is zero (all loudspeakers are turned off), and the reproduction error is equal to the ℓ_2-norm of $\mathbf{p}^{\text{des}}$. That is clearly not a good choice for r. In order to simultaneously reduce the error and power, r should be no greater than N, and since the diagonal elements of $\mathbf{\Sigma}$ are arranged in descending order ($\sigma_1^g \geq \sigma_2^g \geq ... \geq \sigma_N^g$) the best value for r is 1. Therefore, we select the first column of $\mathbf{U}^{\text{ideal}}$ to be parallel with $\mathbf{p}^{\text{des}}$ and the other columns to be unit vectors in the null space of $\mathbf{p}^{\text{des}}$. The new "ideal" ATF matrix is now obtained by combining $\mathbf{U}^{\text{ideal}}$ with the other two SVD matrices. These two SVD matrices, $\mathbf{\Sigma}$ and $\mathbf{V}$, are taken from the benchmark configuration. For benchmark, N omni-directional loudspeakers are placed uniformly on the LR, and the ATF matrix of this configuration is calculated. From Eq. (2.14) the corresponding SVD matrices to the benchmark configuration are computed and they are replaced in the ideal ATF matrix:

$$\mathbf{G}^{\text{ideal}} = \mathbf{U}^{\text{ideal}} \cdot \mathbf{\Sigma}^g \cdot \mathbf{V}^{gH}. \tag{2.23}$$

This "ideal" ATF matrix results in lower reproduction error and normalized power compared to the benchmark ATF matrix, as shown below.

In summary the ideal ATF matrix is calculated through the following steps:

Step 1: Distribute N loudspeakers uniformly on the LR, calculate the resulting ATF matrix, $\mathbf{G}$, and find its SVD $\mathbf{G} = \mathbf{U}^g\mathbf{\Sigma}^g\mathbf{V}^{gH}$.

Step 2: Find the null space of $\mathbf{p}^{\text{des}}$ as follows. SVD is applied to $(\mathbf{p}^{\text{des}})^H$, giving $(\mathbf{p}^{\text{des}})^H = \mathbf{U}^u\mathbf{\Sigma}^u(\mathbf{V}^u)^H$. Columns 2 to M of $\mathbf{V}^u$ form the null space of $(\mathbf{p}^{\text{des}})^H$. This is easily accomplished in MATLAB using the function `null`.

Step 3: Construct $\mathbf{U}^{\text{ideal}}$ as follows. Set its first column to $\mathbf{u}_1^{\text{ideal}} = \mathbf{p}^{\text{des}}/\|\mathbf{p}^{\text{des}}\|_2$ (normalized desired field vector), and the remaining columns to columns 2 to M of $\mathbf{V}^u$ from the previous step.

Step 4: Set the "ideal" ATF matrix $\mathbf{G}^{\text{ideal}}$ equal to $\mathbf{U}^{\text{ideal}}\mathbf{\Sigma}^g\mathbf{V}^{gH}$. It should be noted that matrices $\mathbf{\Sigma}^g$ and $\mathbf{V}^g$ are kept the same as in the initial ATF matrix $\mathbf{G}$, which corresponds to uniformly distributed loudspeakers.

2.5.2 Mathematical interpretation of ideal ATF matrix

Lemma 2.5.1 shows that $\mathbf{G}^{\text{ideal}}$, constructed as above, has the same singular values as $\mathbf{G}$, while its first N left singular vectors are parallel or anti-parallel (parallel and opposite direction) to the columns of $\mathbf{U}^{\text{ideal}}$.

Lemma 2.5.1. *Let* $\mathbf{U}$ *and* $\mathbf{V}$ *be* $M \times M$ *and* $N \times N$ *unitary matrices, respectively, and* $\mathbf{\Sigma}$ *be a* $M \times N$ *diagonal matrix* $\mathbf{\Sigma} = diag(\sigma_1, \sigma_2, ..., \sigma_N)$ *whose diagonal elements are positive and arranged in decreasing order* ($\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_N > 0$). *Let* $\mathbf{G}$ *be multiplication of these three matrices* $\mathbf{G} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$. *Also, let the SVD of* $\mathbf{G}$ *be given by* $\mathbf{G} = \mathbf{U}^g\mathbf{\Sigma}^g\mathbf{V}^g$. *Then* $\mathbf{\Sigma}^g = \mathbf{\Sigma}$ *and for* $1 \leq i \leq N$, $\mathbf{u}_i^g = \pm\mathbf{u}_i$.

Proof. See Appendix A.1. □

Since by construction, $\mathbf{u}_1^{\text{ideal}}$ is parallel to $\mathbf{p}^{\text{des}}$ and other columns of $\mathbf{U}^{\text{ideal}}$ are orthogonal to it, we have by Lemma 2.5.1 that

$$|c_n^{\text{ideal}}|^2 = |(\pm\mathbf{u}_n^{\text{ideal}})^H\mathbf{p}^{\text{des}}|^2 = \begin{cases} \|\mathbf{p}^{\text{des}}\|_2^2 & \text{if } n = 1, \\ 0 & \text{if } n > 1. \end{cases} \tag{2.24}$$

Also by Lemma 2.5.1, the singular values of $\mathbf{G}^{\text{ideal}}$ are the same as those of $\mathbf{G}$. Using these results in Eq. (2.17) and (2.16) shows that for the ideal ATF matrix, the optimized reproduction error and normalized power are given by

$$\|\mathbf{p}^{\text{des}} - \mathbf{p}\|_2^2 = \frac{\gamma^2\|\mathbf{p}^{\text{des}}\|_2^2}{((\sigma_1^g)^2 + \gamma)^2}, \tag{2.25}$$

$$\|\mathbf{s}^{\text{opt}}\|_2^2 = \frac{(\sigma_1^g)^2}{((\sigma_1^g)^2 + \gamma)^2}\|\mathbf{p}^{\text{des}}\|_2^2. \tag{2.26}$$

The main question now is - how do the optimized reproduction error and normalized power of the ideal ATF matrix $\mathbf{G}^{\text{ideal}}$ in Eqs. (2.25) and (2.26) compare to those of the initial ATF matrix $\mathbf{G}$ in Eqs. (2.17) and (2.16)? The following theorems provide the answers and demonstrate the advantage of using $\mathbf{G}^{\text{ideal}}$.

Theorem 2.5.2. *(Reduction of error) At the same level of normalized power, the reproduction error Eq. (2.25) achieved by* $\mathbf{G}^{ideal}$ *is no larger than that achieved by* $\mathbf{G}$ *in Eq. (2.17).*

Proof. See Appendix A.2. □

Theorem 2.5.3. *(Reduction of power) At the same reproduction error, the normalized power in Eq. (2.26) required by* $\mathbf{G}^{ideal}$ *is no larger than that required by* $\mathbf{G}$ *in Eq. (2.16).*

Proof. See Appendix A.2. □

2.5.3 SVD interpretation of the ideal ATF matrix

An SFR system comprises signal processing that operates on the free space links between the loudspeakers and the microphones. Signal processing inputs and outputs are normally considered as voltages, but here these are interpreted as (linearly proportional to) the transmitted complex amplitude, s_n, and the received pressure waves, p_m in free space. The SVD, $\mathbf{G} = \mathbf{U}^g\mathbf{\Sigma}^g(\mathbf{V}^g)^H$, can be depicted as in Fig. 2.1, where the diagonal nature of $\mathbf{\Sigma}^g$ is seen as parallel, independent channels. In multiple input, multiple output (MIMO) radio communications, these are called eigenchannels (or singular channels), and the right and left singular vectors, $\mathbf{U}^g$ and $\mathbf{V}^g$, are implemented as antenna weights, e.g., [67]. Our situation is different to radio MIMO because here the eigenvectors are used simply as mathematical descriptions of the free space channel, and so the singular vectors are not part of the signal processing as in the radio case. The input signals are the N loudspeaker signals. The signal processing description for these is referred to as “complex excitations" because these quantities are proportional to the excitation (voltages) of the loudspeakers, and they are complex because they have magnitude and phase. In the SVD interpretation, the outputs are the pressures created at the M sampling points based on Eq. (2.4). (The signal processing electronics treats these signals as voltages from omnidirectional microphones.) The N inputs are filtered (preprocessed) by matrix $\mathbf{V}^g$, passed through the N independent channels with each channel scaling its corresponding signal by σ_n^g, and then filtered (post-processed) by matrix $\mathbf{U}^g$. The gains of $\mathbf{U}^g$ and $\mathbf{V}^g$ are unity, so the singular values σ_n^g of the ATF matrix can be interpreted as the amplitude gains of the orthogonal spatial channels. To achieve a given reproduction error, a system with higher singular values requires less input power. Power limitation has an important role in SFR systems, and since the upper channels have higher gains ($\sigma_1^g \geq \sigma_2^g \geq ... \geq \sigma_N^g$), power concentration on the upper channels results in less reproduction error under the same power constraint.

To understand how the ATF matrix $\mathbf{G}$ affects the reproduction error in Eq. (2.17), note that the second term in Eq. (2.17), $\sum_{n=N+1}^{M} |c_n^g|^2$, is related to the part of the desired field vector $\mathbf{p}^{\text{des}}$ that falls into the null space of $\mathbf{G}$. Changing the regularization factor γ, and thereby potentially allowing for higher power, does not affect this term. Therefore, the ATF matrix should ideally be such that $c_n = 0$ for $N + 1 \leq n \leq M$. Since c_n is the projection

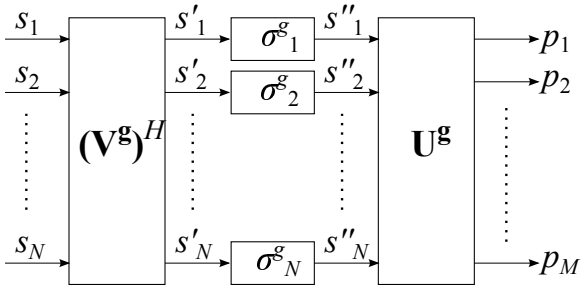


Figure 2.1: Block diagram of an N-input-M-output system describing the free space links.

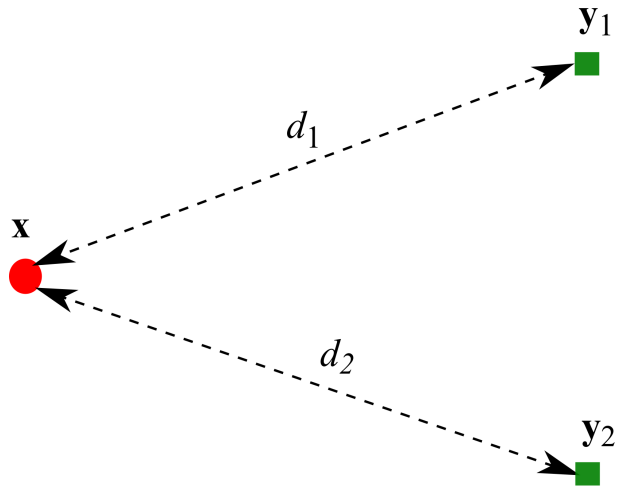


Figure 2.2: Free space sound field reproduction at two points, shown as squares.

of the desired field vector onto the n-th column of $\mathbf{U}^g$, $c_n^g = (\mathbf{u}_n^g)^H \mathbf{p}^{\text{des}}$, the preference is to have the last $M - N$ columns of $\mathbf{U}^g$ to be orthogonal to $\mathbf{p}^{\text{des}}$.

Since the uppermost channel in Fig. 2.1 has the highest gain ($\sigma_1^g \geq \sigma_2^g \geq ... \geq \sigma_N^g$), then from the efficiency point of view, the best use of the ATF matrix is to focus all the signal power on this uppermost channel, since that would require the least input power. This can be achieved if s'_n is equal to zero for $n > 1$. This means that s''_n is equal to zero for $n > 1$, so all elements of $\mathbf{s}''$ at the input to the $\mathbf{U}^g$ block in Fig. 2.1 are zero, except the first one, s''_1. Under this condition, the reproduced sound field vector $\mathbf{p} = \mathbf{U}^{\text{ideal}}\mathbf{s}''$ would be parallel to the first column of $\mathbf{U}^{\text{ideal}}$.

2.5.4 Realizability of the ideal ATF matrix

In this section, through two simple examples, it is explained how the ideal ATF matrix may be realized and why this would not be possible in general to be realizable.

Consider a scenario where the desired field is from a single omni-directional point source, shown as a dot in Fig. 2.2. Suppose the sound field is to be reconstructed at $M = 2$ points, $\mathbf{y}_1$ and $\mathbf{y}_2$, shown as squares in Fig. 2.2 using a single secondary source, i.e. $N = 1$. In

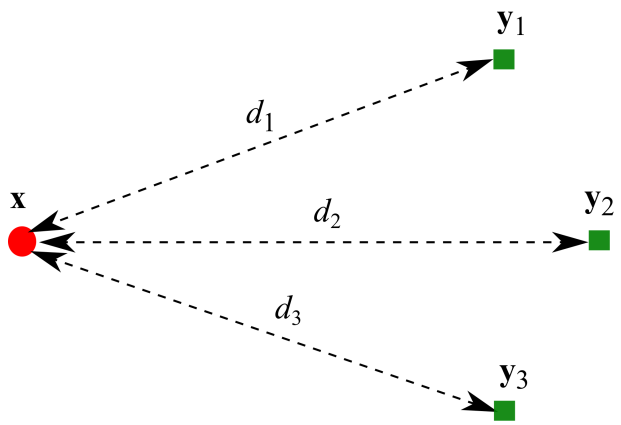


Figure 2.3: Free space sound field reproduction at three points, shown as squares.

this case, $\mathbf{U}^{\text{ideal}}$ is a 2×1 vector parallel with the desired field vector $\mathbf{p}^{\text{des}}$. $\mathbf{G}^{\text{ideal}}$ will be realizable if the secondary source is placed in such a way as to generate the same field as the primary source at $\mathbf{y}_1$ and $\mathbf{y}_2$. The locus of such points in the 3-D space is the intersection of the two spheres centered at $\mathbf{y}_i$, with radii d_i, $i = 1, 2$, respectively, where d_i is the distance from $\mathbf{y}_i$ to the primary source. From 3-D trilateration [68], the intersection is the perimeter of a circle. Therefore, if the secondary source is located at one of these points the ideal ATF matrix would be realizable.

In the second example, suppose the sound field is to be regenerated at $M = 3$ non-collinear points, $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3$, shown as squares in Fig. 2.3 using a single secondary source, i.e., $N = 1$. In this case, $\mathbf{U}^{\text{ideal}}$ is a 3×1 vector parallel with the desired field vector $\mathbf{p}^{\text{des}}$. Again, $\mathbf{G}^{\text{ideal}}$ will be realizable if the secondary source is placed in such a way as to generate the same field as the primary source at $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3$. As the previous example, the locus of such points in the 3-D space is the intersection of the three spheres centered at $\mathbf{y}_i$, with radii d_i, $i = 1, 2, 3$, respectively. The intersection of these spheres is at most two points, one of which is the location of the primary source [68].

Therefore, an "ideal" ATF matrix severely constrains the possible locations of secondary sources, even in the simplest scenarios; resulting location(s) will not be on the LR, except in very special cases. If the number of sampling points M is higher than three, as is usually the case, the only location allowed by the "ideal" ATF is the location of the primary source, which is not on the LR. Hence, in practice, "ideal" ATF is not realizable, so a method is needed to approximate the "ideal" ATF by a physically-realizable ATF. This is discussed in the next section.

2.5.5 SVD-based placement algorithm

The mismatch between the desired field vector $\mathbf{p}^{\text{des}}$, which could be arbitrary, and the constraints on realizable ATF matrices, which depend on the number of loudspeakers, their ra-

diation patterns, and their candidate locations, creates the need to approximate the "ideal" ATF matrix within design constraints. In this chapter, our focus is on omni-directional loudspeakers, whose ATFs are in Eq. (2.3).

From a computational point of view, if the number of loudspeakers was equal to or larger than the number of sampling points of the desired field, i.e. $N \geq M$, then in principle, exact reconstruction of the desired field could be achieved at those M sampling points in the absence of the power constraint, provided $\mathbf{p}^{\text{des}}$ resides in an M-dimensional subspace of span$\{\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_N\}$ where $\mathbf{g}_n$ is the ATF vector of the n-th loudspeaker. Of course, reconstruction between the M sampling points would not necessarily be exact owing to potential spatial aliasing. In practice, power constraint is likely unavoidable and the number of loudspeakers may be smaller than the number of sampling points of the desired field ($N < M$), which means that even at the M points where $\mathbf{p}^{\text{des}}$ is specified, reconstruction error is unavoidable in general. Hence, in order to reduce the reconstruction error, this section presents an algorithm to approximate the "ideal" ATF, which has the minimum reconstruction error under power limitation, by finding suitable locations for N loudspeakers among a candidate set of N_v locations on the LR.

This algorithm works as follows. The first column of the approximated ATF matrix is found as the ATF of the candidate location that is most correlated with the first column of $\mathbf{G}^{\text{ideal}}$. To find the n-th ($n > 1$) column, we select N_c (a function of frequency and power, discussed below) candidate ATFs which have the highest correlation with the n-th column of $\mathbf{G}^{\text{ideal}}$. From these, the candidate ATF whose correlations with all previously selected columns are closest to those of the ideal matrix, is selected as the n-th column. With $N_c = 1$, the algorithm chooses the most correlated vectors with columns of the ideal ATF matrix as the selected ATFs, without taking the relationship among the columns into account. Here, the parameter N_c approximates the correlations among columns of the ideal ATF matrix which are equal to those of the initial ATF matrix. The equality between these correlations implies that increasing N_c leads to more dispersed locations because the corresponding locations from the initial matrix are maximally dispersed. The following paragraphs review why we need to increase N_c and how N_c is selected.

Since $\mathbf{G}^{\text{ideal}} = \mathbf{U}^{\text{ideal}}\mathbf{\Sigma}^{g}\mathbf{V}^{gH}$, the columns of $\mathbf{G}^{\text{ideal}}$ are:

$$\mathbf{g}_n^{\text{ideal}} = \sum_{j=1}^{N} \sigma_j^g v^{g*}_{nj} \mathbf{u}_j^{\text{ideal}}, \tag{2.27}$$

i.e. each column of $\mathbf{G}^{\text{ideal}}$ is a linear combination of the columns of $\mathbf{U}^{\text{ideal}}$, with coefficients given by $\sigma_j^g v^{g*}_{nj}$. It is noted that at low frequencies, the condition number of the initial ATF matrix (benchmark: uniform planar array of N loudspeakers on LR) increases as its columns become more correlated. Now the largest singular value σ_1^g dominates more, and since $\mathbf{G}^{\text{ideal}}$ inherits the same singular values, the columns of $\mathbf{G}^{\text{ideal}}$ also become more

similar. This causes the SVD-based placement method to select closely-spaced locations for the loudspeakers at low frequencies if $N_c = 1$. This strategy is appropriate when $p_{\max}$ is low, because ATFs from neighboring locations are similar, and power emitted from these locations adds up constructively to provide better approximation for the desired sound field. However, when $p_{\max}$ is large, selected locations need not be so closely-spaced, since each loudspeaker can emit sufficient power on its own, without the help from neighboring locations. In this case, it would be beneficial to select farther locations in order to allow the system to approximate the finer structure of the desired field.

To solve this problem, parameter N_c is employed. It acts to increase the distance between loudspeaker locations when frequency is low and $p_{\max}$ is large through the following empirically derived equation:

$$N_c(f, p_{\max}) = \max\left(\lceil F(f) \cdot H(p_{\max}) \rceil, 1\right) \tag{2.28}$$

where

$$F(f) = \begin{cases} 25, & \text{if } f \leq 400 \\ 10, & \text{if } 400 < f \leq 600 \\ 4, & \text{if } 600 < f < 1200 \\ 1, & \text{if } f \geq 1200 \end{cases} \tag{2.29}$$

$$H(p_{\max}) = \min\left(1, \left(\frac{1}{37} \cdot \zeta - \frac{2}{37}\right)_+\right) \tag{2.30}$$

and

$$\zeta = \frac{M \cdot N \cdot p_{\max}}{(4\pi r_{\min})^2 \cdot \|\mathbf{p}^{\text{des}}\|_2^2}. \tag{2.31}$$

In this equation, $r_{\min}$ is the minimum distance between the LR and the listening area, and $(x)_+$ is equal to x if $x > 0$ otherwise it is zero. ζ is a dimensionless quantity, and it is the upper limit of the ratio of the power of the reproduced sampled field to the desired one, viz., $\|\mathbf{p}\|_2^2/\|\mathbf{p}^{\text{des}}\|_2^2$:

$$\begin{aligned} \frac{\|\mathbf{p}\|_2^2}{\|\mathbf{p}^{\text{des}}\|_2^2} = \frac{\sum_m |\sum_n s_n g'_{m,n}|^2}{\|\mathbf{p}^{\text{des}}\|_2^2} &\leq \frac{\sum_m \left(\sum_n |s_n|^2 \sum_n |g'_{m,n}|^2\right)}{\|\mathbf{p}^{\text{des}}\|_2^2} \\ &= \frac{\left(\sum_n |s_n|^2\right)\left(\sum_{n,m} |g'_{m,n}|^2\right)}{\|\mathbf{p}^{\text{des}}\|_2^2} \leq \frac{p_{\max} \cdot M \cdot N}{(4\pi r_{\min})^2 \|\mathbf{p}^{\text{des}}\|_2^2} = \zeta. \end{aligned} \tag{2.32}$$

In this equation, the first inequality is the Cauchy-Schwarz, $p_{\max}$ is the upper bound of the normalized power $\|\mathbf{s}\|_2^2 = \sum_n |s_n|^2$, and $1/(4\pi r_{\min})$ is the maximum of $g'_{m,n}$'s.

The function N_c is robust in the sense that the outcome is not sensitive to the empirical constants or the configuration of LR and the listening area. As $p_{\max}$ increases, ζ increases and makes $H(p_{\max})$ closer to 1. This in turn makes N_c closer to $F(f)$, which is a step-wise

decreasing function of frequency f. Hence, when $p_{\max}$ is large, N_c will be large at low frequencies, and small at high frequencies. On the other hand, as $p_{\max}$ decreases, η also decreases and makes $H(p_{\max})$ closer to 0. This in turn makes $N_c = 1$. Specific parameters in Eqs. (2.28)-(2.30) are found empirically.

Note that the exhaustive search through all possible combinations of loudspeaker positions is infeasible for practical problem sizes. For example, in our simulations, $N = 25$, $N_v = 900$, so $\binom{N_v}{N} > 10^{48}$. Therefore, SVD-based placement is accomplished via Algorithm 4, whose complexity is $O(N \cdot N_v \cdot N_c)$. Since the problem is non-convex and the algorithm does not perform an exhaustive search, the solution is suboptimal in general. Once the loudspeaker locations are selected, and the final ATF matrix $\mathbf{G}$ is constructed, the excitation vector is computed from Eq. (2.9).

Algorithm 4 SVD-based placement

Input: N_v ▷ number of candidate locations

Input: $\{\mathbf{x}_n\}_{n=1}^{N_v}$ ▷ candidate loudspeaker locations

Input: N ▷ allowed number of loudspeakers

Input: N_c ▷ approximation parameter

Input: M ▷ number of sampling points

Input: $\{\mathbf{y}_m\}_{m=1}^{M}$ ▷ locations of sampling points

Output: A^{svd} ▷ selected loudspeaker locations

1: Set $A^{\mathrm{svd}} = \emptyset$.

2: Using Eq. (2.3), calculate the ATF of each candidate source ($\mathbf{x}_n$) at M sampling points ($\mathbf{y}_m$). Denote these ATF vectors $\mathbf{h}_1$, $\mathbf{h}_2$,..., $\mathbf{h}_{N_v}$, where $\mathbf{h}_l$ is the ATF of the l-th candidate location $\mathbf{x}_l$, at M sampling points.

3: **for** $k = 1$ to N **do**

4: **if** $k == 1$ **then**

5: Go to Step 9.

6: **else**

7: Calculate inner products $(\mathbf{g}_k^{\mathrm{ideal}})^H \mathbf{g}_q^{\mathrm{ideal}}$ for $1 \leq q < k$, and arrange the results in a $(k-1)$-dimensional vector $\mathbf{v}$.

8: **end if**

9: Calculate $d(l) = |(\mathbf{g}_k^{\mathrm{ideal}})^H \mathbf{h}_l|$ for $1 \leq l \leq N_v$.

10: **if** $k == 1$ **then**

11: Select index l^* that maximizes $d(l)$.

12: $A^{\mathrm{svd}} = A^{\mathrm{svd}} \cup \{\mathbf{x}_{l^*}\}$.

13: Let $\mathbf{h}^{(k)} = \mathbf{h}_{l^*}$.

14: **else**

15: Find the N_c highest values of $d(l)$.

16: Let $\mathbf{h}_{(j)}, j = 1, 2, ..., N_c$, be the ATFs of the candidate locations found in Step 15. Construct N_c vectors $\mathbf{f}_j$ containing the following inner products:

$$\mathbf{f}_j = \left((\mathbf{h}^{(1)})^H \mathbf{h}_{(j)}, (\mathbf{h}^{(2)})^H \mathbf{h}_{(j)}, ..., (\mathbf{h}^{(k-1)})^H \mathbf{h}_{(j)} \right).$$

17: Find index j^* that minimizes $\|\mathbf{v} - \mathbf{f}_j\|_1$.

18: $A^{\mathrm{svd}} = A^{\mathrm{svd}} \cup \{\mathbf{x}_{j^*}\}$.

19: Let $\mathbf{h}^{(k)}$ be the ATF of $\mathbf{x}_{j^*}$ at M sampling points.

20: **end if**

21: **end for**

22: **return** A^{svd}

2.6 CMP-based placement

Matching Pursuit (MP) is an iterative algorithm to approximate a desired vector as a linear combination of vectors from a given set, known as a dictionary [64]. It has been widely used in signal processing [69–74]. Although MP is applicable in any Hilbert space, our interest here is in the finite-dimensional complex vector space version. In this section, the matching pursuit algorithm from [64] is briefly reviewed and then modified to the Constrained Matching Pursuit algorithm for including power limitation. The CMP algorithm is then deployed to find appropriate locations for the loudspeakers.

2.6.1 Matching Pursuit

Let $\mathbb{C}^M$ be an M dimensional complex vector space, $\mathfrak{D} \subset \mathbb{C}^M$ be a finite dictionary set of unit vectors, and $\mathbf{b}_i \in \mathfrak{D}$ be a dictionary member, with $\|\mathbf{b}_i\|_2 = 1$. Let $\mathbf{a} \in \mathbb{C}^M$ be the vector that needs to be approximated as a linear combination of $N < M$ vectors from $\mathfrak{D}$. The MP algorithm builds up the linear approximation through the following steps:

(i) *Initialization:* Let $n = 1$ and $R^n\mathbf{a} = \mathbf{a}$, where symbol $R^n\mathbf{a}$ represents the approximation error vector at the beginning of the n-th iteration.

(ii) *Selection of a dictionary member:* Find the dictionary member that has the maximum inner product with $R^n\mathbf{a}$:

$$\mathbf{b}^{(n)} = \arg\max_{\mathbf{b}\in\mathfrak{D}} |(\mathbf{b})^H R^n\mathbf{a}|, \tag{2.33}$$

where $\mathbf{b}^{(n)}$ denotes the dictionary member selected at the n-th iteration.

(iii) *Scaling coefficient computation:* Compute the scaling coefficient of the selected dictionary member as

$$\alpha_n = (\mathbf{b}^{(n)})^H R^n\mathbf{a}. \tag{2.34}$$

Therefore, the vector approximated at this iteration, $\mathbf{a}^{(n)}$, is:

$$\mathbf{a}^{(n)} = \left((\mathbf{b}^{(n)})^H R^n\mathbf{a}\right)\mathbf{b}^{(n)} = \alpha_n\mathbf{b}^{(n)}. \tag{2.35}$$

(iv) *Approximation error vector:* Calculate the approximation error vector as

$$R^{n+1}\mathbf{a} = R^n\mathbf{a} - \mathbf{a}^{(n)}. \tag{2.36}$$

This becomes the desired vector to be approximated at the next iteration. Stop if $R^{n+1}\mathbf{a} = \mathbf{0}$ or $n = N$, otherwise increase n by 1 and go to Step (ii).

After N iterations, the linear approximation of the desired vector in terms of the dictionary members is:

$$\mathbf{a} = \sum_{n=1}^{N} \left((\mathbf{b}^{(n)})^H R^n \mathbf{a} \right) \mathbf{b}^{(n)} + R^{N+1}\mathbf{a}, \tag{2.37}$$

where $R^{N+1}\mathbf{a}$ is the approximation error vector.

One of the key results in [64] for MP in a finite dimensional vector space is an upper bound on the ℓ_2 norm of the approximation error vector at the n-th iteration:

$$\|R^n \mathbf{a}\|_2 \leq \|\mathbf{a}\|_2 (1 - I^2)^{n/2}, \tag{2.38}$$

where

$$I = \inf_{\mathbf{a} \in \mathfrak{C}^M} \sup_{\mathbf{b} \in \mathfrak{D}} |\mathbf{a}^H \mathbf{b}| / \|\mathbf{a}\|_2 > 0. \tag{2.39}$$

Hence, approximation error decays exponentially with the number of iterations.

2.6.2 Constrained Matching Pursuit

The MP algorithm outlined above selects at each iteration the vector from the dictionary that is most correlated (has highest inner product) with the current approximation error vector, and assigns it a coefficient (α_n) equal to the inner product. In the absence of any constraint on the scaling coefficients (α_n), this is a good strategy. In our application, dictionary members ($\mathbf{b}_i$) are the ATFs of candidate vector locations at the M sampling points. While selecting the ATFs that are most correlated with the desired field and its approximation residuals is still a good strategy, scaling coefficients must be constrained in case of power limitations.

In the Constrained Matching Pursuit (CMP) algorithm, the coefficients assigned to dictionary members in Step (iii) are limited such that $\sum_{n=1}^{N} |\alpha_n|^2 \leq p_{\max}$, where $p_{\max}$ is the maximum normalized power of the loudspeaker array. For this purpose, Step (ii) and (iii) of the MP algorithm are merged and modified as follows to account for this limitation:

(ii) and (iii) *Selection of a dictionary member and Scaling coefficient computation in CMP:* Dictionary member and the scaling coefficient are selected such that:

$$(\mathbf{b}^{(n)}, \alpha_n) = \underset{\mathbf{b} \in \mathfrak{D}, |\alpha|^2 \leq p_n}{\arg\min} \|\alpha \mathbf{b} - R^n \mathbf{p}^{\text{des}}\|_2^2, \tag{2.40}$$

where $\{p_n\}_{n=1}^{N}$ is a sequence of non-negative numbers selected such that $\sum_{n=1}^{N} p_n = p_{\max}$. If the magnitudes of the dictionary members are equal, the one that is most correlated with the current error vector will be selected, as in MP, and the corresponding coefficient will be selected through the following optimization problem:

$$\alpha_n = \underset{|\alpha|^2 \leq p_n}{\arg\min} \|\alpha \mathbf{b}^{(n)} - R^n \mathbf{a}\|_2^2. \tag{2.41}$$

The solution to the minimization problem in Eq. (2.41) is

$$\alpha_n = \begin{cases} \sqrt{p_n}\frac{(\mathbf{b}^{(n)})^H R^n \mathbf{a}}{|(\mathbf{b}^{(n)})^H R^n \mathbf{a}|} & \text{if } \sqrt{p_n} \leq |(\mathbf{b}^{(n)})^H R^n \mathbf{a}|, \\ (\mathbf{b}^{(n)})^H R^n \mathbf{a} & \text{else.} \end{cases} \tag{2.42}$$

Hence, if the magnitude of the inner product between the selected dictionary member and the approximation error vector is less than $\sqrt{p_n}$, the scaling coefficient will be the same as in the MP, otherwise it is scaled down to have magnitude $\sqrt{p_n}$. The choice of p_n is discussed in the next section.

2.6.3 Mathematical implications of CMP

Owing to different scaling coefficients, CMP produces different approximation error vectors compared to MP. A question arises at this point – how does the approximation error reduce in CMP? The following theorem provides the answer.

Theorem 2.6.1. *(Approximation error in CMP) At the n-th iteration, an upper bound on the ℓ_2 norm of the approximation error in CMP is given by*

$$\|R^{n+1}\mathbf{a}\|_2 \leq \|\mathbf{a}\|_2 \cdot \left(1 - \frac{p_{\min}}{\|\mathbf{a}\|_2^2} I^2\right)^{n/2}, \tag{2.43}$$

where $p_{\min} = \min_n\{p_n\}$ and I is given by Eq. (2.39).

Proof. See Appendix A.3. □

This result shows that the upper bound on the approximation error in CMP also decays exponentially, but the rate of decay is different from (and usually much lower than) that of MP from Eq. (2.38). For each iteration, in order to allow the selected vector to match the current approximation error, its coefficient should be as large as the error vector's magnitude. Since this is not possible in general due to power constraint, we make the coefficients follow an exponential decay, as indicated by the upper bound on the error. Specifically, we select $p_n = Bp_0^n$, where B is selected such that $\sum_{n=1}^{N} p_n = p_{\max}$ and p_0 is given by the base of the exponential term in Eq. (2.43), with $p_{\min} = Bp_0^N$. The value of p_0 is found by numerically solving:

$$p_0 = \left(1 - \frac{p_{\min}}{\|\mathbf{a}\|_2^2} I^2\right) = 1 - p_0^N \frac{p_{\max}(1-p_0)}{p_0 - p_0^{N+1}} \cdot \frac{I^2}{\|\mathbf{a}\|_2^2}. \tag{2.44}$$

When $I \ll 1$, which is the case in all our simulation scenarios, p_0 can be approximated by

$$p_0 \approx \left(1 - \frac{p_{\max}}{N\|\mathbf{a}\|_2^2} I^2\right). \tag{2.45}$$

The value of I should be computed from Eq. (2.39). However, since it is impractical to scan the entire M-dimensional complex vector field $\mathbb{C}^M$, we used the following approach to find I based on Eq. (2.39) for each frequency. Within a ball of radius 10 wavelengths (an arbitrary distance to be far enough from the sampling points so that increasing this makes no difference), centered at the middle of the listening area, possible locations of the primary source are sampled such that the nearest neighbor distance is no larger than half the wavelength. From each of these points, a desired field vector is generated within the listening area. The baseline MP algorithm [64] is then performed with N iterations on this desired field vector. The desired field vector, and all error vectors resulting from the iterations of MP, are used as vector $\mathbf{a}$ in Eq. (2.39). Then, the value of I is computed from Eq. (2.39) by looping through vectors $\mathbf{b}$ in the dictionary. Note that the dictionary is frequency-dependent, so the process needs to be repeated at all frequencies of interest. Then the minimum value of Eq. (2.39) over the desired frequency range is used as I in the expression for p_0 in Eq. (2.45). Therefore, for each system geometry, we need to calculate the value of I before employing the placement method. This value is then kept fixed while the other system parameters, such as frequency and location of the primary source, change.

2.6.4 CMP-based placement algorithm

Selection of loudspeaker locations based on CMP is described in Algorithm 5. Note that Algorithm 5 is only used to select loudspeaker locations. Once the locations are selected, the ATF matrix $\mathbf{G}$ of this array is constructed, and the excitation vector is computed from Eq. (2.9).

Algorithm 5 CMP-based placement

Input: N_v ▷ number of candidate locations

Input: $\{\mathbf{x}_n\}_{n=1}^{N_v}$ ▷ candidate loudspeaker locations

Input: N ▷ allowed number of loudspeakers

Input: M ▷ number of sampling points

Input: $\{\mathbf{y}_m\}_{m=1}^{M}$ ▷ locations of sampling points

Output: A^{cmp} ▷ selected locations of loudspeakers

1: Set $A^{\text{cmp}} = \emptyset$.

2: Using Eq. (2.3), calculate the ATF of each candidate source ($\mathbf{x}_n$) at M sampling points ($\mathbf{y}_m$) and place them in set $\mathfrak{D} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{N_v}\}$, where $\mathbf{h}_l$ is the ATF of the l-th candidate location $\mathbf{x}_l$, at M sampling points. These vectors are considered as dictionary members ($\mathbf{b}_i$'s) in CMP.

3: Initialize $n = 1$, $R^n\mathbf{p}^{\text{des}} = \mathbf{p}^{\text{des}}$ and calculate p_n's form Eq. (2.45) for $1 \leq n \leq N$.

4: **for** $n = 1$ to N **do**

5: **if** $R^n\mathbf{p}^{\text{des}} == \mathbf{0}$ **then**

6: Go to Step 14.

7: **else**

8: Select $\mathbf{h}^{(n)}$ using Eq. (2.33). The corresponding location of the candidate source, $\mathbf{x}_{(n)}$, whose ATF is $\mathbf{h}^{(n)}$, is the location selected for the n-th loudspeaker.

9: $A^{\text{cmp}} = A^{\text{cmp}} \cup \{\mathbf{x}_{(n)}\}$.

10: Calculate the complex coefficient α_n from Eq. (2.42).

11: Calculate the new error vector as

$$R^{n+1}\mathbf{p}^{\text{des}} = R^n\mathbf{p}^{\text{des}} - \alpha_n\mathbf{h}^{(n)}.$$

12: **end if**

13: **end for**

14: **return** A^{cmp}

2.7 Conclusion

This chapter presented two novel methods for optimizing the locations of loudspeakers: SVD-based placement and CMP-based placement. The first method adjusts the locations of the loudspeakers in order to approximate the ideal ATF matrix. The ideal ATF matrix is obtained from the singular value decomposition of the ATF of the benchmark configuration. In this method, based on the maximum normalized power, the ATF matrix is approximated either column by column or by considering the relationship among the columns of the ATF

matrix. Therefore, although the effect of power is not directly considered in this method, it is considered in determining parameter N_c.

The second method seeks the locations of loudspeakers based on the CMP algorithm by taking the maximum normalized power into account. Therefore, in contrast to the SVD-based placement method, power limitation is directly taken into account in this iterative algorithm. In this algorithm the error vector is initialized as the desired vector. At each iteration, the most correlated ATF, from the candidate ones, with the desired vector is selected. The coefficient corresponding to the selected ATF is calculated by considering the power limitation. Then, the difference between the input vector and the reconstructed vector at the current iteration is calculated as error vector which is the input vector of the next iteration. This process continues until N ATFs (locations) are selected.

The existing placement methods, the Lasso-based and the GS-based placement algorithms, were also reviewed in this chapter. The Lasso-based algorithm finds the locations of loudspeakers by minimizing error while the ℓ_1-norm of loudspeaker excitations is limited. This method results in sparse representation of the desired vector in terms of the candidate ATFs. Therefore, the candidate ATFs with non-zero coefficients are selected as loudspeaker locations. The GS-based algorithm finds a number of loudspeaker locations whose ATFs are less correlated. This algorithm differs from other placement algorithms in that it optimizes the loudspeaker locations just based on the system configuration and regardless of any information about the desired vector.

In the next chapter, the performance of all placement algorithms will be compared from different points of view such as error performance, computational complexity, and the reproduced field outside of the listening area. All of these placement algorithms calculate the optimized locations for a single tone primary source. It means that the optimum locations at one frequency is not the optimum ones at other frequencies. In other words, the loudspeaker locations should be recalculated by changing the frequency of the primary source which is not practical. This problem is also studied in the next chapter.

Chapter 3

Comparison among Placement Methods

3.1 Introduction

The performance of the placement algorithms, explained in Chapter 2, is evaluated in this chapter through simulations on two different configurations, 2-D and 3-D. In the 3-D configuration, shown in Fig. 3.1, secondary sources are placed on a square region as a planar array, and the listening area is a cubic region. The desired field is assumed to originate from a primary source which is located somewhere behind the planar array. The 2-D configuration, as shown in Fig. 3.17, is composed of a circular array of loudspeakers as secondary sources and a square region at the center of the circular array as a listening area. Using the algorithms described in the previous chapter, first the locations of loudspeakers are sought on the loudspeaker region, and then the LS solution is employed to find the complex amplitudes of the loudspeakers.

The placement algorithms are evaluated from different points of view. First, the computational complexity of each algorithm is calculated in Section 3.2. Second, the effect of power is investigated on the reproduced sound field inside and outside of the listening area experimentally in Section 3.3. Third, in Section 3.4 the following question is answered: "How the optimum locations of loudspeakers change based on the frequency and maximum normalized power"? Then, the optimized locations of loudspeakers resulting from SVD-, CMP-, and Lasso-based algorithms are illustrated, and the behavior of each algorithm is explained. Finally, the error performance of the all systems will be compared in Section 3.5 for different system parameters such as frequency, power, size of the listening areas, and the locations of the primary source. The contributions of this chapter are:

- Investigating the role of the power limitation in the performance of the SFR system.

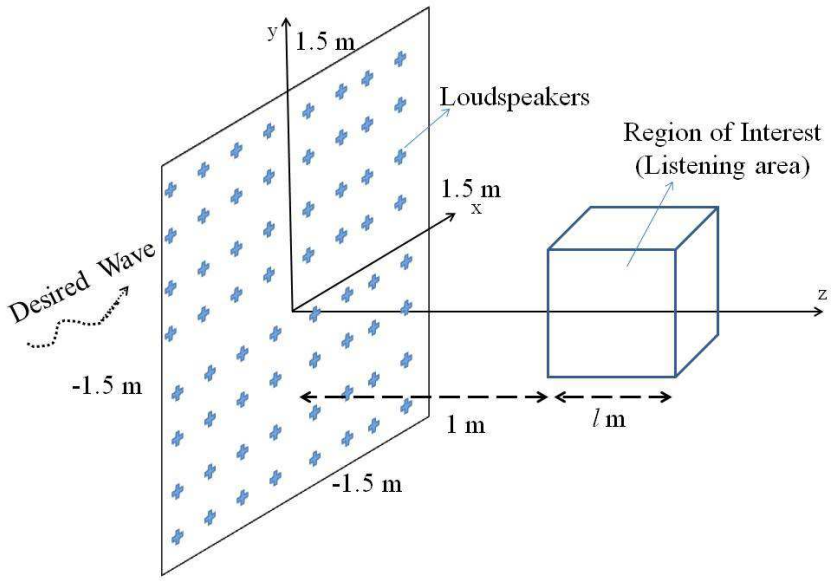


Figure 3.1: The 3-D configuration of interest for a sound field reproduction system.

- Qualitative analysis of the placement methods and inspecting how the optimum locations change in terms of the frequency and power.
- Comparing the error performance of the placement algorithms.

3.2 3-D SFR configuration

The 3-D configuration studied in this chapter is shown in Fig. 3.1. According to this figure, the loudspeaker region is a 3 m $\times$ 3 m square located on the (x, y)-plane. The region of interest is a 1 m $\times$ 1 m $\times$ 1 m cube, which is placed 1 m away from the LR. Let $N = 25$ be the number of loudspeakers, $N_v = 625$ be the number of candidate locations on the LR, and $M = 125$ be the number of sampling points distributed uniformly in the cubic listening area. Unless otherwise stated, it is assumed that the primary source is a point source located at $(0, 0, -8)$ m with a complex amplitude of $8e^{i\angle 0}$. Note that the placement methods under study are applicable to other LR and listening area geometries as well.

In the direct approximation method, the complex amplitudes of loudspeakers are determined such that the reproduction error is minimized at sampling points, and other points of the listening area are not under control. In this chapter, to evaluate the performance of various SFR methods, the reproduction error is calculated at $M_2 = 125,000$ points, uniformly distributed in the listening area. The error in dB is:

$$\text{Error (dB)} = 10 \log_{10} \left(\frac{\|(\mathbf{p}^{\text{des}})_{M_2} - (\mathbf{p})_{M_2}\|_2^2}{\|(\mathbf{p}^{\text{des}})_{M_2}\|_2^2} \right), \tag{3.1}$$

where $(\mathbf{p}^{\text{des}})_{M_2}$ and $(\mathbf{p})_{M_2}$ are the vectors containing M_2 samples from the desired and reproduced sound fields, respectively. Note that $M_2 \gg M$, so the error is evaluated at densely placed sampling points.

In SFR for stereo, frequencies of interest are up to 1500 Hz [15, 38, 75]. The reason is that at higher frequencies, the Inter-aural Time Difference (ITD) of the signal envelope, rather than the fine structure of the sound field itself, is reported to play a more important role in source localization. In SFR for noise suppression, passive controls are more efficient at high frequencies than active ones [76, 77] because the absorption of most materials is higher at higher frequencies. Hence, in both of these SFR applications, the main interest for SFR is at lower frequencies. For this reason, the frequency of the primary source in our simulations is in a low-frequency range, from 200 Hz to 2000 Hz.

The distance between the M sampling points at which the desired field is specified is 25 cm in our simulations, and it is kept fixed for different frequencies. The distance between the listening area and the primary source is 9 m. It should be noted that to avoid aliasing in SFR the distance between sampling points should be less than $\lambda/2$ when the primary source is far enough from the listening area. By decreasing the distance between the primary source and the listening area, to avoid the spatial aliasing, the required distance between the sampling points decreases. In [51], it has been proven that in order to reproduce a sound field along a line, the distance between sampling points can be derived using the following equation:

$$SNR \geq \frac{\pi}{4E_i\left(2d\sqrt{(\frac{\pi}{d_s})^2 - (\frac{2\pi f}{C})^2}\right)} \tag{3.2}$$

where SNR is the signal to noise (reconstruction noise) ratio, $E_i(x) = \int_x^\infty \frac{e^{-t}}{t}\,\mathrm{d}t$, d is the distance between primary source and listening area, and d_s is the distance between sampling points. In our configuration, the minimum distance between a line from the listening area and primary source is 9 m. For $SNR > 10^{10} (= 100$ dB), from Eq. (3.2):

$$\frac{\pi}{4E_i\left(2d\sqrt{(\frac{\pi}{d_s})^2 - (\frac{2\pi f}{C})^2}\right)} \geq 10^{10} \quad \Longrightarrow \quad \left(2d\sqrt{(\frac{\pi}{d_s})^2 - (\frac{2\pi f}{C})^2}\right) \geq 20.21 \tag{3.3}$$

Therefore the distance between sampling points should obey the following equation:

$$d_s \leq \frac{1}{\sqrt{(\frac{20.21}{2\pi\times 9})^2 + (\frac{2f}{C})^2}} \tag{3.4}$$

Based on this equation, if $f > 183$ Hz, $(\frac{2f}{C})^2$ is much larger than $\frac{20.21}{2\pi\times 9} = 0.357$, so to avoid aliasing the distance between sampling points should be less than $\frac{C}{2f} = \lambda/2$ which is the Nyquist sampling rate. Fig. 3.2 shows the distance between sampling points from Eq. (3.4) and from the Nyquist rate. According to this figure, these graphs are on top of each other for $f > 200$ Hz. Therefore, it can be concluded that the Nyquist sampling rate

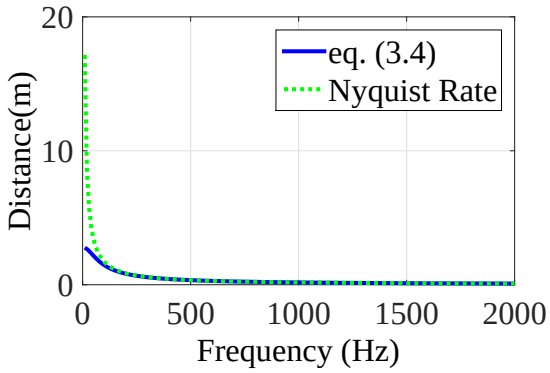


Figure 3.2: Maximum distance between sampling points from Eq. (3.4) and Nyquist's rate.

works for our proposed structure because the distance between sampling points and primary source is large enough. As mentioned earlier our system is under-sampled for frequencies greater than 686 Hz. However, for evaluation, the reproduction error is calculated at M_2 sampling points spaced by 2 cm, which corresponds to half the wavelength at 8403 Hz. Therefore, the reported errors are reliable for the frequency range covered here, since the error field is well over-sampled.

The required number of operations, including summation, multiplication, division, comparison, for placement algorithms – SVD, CMP, Lasso, and GS– are given in Table 3.1 for three different values of (M, N, N_v). Considering the parameters selected for our simulations, the required number of operations for Lasso-based placement is 35.7, 21.06, and 17.8 times those of the CMP-based, SVD-based, and GS-based algorithms, respectively, when the number of iterations, N_{it}, is 1. Therefore the computational complexity of Lasso is much higher than other placement algorithms and it increases linearly with the number of iterations. The reason is that the computational complexity of the Lasso versus N_v is $O(N_v^2)$ and $N_v \gg N$, while the computational complexity of CMP-, GS-, and SVD-based placement versus N_v is $O(N_v)$.

The computational complexity of these algorithms is important if the locations of loudspeakers are dynamic DoFs. For example, suppose an array of N_v loudspeakers is employed, and N loudspeakers out of N_v loudspeakers are to supposed be selected for SFR in real-time based on the desired field by placement methods. However, if loudspeaker locations are static DoFs, which is the case in practical SFR systems, the computational complexity of placement algorithms are not a crucial factor meaning that they are determined offline (before system operation) and kept fixed during system operation.

3.3 Effects of power constraint on SFR systems

Limiting power in SFR systems is required to control the sound field outside of the listening area although it may increase reproduction error. In active noise cancellation, one is

Table 3.1: Required number of operations for placement algorithms

(M, N, N_v)	MP-based	SVD-based	Lasso-N_{it}	GS-based
$(125, 16, 900)$	3.6×10^6	5.80×10^6	$2.03 \times 10^8 \times N_{it}$	7.3×10^6
	(x)	$(1.6x)$	$(55.8N_{it}x)$	$2x$
$(64, 25, 625)$	2.03×10^6	3.5×10^6	$5.01 \times 10^7 \times N_{it}$	4.1×10^6
	(x)	$(1.7x)$	$(24.6\ N_{it}x)$	$2x$
$(125, 25, 900)$	5.6×10^6	9.6×10^6	$2.03 \times 10^8 \times N_{it}$	1.15×10^7
	(x)	$(1.7x)$	$(35.7N_{it}x)$	$2x$

required to reproduce the sound field in the listening area providing that the sound field does not change considerably outside of the listening area. Therefore, one of the important parameters is the maximum normalized power, $p_{\max}$. Based on Eqs. (2.7) and (2.8), the regularization parameter is a non zero value, which is found in order to limit the power of loudspeakers to $p_{\max}$. As seen from Eq. (2.17), this parameter increases the error in the SFR system. For a general ATF matrix, according to Eq. (2.17), two different cases occur in practice:

Case 1: $\sum_{m=N+1}^{M} |c_m^g|^2 \gg \sum_{n=1}^{N} |c_n^g|^2$. This implies that the desired field falls almost entirely in the null space of the ATF matrix. Therefore, the reproduction error does not change significantly by changing the normalized power (or γ) since it is approximated by $\sum_{m=N+1}^{M} |c_m^g|^2$ (which is independent of γ). In this case, because $|c_i^g|$'s for $1 \leq i \leq N$ are negligible, the normalized power of the loudspeaker array in Eq. (2.16) is also small. It means that, in this case, the SFR error is large while the normalized power of the loudspeaker array is small. In the extreme case, all columns of the ATF matrix are perpendicular to the desired field, all loudspeakers are off, the consuming power of the loudspeakers is zero, and the reproduction error is equal to 1 (0 dB).

Case 2: $\sum_{m=N+1}^{M} |c_m^g|^2$ and $\sum_{n=1}^{N} |c_n^g|^2$ are comparable. In this case the second part of error from Eq. (2.17), $\sum_{m=N+1}^{M} |c_m^g|^2$, is inevitable while the error related to the first term in Eq. (2.17) can be decreased by decreasing the regularization parameter (increasing the power of the loudspeakers). Hence, smaller γ leads to smaller error and larger power. The selected value for γ in order to control the first term in Eq. (2.17) highly depends on the singular values of the ATF matrix. Consider two well-conditioned ATF matrices which result in equal c_n's but the singular values of the first matrix are larger than those of the second matrix. In order to keep the first term in Eq. (2.17) fixed, the first ATF matrix need larger value for γ, and consequently smaller power for the loudspeakers. However, the second system should employ smaller γ and spend larger power for the loudspeakers.

Therefore, in *Case 2*, $p_{\max}$ plays an important role in the reproduction error and the reproduced field outside of the listening area. The effect of limiting the power of the loudspeakers on the reproduced sound field can be justified as follows:

Let p be the pressure at a point caused by the loudspeaker array, and p_j be the pressure caused by the j-th loudspeaker in the array. Humans hear the intensity of a sound field at each point, which is proportional to the squared pressure, p^2

$$p^2 = (\sum_{j=1}^{N} p_j)^2 = \sum_{j=1}^{N} |p_j|^2 + \sum_{j,l\neq i} p_j p_l^*. \tag{3.5}$$

The pressure at each point changes between zero for destructive superposition (when the second term is equal to $-\sum_{j=1}^{N}|p_j|^2$) of waves and $(\sum_{i=1}^{N}|p_j|)^2$ for constructive superposition (when second term is positive). Based on the law of the conservation of energy, the constructive superposition at some points results in a destructive one at other points. In SFR, the excitation of loudspeakers are determined such that the pressure values inside the listening area is close to the desired values. Therefore, the pressure values inside the listening area are known and fixed. Limiting the power of loudspeakers is equivalent to limiting the first term of Eq. (3.5) and consequently increasing the second term inside the listening area in order to achieve the desired pressure. It implies that limiting the power of the loudspeakers leads to a constructive sound field inside the listening area, and most likely the destructive or partially constructive superposition outside of the listening area.

The following experiments show the effect of $p_{\max}$ on SFR performance. The primary point source with complex amplitude of $A = 8e^{i\angle 0}$ is at $(0, 0, -8)$ m. A cross-section of its sound field is shown in Fig. 3.3(a). The cross-section of the cubic region of interest is shown as the square. Three SFR systems are designed with $N = 25$ loudspeakers placed on the LR using the CMP-based algorithm with $p_{\max} \in \{0.5, 10, \infty\}$, where $p_{\max} = \infty$ means unconstrained power. The cross-sections of the real part of the pressure fields for the three systems are shown in Figs. 3.3(b), (d), and (f), respectively, with the corresponding squared absolute value of the error fields shown in Figs. 3.3(c), (e), and (g). Since all analysis are in frequency domain the results of Figs. 3.3(c), (e), and (g) are proportional to the time-averaged error energy at each point by Parseval's theorem. As seen in Figs. 3.3(b), (d), and (f), all three systems approximate the desired field inside the region of interest (square). The approximation gets better as $p_{\max}$ increases, but the most apparent difference in the produced sound fields is outside the region of interest, where the sound can be highly intensified (note the bright and dark regions in Figs. 3.3(d), and (f)), which shows up as error in Figs. 3.3(c), (e), and (g). In most situations, $p_{\max}$ should be kept as small as possible while achieving some required SFR approximation inside the listening area.

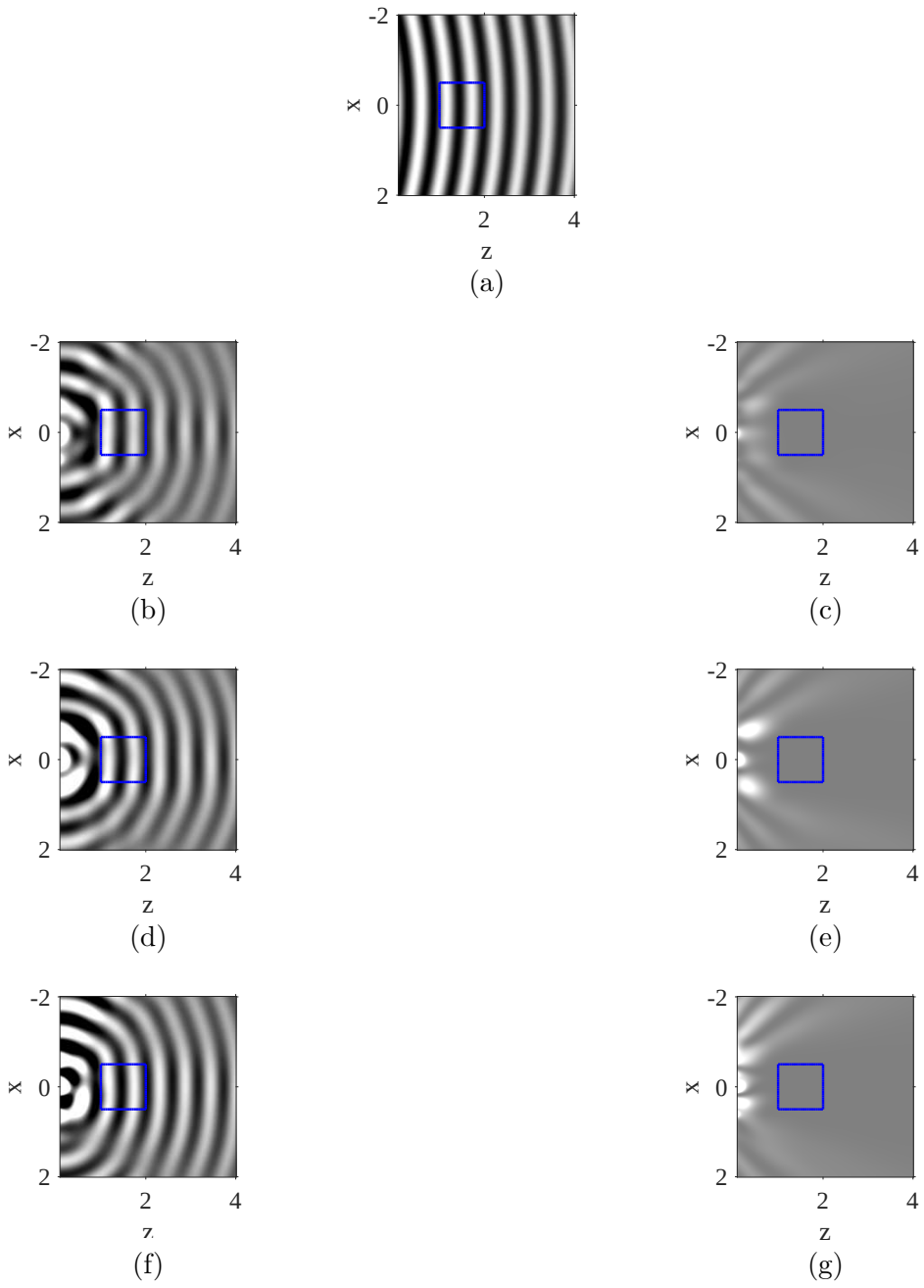


Figure 3.3: (a) Real part of the desired field, Real part of the reproduced field and squared absolute value of the error field with (b),(c) $p_{\max} = 0.5$, (d), (e) $p_{\max} = 10$, and (f), (g) unconstrained power.

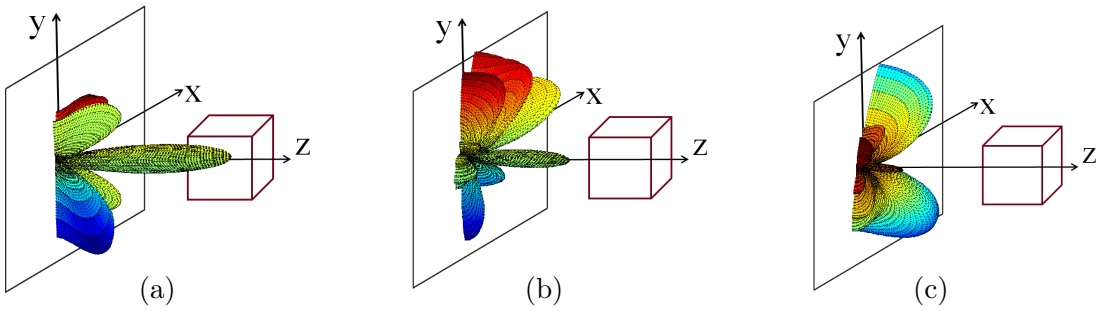


Figure 3.4: Forward direction radiation pattern of loudspeaker array at $f = 600$ Hz for (a) $p_{\max} = 0.5$, (b) $p_{\max} = 10$, and (c) unconstrained power.

To gain a deeper understanding of the reasons behind the behavior shown in Fig. 3.3, the far-field radiation patterns of the loudspeaker arrays in the three systems are shown in Fig. 3.4 along with the cubic region of interest. As seen in the figure, the system with the lowest $p_{\max}$ (Fig. 3.4(a)) provides the best directionality, with the dominant main lobe pointing towards the region of interest. As $p_{\max}$ increases, the resulting array pattern may provide improved approximation to the desired field inside the region of interest, but this comes at the cost of much energy wastage in other directions, with associated unwanted "hotspots".

Another important role of power limitation in SFR systems is in controlling the regularization parameter γ in Eqs. (2.8)-(2.9), which results in improved system robustness [15, 78, 79]. Without the power constraint, the regularization parameter essentially becomes zero in Eqs. (2.8)-(2.9). In this case, depending on the system configuration, the condition number of $\mathbf{G}^H\mathbf{G}$ may be too large, which would make matrix inversion in Eq. (2.9) problematic and would have a detrimental effect on system robustness. It is worth mentioning that in our simulations, to ensure that the normalized power is less than $p_{\max}$, the regularization factor is found through the Newton method in [39].

In the next experiment, the problem of power allocation to different iterations of the CMP-based placement algorithm is investigated. Specifically, three cases are compared: exponential allocation $p_n = Bp_0^n$ with $p_0 \in \{0.1, 0.6\}$ and p_0 from Eq. (2.44). The reproduction error as a function of $p_{\max}$ is shown in Fig. 3.5 for operating frequencies 600 Hz and 1200 Hz. For $p_0 = 0.1$, most of the power is allocated to the first few iterations, depleting the power budget for subsequent iterations. This is not a good strategy under a power constraint. The error improves when p_0 increases to 0.6, but of the three schemes compared here, the best performance is achieved using the p_0 from Eq. (2.45). As mentioned in Section 2.6.2, this is not necessarily the optimal way to allocate power in CMP, but it is sufiňĄciently good to demonstrate the overall superiority of CMP against other loudspeaker placement methods in this study, as will be shown in the remainder of this chapter.

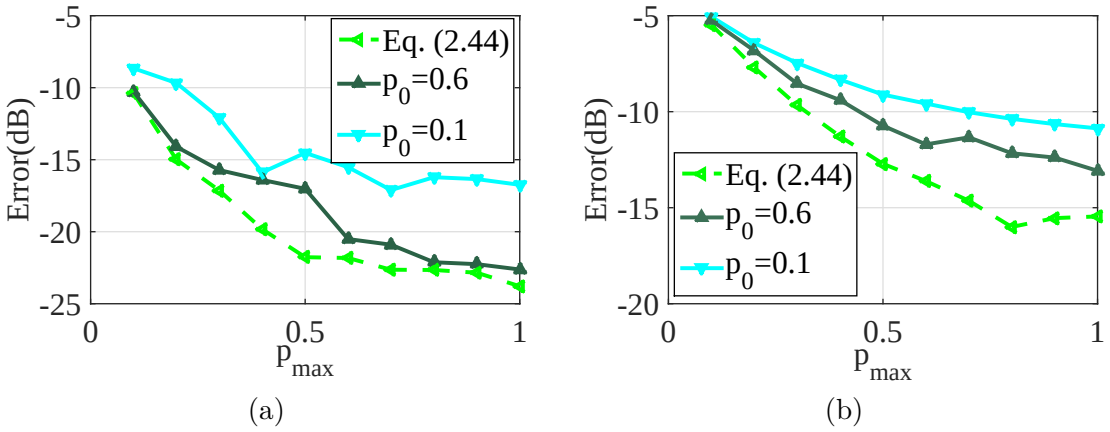


Figure 3.5: Error performance versus power for frequencies for (a) 600 Hz, (b) 1200 Hz.

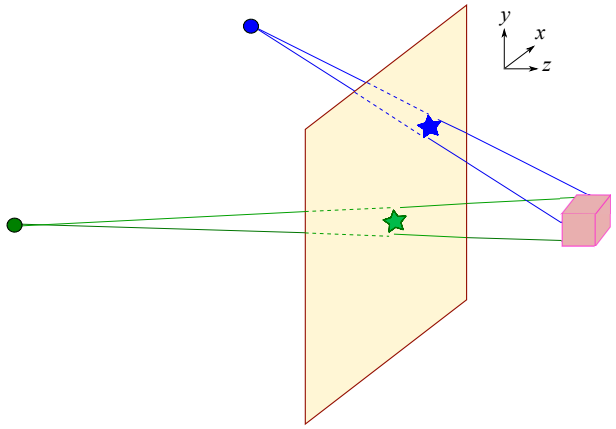


Figure 3.6: Illustration of ray-cuts (shown as stars) on the LR for two primary sources.

3.4 Qualitative analysis of loudspeaker placement

Finding optimal locations for the loudspeakers is computationally challenging because the error function is not convex in terms of the locations [15], while the number of possible combinations is large. For example, with $N = 25$ and $N_v = 900$, the number of possible combinations is $\binom{N_v}{N} = \binom{900}{25} > 10^{48}$. In this section, some intuition is discussed for the best locations for the loudspeakers as a function of the parameters of the desired field such as the location(s) of primary source(s), operating frequency, and $p_{\max}$.

Consider Fig. 3.6, which shows two primary sources as circles. Rays are drawn from these primary sources to the cubic region of interest, and the positions where they cut the square LR are indicated by stars. These positions will be referred to as *ray-cuts*.

Suppose there is only one primary source, say the lower circle, and that only one secondary source can be used to regenerate the sound field produced by the primary source. Where should the secondary source be placed? Intuitively, the corresponding ray-cut (lower star) should be a good position for the secondary source, because it lies on the direct path

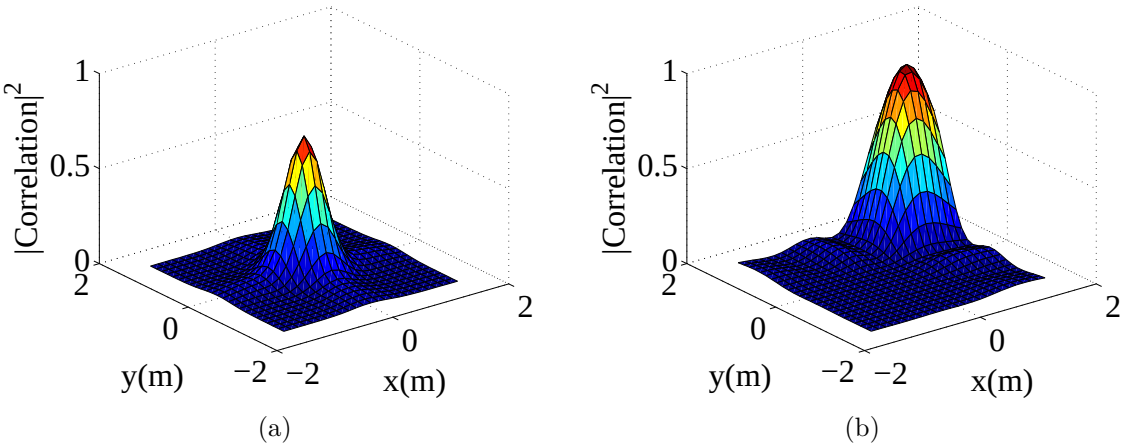
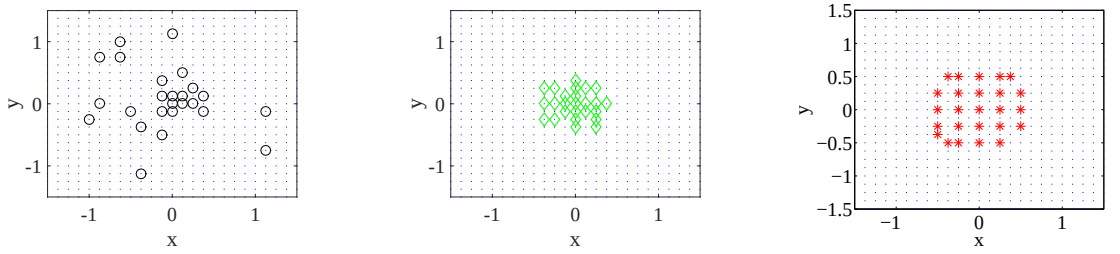
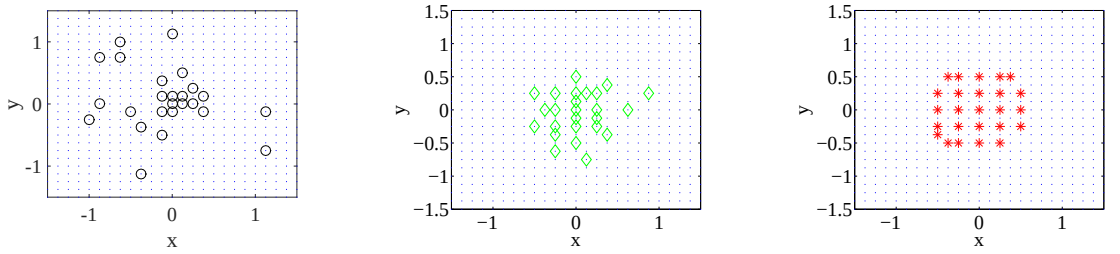


Figure 3.7: Correlation between the desired and produced sound field as a function of the position of the secondary source within LR, when the primary source is at (a) $(0, 0, -8)$ m and (b) $(4, 4, -4)$ m.

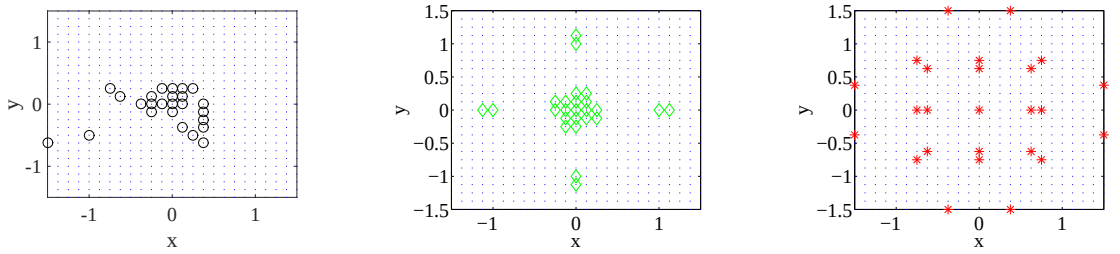
from the primary source to the region of interest. This intuition is confirmed by the following experiment, where the correlation is computed between the primary sound field and the field produced by a secondary source placed at various locations in the LR. With the system parameters as before (except that only $N = 1$ secondary source is allowed) and $f = 600$ Hz, the primary source is placed at $(0, 0, -8)$ m (similar to the lower circle in Fig. 3.6) and the correlation between the desired and produced sound fields as a function of the location of the secondary source on the LR is computed. The result is shown in Fig. 3.7(a). As expected, the correlation is highest when the secondary source is placed at $(0, 0)$ m on the LR, the ray-cut (lower star) in Fig. 3.6. The analogous correlation plot for the other secondary source (upper circle) from Fig. 3.6 is shown in Fig. 3.7(b), and it is seen that the correlation is highest at the corresponding ray-cut for this source, which is towards the upper-right corner of the LR.



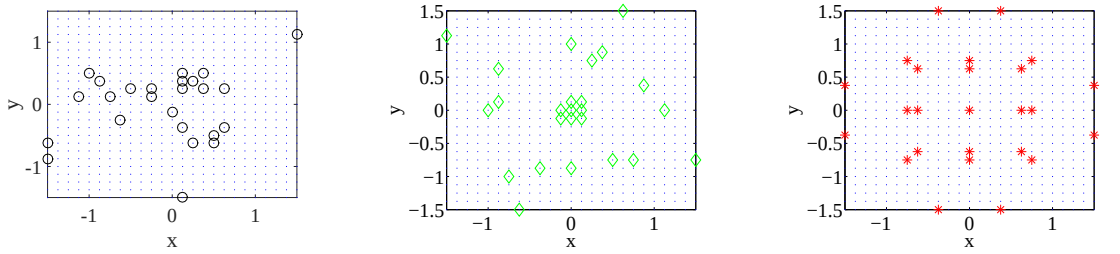
(a) SVD $(1400, 0.1)$ (b) CMP $(1400, 0.1)$ (c) Lasso $(1400, 0.1)$



(d) SVD $(1400, 1)$ (e) CMP $(1400, 1)$ (f) Lasso $(1400, 1)$



(g) SVD $(400, 0.1)$ (h) CMP $(400, 0.1)$ (i) Lasso $(400, 0.1)$



(j) SVD $(400, 1)$ (k) CMP $(400, 1)$ (l) Lasso $(400, 1)$

Figure 3.8: Loudspeaker placement from SVD (left column), CMP (middle) and Lasso (right) for $f \in \{400, 1400\}$ Hz and $p_{\max} \in \{0.1, 1\}$.

While loudspeaker placement is relatively easy and intuitive in these simplest cases, it becomes much more challenging as the number of primary and/or secondary sources increase. Intuitively, it can still be expected that ray-cuts would be good choices for loudspeaker placement, but it is not clear which other locations, beside ray-cuts, would be the best choices for reducing the sound field reproduction error. For sound field reproduction under a power constraint, the optimum placement of loudspeakers depends not only on the position of the primary source(s) and the frequency of the desired field, but also on the available power. In the following paragraphs, it is explained how the distances between the selected locations change with frequency and power of the primary source.

Lower frequencies: at these frequencies, if the maximum normalized power is large enough, the neighboring (closest) ATFs to the first selected ATF (which is close to ray-cut) will not be selected. The reason is that since the wavelength is much larger than the distance between the candidate locations, the ATF of these points are similar to the ATF of the first selected loudspeaker. For example, in our structure, the distance between neighboring candidate locations are 12.5 cm, but the wavelength is greater than 1.14 m for $f < 300$ Hz. In this situation, the optimum distance between loudspeakers is large. However, if the maximum normalized power is small, the neighboring locations are optimal. The reason is that because of the power limitation the ATF corresponding to the first selected location cannot describe the desired vector completely. Therefore, since the neighboring ATFs are similar, the consequent ATFs, which are similar to the first one, help the first selected ATF to completely describe the desired vector.

Higher frequencies: At these frequencies, the distance between the neighboring candidate locations is comparable with the wavelength (34 cm for $f = 1000$ Hz), so the neighboring ATF are not similar. Therefore, in this case, the loudspeakers are more concentrated around the first selected location in comparison with the previous case, and they change with $p_{\max}$ the same as in the previous case.

If the number of primary sources increases, the optimum locations cluster around the ray-cuts corresponding to each primary source. The number of loudspeakers in each cluster depends on the magnitude of each primary source, the larger magnitude, the more loudspeakers in its corresponding cluster.

Form the above discussion, it can be concluded that for the complicated scenarios, a placement algorithm is required to find the optimum locations for loudspeakers. The following experiment offers further insight into the selected locations by the placement algorithms.

A primary point source with complex amplitude of $A = 8e^{i\angle 0}$ is placed at $(0, 0, -8)$ m. Several SFR systems are designed by placing $N = 25$ loudspeakers on the LR using CMP-based, Lasso-based, and SVD-based placement algorithms for frequencies $f \in \{400, 1400\}$ Hz and $p_{\max} \in \{0.1, 1\}$. The resulting placements are shown in Fig. 3.8. The middle column (Figs. 3.8(b), (e), (h), (k)) shows CMP-based placement, the right column (Figs. 3.8(c), (f),

(i), (l)) shows the Lasso-based placement, and the left column (Figs. 3.8(a), (d), (g), (j)) shows SVD-based placement for the four pairs $(f, p_{\max})$.

Fig. 3.8 indicates that all methods select the locations that are clustered around the ray-cut (center of the LR in this case), which agrees with the intuition noted above. The behavior of each algorithm is explained as follows:

CMP-based algorithm: In this algorithm, at higher frequencies all loudspeakers are arranged around the "ray-cuts", and based on the maximum normalized power the concentration around this point changes. At lower frequencies, for large $p_{\max}$, the distance between the loudspeakers is large with a single cluster around the ray-cut, while for small $p_{\max}$ the loudspeakers are arranged around the ray-cut in different clusters. The reason is that, in this algorithm, at the n-th iteration the inner product between the selected dictionary member and residual vector are compared with the allocated power to the n-th iteration (p_n). The scaling coefficient of the selected dictionary member is the minimum value of p_n and the projection (inner product) of the residual error on the selected dictionary member. Therefore, the selected locations depend on p_n and $p_{\max}$. Assume p_n is less than the inner product between the selected dictionary member and the residual error. Then, the magnitude of the scaling coefficient of the n-th vector is equal to p_n which is less than the required magnitude (the inner product between two vectors). As a result, only a small part of the selected dictionary member is removed from the desired vector, and the selected dictionary member of the next iteration would be close to the selected dictionary member at the n-th iteration in order to compensate the effect of power limitation. This process continues until the selected dictionary members completely describe the desired vector at the most correlated direction, and this forms the first cluster. After this step, another dictionary member is selected, and again another cluster is formed. Therefore, the CMP-based method arranges loudspeakers in different clusters for small $p_{\max}$. When $p_{\max}$ is large enough, after forming the first cluster around the ray-cut, the projection of the desired vector on the selected dictionary member would be larger than the assigned power (p_n), so after several iterations the clusters do not form around the selected locations.

Lasso-based algorithm: The Lasso algorithm finds the coefficients of all ATF vectors and updates them iteratively. For this purpose, at each iteration, to update the coefficients of the n-th ATF, the error vector is calculated assuming that the corresponding coefficient to the n-th ATF vector is zero. Then, the inner product between the error and the n-th ATF vector is calculated. If it is less than a threshold, the corresponding coefficient to the n-th ATF vector is set to zero, otherwise a coefficient is assigned to that ATF vector. Suppose that all ATF vectors are orthogonal, in this case, the error vector, which is calculated without considering the n-th ATF vector, is in parallel with the ATF vector. Therefore, the inner product would be a large number, and a nonzero number is assigned to the n-th ATF vector. Now suppose that there is a vector among the candidate ATFs with high correlation with the n-th ATF vector. In this case, this vector plays the role of the n-th ATF vector

in reconstruction, so the error vector would be nearly perpendicular to the n-th vector. It means that the inner product would be a small number, and the coefficient assigned to the n-th ATF is zero. Therefore, if there is another candidate vector whose correlation with n-th ATF vector is large, the projection of the n-th ATF vector onto the error vector would be small, and the corresponding coefficient would be zero. It is concluded that the Lasso solver removes the ATFs which are correlated and focuses the whole power in one of them. At lower frequencies, since the correlation among the ATF vectors are high, the selected locations are dispersed. However, by increasing the frequency the selected locations are more concentrated around the ray-cut.

The difference between Lasso-based and CMP-based methods is that in the CMP-based method $p_{\max}$ is taken into account while in the Lasso-based method the selected locations do not depend on $p_{\max}$. Therefore, the configuration provided by the CMP-based algorithm is more adaptable to constraints.

SVD-based method: Here, the inner products between the columns of the ideal ATF matrix and the candidate ATFs are calculated. The columns of the ideal ATF matrix are equal to a combination of the desired vector and its perpendicular vectors.

$$\mathbf{g}_n^{\text{ideal}} = \sum_{j=1}^{N} (v_{nj}^g)^H \sigma_j^g \mathbf{u}_j^{\text{ideal}}.$$

The scaling coefficient of the desired vector is $v_{nj}^* \sigma_1$. This coefficient is large because σ_1 is larger than other singular values. Hence, most of the selected locations are around the "ray-cut." However, for some values of v_{1j}^g, the scaling coefficient of the desired vector is not the largest which means that the selected locations would be further away from the ray-cuts.

In addition, the SVD-based method approximates the covariance matrix. The covariance matrix of the ideal matrix is equal to that of the benchmark configuration (shown in Fig. 3.8(a)), and the effect of the covariance matrix in the approximation process increases by increasing N_c. Therefore, increasing N_c leads to more similar placement to the benchmark. Based on the intuition behind the optimum locations of loudspeakers, at lower frequencies with large power the distance between loudspeakers increases. Therefore, for this case, N_c should be a large number. In our simulations, N_c is selected experimentally between 1 and 25 as given in Eq. (2.28). It means that as with the CMP-based method, $p_{\max}$ is taken into account in this algorithm.

To show the effect N_c on the performance of SVD-based placement, the error with $N_c = 1$ is compared against the error when N_c is calculated from Eq. (2.28). The reproduction errors are shown in Figs. 3.9(a) and 3.9(b) for $f = 400$ Hz and $f = 700$ Hz, respectively, as a function of the maximum normalized power $p_{\max}$. From Eq. (2.28), parameter N_c changes between 1 and 23 as a function of $p_{\max}$. The figures show that employing N_c, especially at lower frequencies, and a larger $p_{\max}$, improves the performance of the SVD-based method.

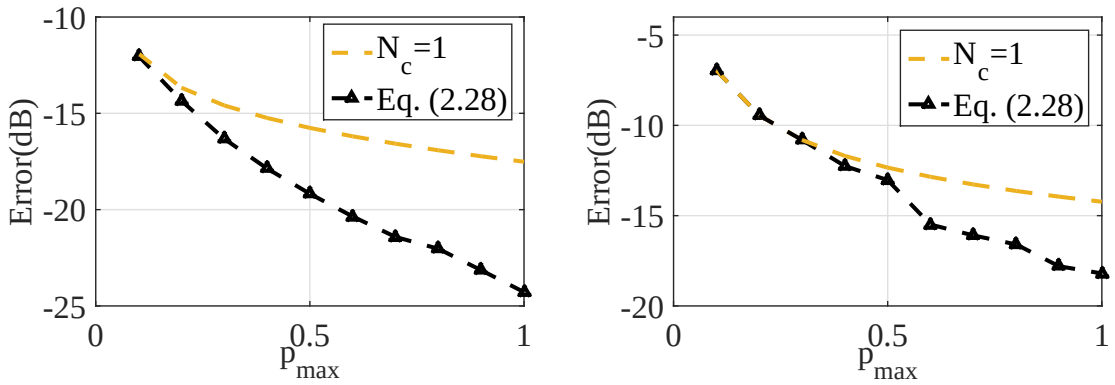


Figure 3.9: Reproduction error of the SVD-based placement with $N_c = 1$ and N_c from Eq. (2.28) at (a) $f = 400$ Hz and (b) $f = 700$ Hz.

3.5 Error performance

In this section, the SFR performance of the configuration of Fig. 3.1 under different conditions will be investigated. In the first few experiments, it is assumed that the frequency and amplitude of the primary source are fixed, while in the last three experiments, frequency and/or amplitude of the primary source are changing.

In the first experiment, we compare the SFR performance of the five placement methods – benchmark, SVD, CMP, Lasso, and GS – for a single-tone primary source located at $(0, 0, -8)$ m. The experiment is performed at various frequencies and power limits. Once the placement is computed for a particular frequency and power limit, loudspeaker excitations are obtained from Eq. (2.9). The results are shown in Fig. 3.10(a) versus frequency when $p_{\max} = 0.3$ and in Fig. 3.10(b) versus $p_{\max}$ at $f = 700$ Hz.

In all cases, the reproduction error increases at higher frequencies. This is because the desired field $\mathbf{p}^{\text{des}}$ is undersampled at higher frequencies, so the synthesized field cannot accurately reproduce it between the sampling points (recall the error field is over-sampled to reveal this). Also, as expected, the error decreases as $p_{\max}$ increases, because higher available power allows for better approximation of the desired field.

Based on the comparisons of Fig. 3.10 several observations can be made. First, the SVD-based placement is better than the benchmark across the range of frequencies and powers, except at very low frequencies and low power, where the benchmark is slightly better. The reason is that under these conditions, SVD-based placement selects mostly locations near the ray-cut (see Fig 3.8(g)), while the benchmark has more dispersed loudspeakers, which helps reduce the error in certain parts of the listening volume. Second, the performance of Lasso is better than SVD at higher frequencies, as is expected from [15], where it was shown to offer solid performance in the case of undersampling. However, as $p_{\max}$ decreases, SVD starts performing better than Lasso, especially at low frequencies. The reason is that the

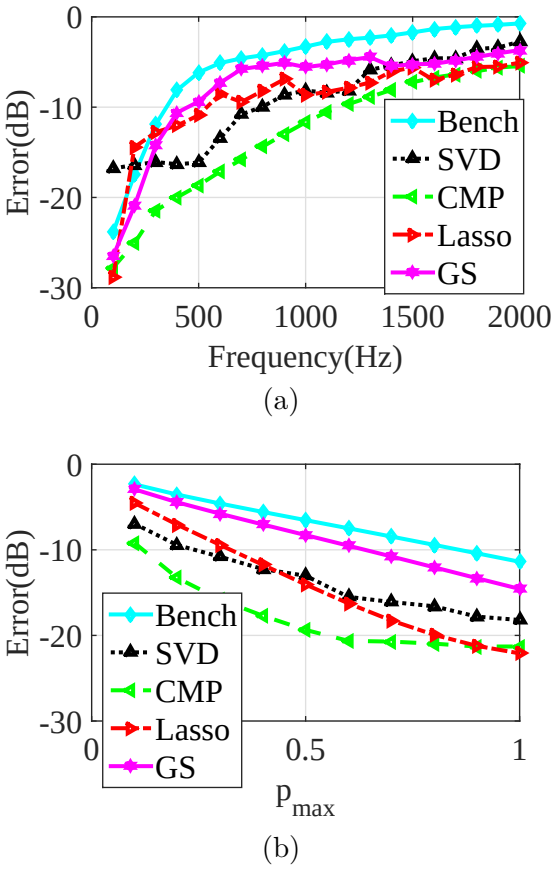


Figure 3.10: Error performance for benchmark, SVD-based, CMP-based, Lasso-based, and the method in [58] versus (a) frequency at $p_{\max} = 0.3$, and (b) versus $p_{\max}$ at $f = 700$ Hz.

SVD-based placement considers power limitation (in calculation of N_c), whereas Lasso does not. Third, CMP is considerably better than Lasso in placing loudspeakers under power limitation and at low frequencies. As the frequency and $p_{\max}$ increase, the performance gap reduces. Finally, the performance of GS is worse than all other placements except the benchmark. The reason is that GS-based placement method takes neither the desired field nor the power limitation into account when selecting loudspeaker locations.

The reproduction error as a function of the size of the cubic listening area is shown in Table 3.2 at $f = 600$ Hz and $p_{\max} = 0.5$ for all placement methods. The performance of all placement methods worsens as the size of the cubic region increases. The reason is twofold. First, since the distance between sampling points is fixed (25 cm), the number of sampling points increases as the size of the cubic region increases. Therefore, the number of equations in Eq. (2.4) increases while the number of unknowns remains the same, so the system becomes more over-determined and overall error increases. Second, as the size of the cube increases and becomes more similar to the dimension of the LR, ray-cuts cover a larger portion of the LR and the loudspeaker placement produced by the different

Table 3.2: Reproduction error in dB as a function of the dimension of the cubic listening area (L) at $f = 600$ Hz and $p_{\max} = 0.5$.

L (m)	CMP	SVD	Lasso	GS	Bench.
0.5	-29.74	-31.85	-18.72	-27.85	-16.99
1.5	-12.47	-8.36	-10.38	-6.35	-4.84
2	-7.15	-4.86	-6.57	-4.86	-2.91
2.5	-4.80	-3.20	-4.33	-3.84	-1.82
3	-3.45	-2.28	-2.76	-3.10	-1.56

Table 3.3: Reproduction error in dB versus the frequency for various locations of the primary source at $f = 600$ Hz and $p_{\max} = 0.5$.

Location (m)	CMP	SVD	Lasso	GS	Bench.
$(1.9, 0, -7.7)$	-21.05	-14.86	-16.44	-11.94	-6.76
$(0, -2.8, -7.4)$	-20.57	-15.29	-16.47	-9.16	-6.53
$(3.2, 3.2, -6.5)$	-21.02	-14.73	-13.52	-7.86	-6.34
$(4.8, 0, -6.8)$	-20.99	-13.62	-14.56	-8.64	-7.41
$(4.1, -4.1, -5.4)$	-20.26	-13.43	-12.42	-7.22	-6.85

methods approaches uniform, so the performance becomes closer to that of the benchmark configuration.

As mentioned above, power constraint plays an important role in SFR. However, evaluation of the reproduction error without the power constraint offers insight on how different placement algorithms work. Recall that the first term in Eq. (2.17) results from the power constraint and the second term is created by the parts of the desired field that fall into the null space of the ATF matrix, which cannot be reproduced even with unlimited power. In the next experiment, the SFR systems that are designed without the power limitation are examined. The excitation vector in all cases is obtained with $\gamma = 10^{-6}$ in Eq. (2.9) (arbitrary choice, not particularly sensitive). The reproduction error is shown in Fig. 3.11(a), while the normalized power, $\|\mathbf{s}\|_2^2$, is shown in Fig. 3.11(b) on the logarithmic scale. According to this figure, the performance of the Lasso-based placement method is better than all other placement methods when the power is not limited, which means that the ATFs selected by Lasso better represent the desired vector. However, Lasso-based placement also needs higher power, at higher frequencies, than CMP to achieve this reduced error, as shown in Fig. 3.11(b). Another point of interest is that the consumed power by the GS-based method is the least among the placement methods, because in this placement method, the columns of the ATF matrix are selected to be less correlated, which results in a well conditioned ATF matrix with large singular values and consequently lower consuming power.

In the next experiment, the performance of the placement algorithms is examined for different locations of the primary source. The results are shown in Table 3.3. The distance between the primary source and the origin is 8 m in all cases, and $f = 600$ Hz and $p_{\max} = 0.5$

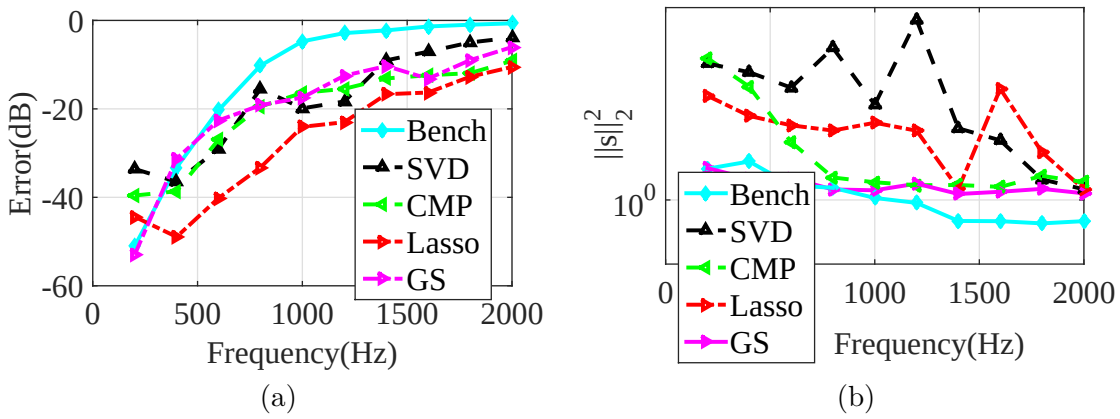


Figure 3.11: Results without power limitation: (a) Reproduction error versus frequency, (b) Total power versus frequency, for $\gamma = 10^{-6}$ in Eq. (2.9), Lasso requires considerably higher power.

are used in all cases. The CMP- and Lasso-based methods again outperform other methods for all locations of the primary source. Among the placement methods, the performance of the GS-based method is poorer because the selected locations by this algorithm do not depend on the location of the primary source.

The next experiment examines the effect of the number of loudspeakers (N) and the number of candidate locations (N_v) on the reproduction error under a power limit. Fig. 3.12(a) shows the reproduction error as a function of N for the various placement methods when $N_v = 625$. As expected, the performance of all placements improves as the number of loudspeakers increases. CMP and SVD offer similar performance, with CMP being slightly better. In short, for the configuration of interest, the performance of Lasso is between that of SVD and benchmark for low and medium values of N, and becomes better than SVD, CMP, and GS for large values of N.

Fig. 3.12(b) shows the reproduction error as a function of the number of candidate locations N_v for SVD-, CMP-, Lasso-, and GS-based placement methods, when $N = 25$. CMP outperforms SVD and GS, but the performance of Lasso oscillates between these two, depending on the number of candidate locations. It turns out that when N_v is odd, and one of the candidate locations coincides with the ray-cut (in this case, the origin), Lasso selects that location for loudspeaker placement, and other locations are selected far from the ray-cut (see Fig. 3.8(i)). If N_v is even then the ray-cut is not included among the candidate locations. In this case Lasso forms a cluster of loudspeakers around the ray-cut, but this reduces the number of loudspeakers available for other areas on the LR, which results in more concentrated loudspeaker locations and improving the error performance under power constraint. The opposite effect exists for Lasso in the absence of power limitation. For example, in Fig. 3.11, the reproduction error of Lasso at $f = 800$ Hz is -33 dB when the number of candidate locations is $N_v = 25 \times 25 = 625$ (i.e., odd number and more dispersed

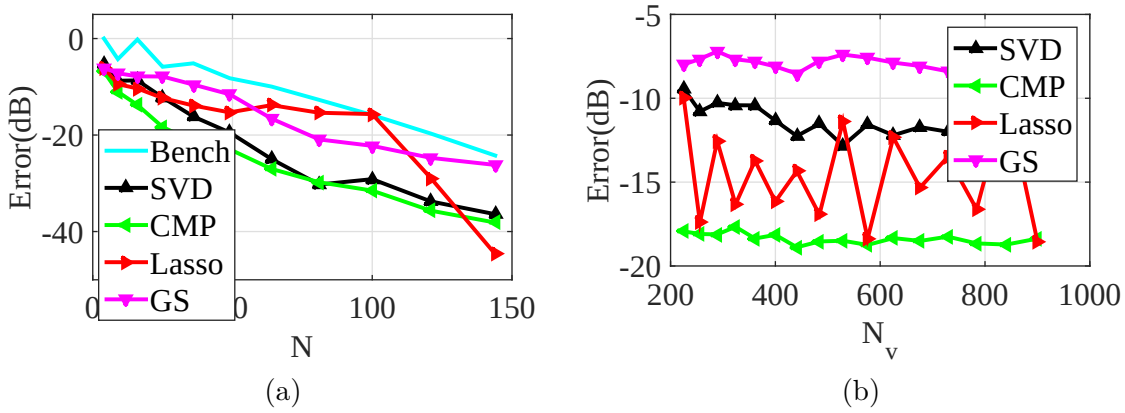


Figure 3.12: Reproduction error at $f = 800$ Hz and $p_{\max} = 0.5$ versus: (a) the number of loudspeakers N when $N_v = 625$, (b) the number of candidate locations N_v when for $N = 25$.

loudspeaker locations). However, if the number changes to $N_v = 24 \times 24 = 576$ (more concentrated loudspeaker locations), the error increases by 12 dB, to -21 dB, when there is no power constraint. This demonstrates how Lasso is more sensitive to the initial set of candidate locations than SVD, CMP, or GS.

So far, it has been assumed that the frequency and amplitude of the primary source are fixed. In the next three experiments, the performance of the system will be evaluated for the case when the frequency or amplitude of the primary source is variable, while its location is fixed at $(0, 0, -8)$ m. First, suppose loudspeaker locations are sought that would work well over a range of frequencies of the primary source, say between 300 Hz and 1300 Hz. Two ways can be envisaged to find the appropriate loudspeaker locations. 1) Find the locations for more than N loudspeakers at each frequency individually, and then select the N common locations among all the sets of locations found at various frequencies. If the common set contains less than N locations, the closest locations among different sets are added. 2) Another way is to select the locations for the mid-point of the frequency range. The latter approach is employed in the experiment in Fig. 3.13. In this figure, the locations are selected for $f = 800$ Hz, and the reproduction error is reported at other frequencies by using the locations selected for $f = 800$ Hz. The excitation vector is computed from Eq. (2.9) at each frequency. For comparison purposes, the thin dashed lines in this figure show that the reproduction error achieved when the locations of loudspeakers are selected at each frequency individually. As shown in the figure, locations selected for $f = 800$ Hz work well at other frequencies as well. Another observation from this figure is the sub-optimality of the placement algorithms. This can be seen at other frequencies, where the locations selected at $f = 800$ Hz sometimes give rise to a smaller reproduction error compared to the locations selected at the corresponding frequencies.

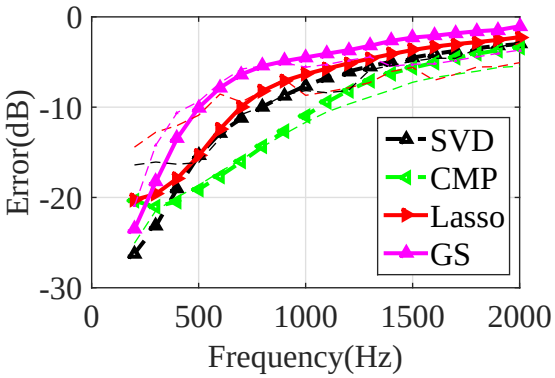


Figure 3.13: Reproduction error for the primary source with variable frequency versus the frequency, when $p_{\max} = 0.3$. The solid lines show the error achieved by placement designed for $f = 800$ Hz, while the thin dashed lines show the error achieved by placement designed for the particular frequency.

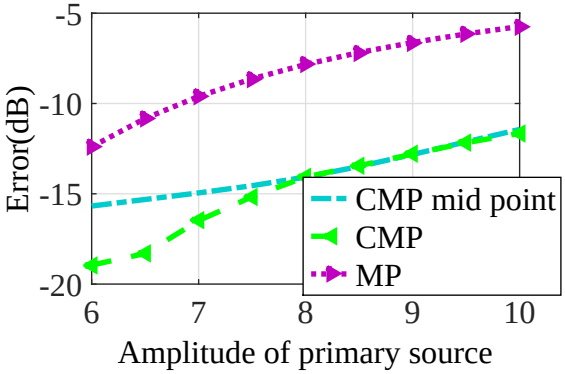


Figure 3.14: Reproduction error for the primary source with variable amplitude versus the amplitudes, when $p_{\max} = 0.3$.

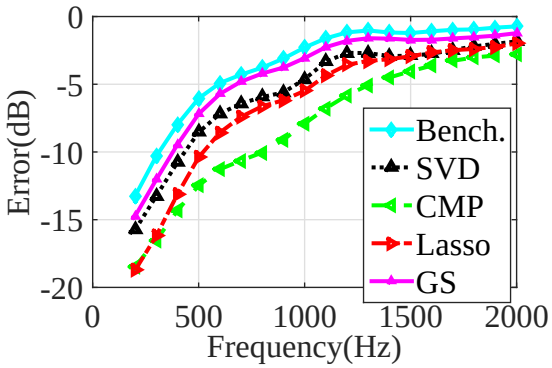


Figure 3.15: Reproduction error versus the frequency for two primary sources with variable frequency, when $p_{\max} = 0.3$.

The locations selected by the Lasso-, GS- and SVD-based algorithms do not depend on the amplitude of the primary source. However, the locations selected by the CMP-based algorithm depend on the amplitude of the primary source because the approximation errors $R^n\mathbf{p}^{\text{des}}$ (Algorithm 5) would change if $\mathbf{p}^{\text{des}}$ is scaled. To overcome this problem, one can also choose the locations for the mid-point of the amplitude range, and update only the complex excitation vector by Eq. (2.9) when the amplitude of the primary source changes. Another possible way is to use the baseline MP algorithm without considering power limitations for placement selection. Both methods are tested in Fig. 3.14, which shows the results when the amplitude of the primary source is between 6 and 10. The reproduction error is shown for the case when the loudspeaker locations are selected by CMP at the mid-point of the amplitude range (in this case 8), selected by CMP at each amplitude separately, and selected by the baseline MP algorithm. Based on this figure, the locations selected by CMP at the mid-point of the amplitude range work reasonably well at other amplitudes as well. They offer 3-5 dB lower error than the locations selected by the baseline MP, and essentially the same performance as the amplitude-specific CMP placement in the upper half of the amplitude range.

Therefore if the range of frequencies and amplitudes of the primary source is known, the locations of loudspeakers can be determined for the suitably chosen representative points within these ranges, such as midpoints. At operation time, the excitations of loudspeakers change according to Eq. (2.9) to minimize the reproduction error.

In the next experiment, it is assumed that two primary sources are located at $(0, 0, -8)$ m and $(0, \sqrt{28}, -6)$ m with equal complex amplitudes of $\sqrt{32}$. It is also assumed that the frequencies of the sources may be changing between 300 Hz and 1300 Hz. Fig. 3.15 shows the reproduction error versus frequency. As in the previous experiment, the locations of the loudspeakers are optimized for the mid-point frequency of 800 Hz, and then the complex excitation of the loudspeaker array are updated at each frequency based on Eq. (2.9). Again, this experiment indicates that the CMP-based method works better than all other placement methods even in the presence of multiple primary sources.

In the final experiment, the numerical robustness of the SFR system is investigated. Let $\text{cond}(\mathbf{A})$ be the condition number of matrix $\mathbf{A}$. For each method, the SFR system's condition number in dB [78], $10\log_{10}\Big(\text{cond}(\mathbf{G}^H\mathbf{G}+\gamma\mathbf{I})\Big)$, is shown in Fig. 3.16(a) for a single-tone primary source located at $(0, 0, -8)$ when $p_{\max} = 0.3$. The same parameter is shown in Fig. 3.16(b) for a primary source with frequency range between 100 and 2000 Hz, for which the loudspeaker locations are optimized at 800 Hz.

The lower the condition number, the more stable is the matrix inversion in Eq. (2.9). According to this figure, the maximum condition number, which can be thought of as a measure of system robustness [79], is comparable for all placement methods. Specifically, the condition number reaches up to about 30dB, or around 10^3, which means that the last 3 digits in the computation may be unreliable [80, 81]. The condition number of the

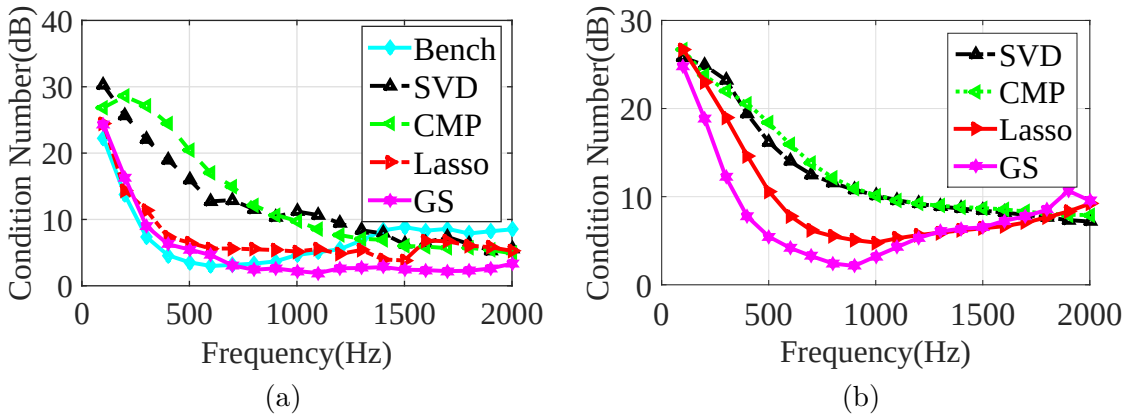


Figure 3.16: System condition number with $p_{\max} = 0.3$ for (a) locations optimized at each frequency, (b) locations optimized at $f = 800$ Hz.

method from [58] is generally lower than other placement methods because in [58], the loudspeaker locations are selected such that the resulting transfer impedance matrices are linearly independent. The condition numbers of SVD- and CMP-based placements are generally higher than other placement methods because, according to Fig. 3.8, the distance between loudspeakers becomes small at low $p_{\max}$, which makes the columns of the resulting ATF matrix more similar and increases the condition number.

At this point, it is also worth mentioning that the filters for finding loudspeaker excitations can be implemented in the time domain, as discussed in [78], rather than the frequency-domain approach discussed in this thesis. Such time-domain filters can be designed to be causal and stable, and employing such filters avoids the matrix inversion in Eq. (2.9), which reduces computational complexity and improves system robustness. For such systems, the results of the last experiment are irrelevant, but the results of the previous experiments are still expected to hold.

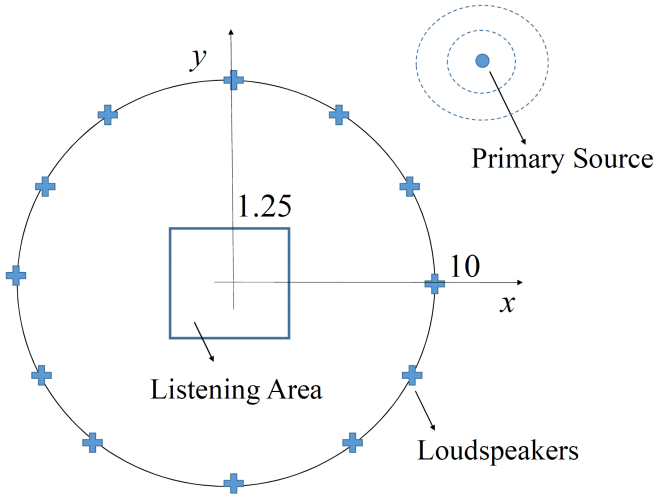


Figure 3.17: The 2-D configuration of interest for a sound field reproduction system.

3.6 2-D SFR configuration

In this section, the placement methods are tested in the context of 2-D SFR. The Loudspeaker Region (LR) is a circle centered at the origin with a radius of 10 m, and the primary source is at $(12, 12)$ m with an amplitude of 4. The listening area is a 2.5 m $\times$ 2.5 m square located at the center of the circle, and 121 sampling points for the desired field are distributed uniformly on a grid of 11×11. The reproduction error is sampled at $40,000$ points on a 200×200 uniform grid in the listening area. The number of loudspeaker is $N = 25$ and the number of candidate locations is $N_v = 400$, distributed uniformly around the circle. A similar configuration was used in 2-D SFR simulations in [15].

Four loudspeaker placement methods are compared against a circular benchmark (CB) configuration, in which 25 loudspeakers are placed uniformly around the circle. Fig. 3.18(a) shows the reproduction error vs. frequency at $p_{\max} = 0.5$, and Fig. 3.18(b) illustrates the error vs. $p_{\max}$ at $f = 700$ Hz. According to both figures, CMP outperforms the other methods. SVD-based placement works better than the Lasso-based one under power limitation (Fig. 3.18(a)), but Lasso improves as the maximum power increases. In fact, when $p_{\max}$ is high enough, the performance of Lasso-based placement approaches that of CMP (Fig. 3.18(b)). The performance of the GS-based placement method is similar to CB at low powers, and it improves by increasing power. All four placement methods provide better SFR performance to that of CB.

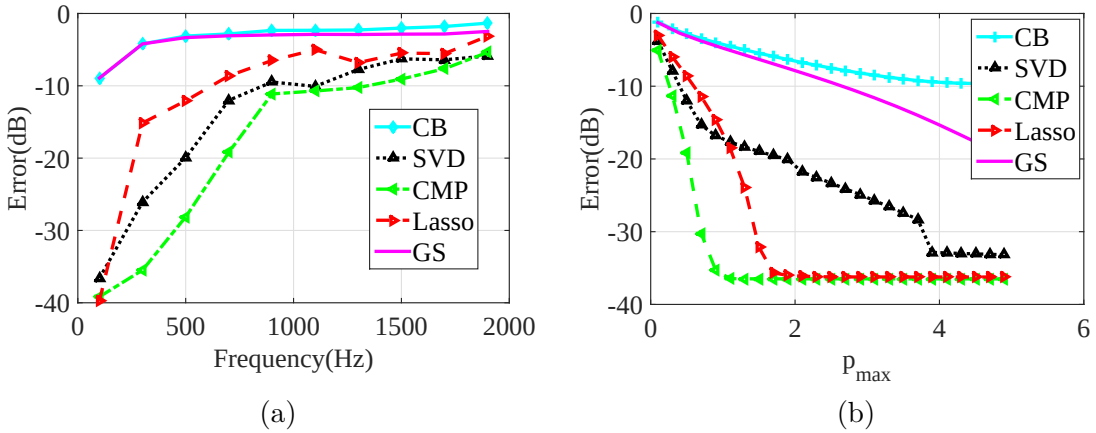


Figure 3.18: Reproduction error in 2-D SFR versus (a) frequency, at $p_{\max} = 0.5$, and (b) $p_{\max}$, for $f = 700$ Hz.

3.7 Conclusion

The SFR error is a non-convex function in terms of the locations of the secondary sources. In this chapter, first, a specific 3-D configuration of interest for SFR (Fig. 3.1) was used to compare different placement methods. Omni-directional loudspeakers were used in a planar array, and a cubic listening region was sampled over a fixed Cartesian grid. The system was tested across a frequency range that made the desired field oversampled at one extreme and spatially aliased at the other. The computational complexity of the placement algorithm showed that Lasso is the most complex among four placement algorithms.

The effect of power on the reproduction error and the reproduced field outside the listening area was investigated. The concept of ray-cut and the intuition behind the placement methods were explained along with the behavior of each algorithm. Lasso-based placement [15] does not take $p_{\max}$ into account in contrast to CMP and SVD algorithms, which produce different placements for different $p_{\max}$. The GS-based method selects the loudspeaker locations regardless of the desired vector and is only based on the system configuration. For this reason it has the poorest error performance among the placement methods. The CMP-based method selects locations by looking at the correlation between the the approximation error and the ATF from a particular location to the region of interest by taking the power constraint into account. The selected locations form different clusters around the most correlated ATFs with the error vector. The emergence of clusters is an interesting phenomenon in CMP with low power limit, which occurs due to the inability of a single selected ATF (dictionary member) to reduce the error sufficiently, so a similar ATF is selected in the subsequent iterations to reduce the error further in the same direction. Lasso effectively removes the correlated vectors and concentrates the energy in one of the ATFs. Therefore, the selected locations are more dispersed at lower frequencies. Simulation results indicated that the CMP-based placement outperforms the benchmark, the SVD-,

Lasso-, and GS- based placements in terms of the reproduction error in various scenarios. A feature of the CMP-based method is that it takes into account the available power of the loudspeaker array and the power of the primary source, which is part of the reason for its superior performance. The placement methods were also studied in an SFR system without the power constraint, and it was found that the uniform placement offers some advantages at low frequencies, below 300 Hz, while the performance of other algorithms is better than uniform placement at higher frequencies. The performance of Lasso-based placement method is better than other placement methods without power limitation.

Finally, a 2-D configuration was tested with the four placement methods, and the results have the same trend as in the 3-D SFR configuration. It means that the results are applicable to various SFR configurations, but the improvement level may be variable from one configuration to another.

In the next chapter, another system parameter, the radiation pattern of loudspeakers, will be optimized by the CMP algorithm, and the system performance will be evaluated when both placement and patterns of loudspeakers are optimized.

Chapter 4

Pattern Selection

4.1 Introduction

Another important factor in sound field reproduction is the radiation patterns of the loudspeakers. First, a novel algorithm is proposed to optimize the radiation patterns of loudspeakers assuming that they are distributed uniformly. The candidate radiation patterns are higher-order loudspeakers as opposed to omnidirectional patterns considered in the previous chapters. The order of all loudspeakers in the array is the same, but their harmonic coefficients (to be described later in this chapter) are optimized using a Constrained Matching Pursuit (CMP) algorithm. Therefore, the resulting radiation patterns may be different from one loudspeaker to the next, and the array contains a diverse set of loudspeaker patterns. This algorithm optimizes the radiation patterns based on the characteristics of the primary source. Hence, the resulting radiation patterns do not operate well if the features of the primary source change. To resolve this problem, a method is proposed to design loudspeaker patterns before the system operation based on possible features of the primary source during the system operation. It means that the patterns assigned to the loudspeakers remain fixed during system operation (static DoFs), and only their complex amplitudes will change as the features of the primary source (e.g. its frequency and location) change. The proposed algorithms are given in Section 4.2.

In Section 4.3, a new method is presented for jointly optimizing loudspeaker placement and radiation patterns. For a known primary source, the loudspeaker locations and patterns are jointly optimized by the CMP method. The excitations of the designed loudspeakers are then determined by direct approximation. Simulations for free space conditions show that the new method yields a lower reproduction error under a power constraint compared to other SFR methods for which only patterns or locations are optimized. The presented method can also work offline based on the features of the primary source. The numerical results are given in Section 4.4.1 for a primary source with fixed frequency and location and in Section 4.4.2 for a primary source with variable frequency and location.

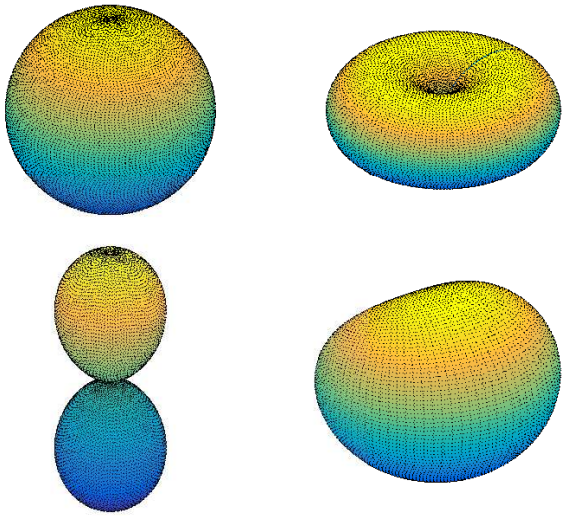


Figure 4.1: The 3-D radiation patterns of the first-order loudspeakers with $\mathbf{c}$ equal to $[1,0,0,0]$ (top left), $[0,1,0,0]$ (top right), $[0,0,1,0]$ (bottom left), and $[1/\sqrt{2},0,0,1/\sqrt{2}]$ (bottom right).

The contributions of this chapter are:

- Introducing a CMP-based pattern selection algorithm.
- Extending the pattern selection algorithm to work for a multi-frequency primary source whose exact location is not known in advance.
- Introducing a joint optimization of placement and pattern algorithm.
- Comparing placement, pattern, and joint optimization algorithms for single tone and multi-frequency primary sources.

4.2 Pattern optimization

4.2.1 Higher-Order Loudspeakers

2-D higher-order loudspeakers are first contemplated for SFR in [62]. Consider an array of L-th order 2-D loudspeakers. The pressure created by the n-th loudspeaker in the array is given by:

$$p(r,\phi,r_n,\phi_n) = \sum_{l=-L}^{L} w_{n,l} H_l(k\|\mathbf{x}-\mathbf{y}\|_2) e^{il\beta_n}, \tag{4.1}$$

where $\mathbf{y}=(r,\phi)$ and $\mathbf{x}=(r_n,\phi_n)$ are the 2-D locations of the observation point and the n-th loudspeaker, respectively, in polar coordinates, $H_l(\cdot)$ is the l-th cylindrical Hankel function, β_n is the polar angle between (r,ϕ) and (r_n,ϕ_n), $k=2\pi/\lambda$ is the wave number, λ is the

wavelength, and $w_{n,l}$'s are harmonic coefficients. $H_0(.)$ is the 2-D zeroth-order loudspeaker, i.e., the ATF of an omnidirectional loudspeaker.

In [62], a circular array of higher-order loudspeakers is employed to recreate a desired sound field on the surface of a disk. For this purpose, the coefficients $w_{n,l}$ are determined based on the HOA method to minimize the reproduction error. If there are N loudspeakers in the array, then using L-th order loudspeakers gives $(2L+1)N$ degrees of freedom, i.e. $(2L+1)N$ coefficients $w_{n,l}$ available for manipulation. Clearly, higher-order loudspeakers ($L > 0$) have the potential to make SFR more accurate compared to zeroth-order loudspeakers ($L = 0$), since there are more degrees of freedom in the system. In the method proposed in [62], for each change in conditions (e.g., change in the frequency of the desired field, or the location of the primary source), all $(2L+1)N$ coefficients need to be updated.

In this thesis we are mostly interested in 3-D SFR. Let $\mathbf{y} = (r, \theta, \phi)$ be the observation point in 3-D spherical coordinates. According to [47, 82], the pressure generated at that point by an arbitrary acoustic source located at the origin can be written as

$$p(r,\theta,\phi) = \sum_{l=0}^{L} \sum_{m_d=-l}^{l} C_{l,m_d} \cdot h_l(kr) \cdot Y_l^{m_d}(\theta,\phi), \tag{4.2}$$

where $h_l(kr)$ is the l-th order Hankel function, k is the wave number, $Y_l^{m_d}(\theta,\phi)$ is the spherical harmonic function of order l and degree m_d, and C_{l,m_d}'s are the harmonic coefficients. The term corresponding to $l = 0$ is omni-directional. For the far-field propagation ($kr \gg 1$):

$$h_l(kr) = \frac{e^{ikr}}{kr}(-i)^{l+1}. \tag{4.3}$$

To be consistent with the formulations and notations of the previuos chapters and Appendix C, we modify Eq. (4.2) for far-field propagation as follows:

$$\begin{aligned} p(f;r,\theta,\phi) &= s(f) \cdot \frac{e^{ikr}}{4\pi r} \cdot \sum_{l=0}^{L} \sum_{m_d=-l}^{l} C'_{l,m_d} Y'^{m_d}_l(\theta,\phi) \\ &= s(f) \cdot g(f;r) \cdot \mathcal{L}(f,\theta,\phi), \end{aligned} \tag{4.4}$$

where $s(f)$ is the complex amplitude of the higher-order loudspeaker in [Pa·m], $g(f;r)$ is the free space Green's function, and $\mathcal{L}(f,\theta,\phi) = \sum_{l=0}^{L}\sum_{m_d=-l}^{l} C'_{l,m_d} Y'^{m_d}_l(\theta,\phi)$ is the radiation pattern of the higher-order loudspeaker. In this equation $C'_{l,m_d} = \sqrt{4\pi} C_{l,m_d}(-i)^{l+1}/k$ are called the expansion coefficients of the higher-order loudspeaker in this thesis, and $Y'^{m_d}_l(\theta,\phi) = \sqrt{4\pi} Y_l^{m_d}(\theta,\phi)$. For a fair comparison between the performance of a system employing higher-order loudspeakers and a benchmark system employing omni-directional loudspeakers, these systems should use the same input power. The ATF of an omni-directional source is given by Eq. (1.4), which is equal to the $l = 0$ term in Eq. (4.4) with $C'_{0,0} = 1$. In order to compare the performance of the two systems under the same power con-

straint, the integrals of the radiation patterns of an omni-directional source ($\mathcal{L}(f,\theta,\phi)=1$) and a higher-order loudspeaker ($\mathcal{L}(f,\theta,\phi)=\sum_{l=1}^{L}\sum_{m_d=-l}^{l} C'_{l,m_d}Y'^{m_d}_{l}(\theta,\phi)$) should be equal over the unit sphere. This leads to $\|\mathbf{c}\|_2^2 \leq 1$ where $\mathbf{c}$ is a $(L+1)^2\times 1$ vector containing the expansion coefficients in increasing order of l and m_d. Under these conditions, the ATF corresponding to the term (l,m_d) is $g(f,r)Y'^{m_d}_{l}(\theta,\phi)$. Note that (r,θ,ϕ)'s are calculated with respect to the location of the loudspeakers.

These loudspeakers can be realized as combinations of monopole sources. For example, the first-order loudspeaker is a combination of one monopole and three dipoles, while each dipole is composed of two monopoles placed very close to each other [47]. Fig. 4.1 illustrates radiation patterns of several first-order loudspeakers ($L=1$) for various values of $C'_{m_d,l}$.

4.2.2 CMP-Based Pattern Selection

The CMP algorithm is used to seek the radiation patterns of loudspeakers while their locations are fixed. For this purpose, the expansion coefficients of higher order loudspeakers are determined to approximate the desired field using the CMP-based algorithm. Let $\mathbf{c}$ be the expansion coefficients of loudspeakers in increasing order of l and m_d:

$$\mathbf{c}=[C'_{0,0},C'_{-1,1},C'_{0,1},C'_{1,1},C'_{-2,2},...,C'_{L-1,L},C'_{L,L}]^T.$$

Also let $\mathbf{B}_n$ be a $M\times(L+1)^2$ matrix, corresponding to the n-th loudspeaker whose elements are the terms $h_l(kr)Y'^{m_d}_{l}(\theta,\phi)$ evaluated at virtual sampling points, in increasing order of l and m_d. This matrix contains the ATFs of each term in Eq. (4.2) from a given loudspeaker location $\mathbf{x}_n=(r_n,\theta_n,\phi_n)$ to each of the virtual sampling points:

$$\mathbf{B}_n=\begin{bmatrix} h_0(kr_{1,n})Y'^{0}_{0}(\theta_{1,n},\phi_{1,n}) & h_1(kr_{1,n})Y'^{1}_{-1}(\theta_{1,n},\phi_{1,n}) & \cdots & h_L(kr_{1,n})Y'^{L}_{+L}(\theta_{1,n},\phi_{1,n}) \\ h_0(kr_{2,n})Y'^{0}_{0}(\theta_{2,n},\phi_{2,n}) & h_1(kr_{2,n})Y'^{1}_{-1}(\theta_{2,n},\phi_{2,n}) & \cdots & h_L(kr_{2,n})Y'^{L}_{+L}(\theta_{2,n},\phi_{2,n}) \\ \vdots & \vdots & \ddots & \vdots \\ h_0(kr_{M,n})Y'^{0}_{0}(\theta_{M,n},\phi_{M,n}) & h_1(kr_{M,n})Y'^{1}_{-1}(\theta_{M,n},\phi_{M,n}) & \cdots & h_L(kr_{M,n})Y'^{L}_{+L}(\theta_{M,n},\phi_{M,n}) \end{bmatrix}. \tag{4.5}$$

Here, $(r_{m,n},\theta_{m,n},\phi_{m,n})$ is the spherical coordinate of the m-th sampling point with respect to the n-th loudspeaker location.

To find the radiation patterns of loudspeakers through the CMP-based algorithm, the dictionary members are formed as follows: First, matrix $\mathbf{B}_n$ is formed for each loudspeaker location, and its columns are placed in the set $\mathfrak{D}$ as dictionary members. Since each $\mathbf{B}$ has $(L+1)^2$ columns and there are N such matrices, set $\mathfrak{D}$ initially contains $N(L+1)^2$ members.

In the CMP-based algorithm, each dictionary member is identified with a possible higher-order pattern at a loudspeaker location. The algorithm, first, finds the most correlated dictionary member with the error field, which is initialized as the desired vector.

Algorithm 6 Loudspeaker pattern optimization

Input: $\mathfrak{D}$ ▷ dictionary
Input: $\mathbf{p}^{\text{des}}$ ▷ desired vector
Input: N ▷ number of loudspeakers in the array
Output: $\{\mathbf{c}^n\}$ ▷ loudspeaker pattern coefficients

1: Set $R^1(\mathbf{p}^{\text{des}}) = \mathbf{p}^{\text{des}}$.
2: **for** $n = 1$ to N **do**
3: Find $\mathbf{d} \in \mathfrak{D}$ that is most correlated with $R^n\left(\mathbf{p}^{\text{des}}\right)$.
4: Find the loudspeaker location that $\mathbf{d}$ corresponds to.
5: Extract all members of $\mathfrak{D}$ that correspond to this location and place them into set $\mathfrak{B}$.
6: Apply the CMP-based algorithm (refer to Section 2.6.2) with $R^n(\mathbf{p}^{\text{des}})$ as the desired vector, $\mathfrak{B}$ as the dictionary, maximum power equal to 1, and $(L+1)^2$ as the number of iterations.
7: The output of this algorithm will be the coefficient vector $\mathbf{c}^n$, containing the expansion coefficients of the n-th loudspeaker, and selected vectors $\{\mathbf{b}^{(j)}\}$ from $\mathfrak{B}$.
8: Set $\widehat{\mathbf{p}}^n = \sum_{j=1}^{(L+1)^2} c_j^n \mathbf{b}^{(j)}$, which is the ATF of the n-th selected loudspeaker.
9: Set $\mathfrak{D} = \mathfrak{D} \setminus \mathfrak{B}$.
10: To calculate the new error vector, find the corresponding coefficient to $\mathbf{b}^{(j)}$ as $\alpha^{(j)}$, by solving Eq.(2.42), where $p_n = p_{\max}/N$.
11: Compute the new error vector:

$$R^{n+1}(\mathbf{p}^{\text{des}}) = R^n(\mathbf{p}^{des}) - \alpha^{(j)}\widehat{\mathbf{p}}^n. \tag{4.6}$$

12: **end for**
13: **return** $\{\mathbf{c}^n\}$.

The selected dictionary member corresponds to one of the loudspeaker locations. Matrix $\mathbf{B}$ corresponding to the selected loudspeaker is considered as a new dictionary. In order to find the expansion coefficients of the selected loudspeakers, the CMP-based algorithm is applied with matrix $\mathbf{B}$ as dictionary, with $p_{\max} = 1$ as power constraint, and the error vector as the input vector. After that the expansion coefficients are calculated, the error vector is updated, and this process continues until the expansion coefficients of all loudspeakers are determined. The details of the CMP-based pattern selection algorithm is given in Alg. 6.

In fact, this algorithm is composed of two CMP algorithms. Step 6 is the inner part which finds the expansion coefficients of loudspeakers by the CMP method. As mentioned earlier the maximum power of the expansion coefficients is limited to 1. In the CMP algorithm this power is distributed uniformly among the iterations. It means that for the n_t-th iteration, $p_{n_t} = \frac{1}{(L+1)^2}$. This step finds the harmonic coefficients of the selected loudspeaker (or the ATF of the selected loudspeaker), and returns the results to the outer CMP algorithm. Then, in Step 10, the coefficients corresponding to the selected loudspeaker (ATF) are obtained using the CMP equations in Section 2.6.2 in order to calculate the new error vector in the outer loop.

Algorithm 7 Extended loudspeaker pattern optimization

Input: $\{f_1, f_2, ..., f_Y\}$ ▷ frequencies of interest
Input: $\{\mathbf{x}_1^t, \mathbf{x}_2^t, ..., \mathbf{x}_W^t\}$ ▷ possible locations of the primary sources
Input: $\{n_{w,y}\}$ ▷ number of loudspeakers per frequency-location pair
Output: $\{\mathbf{c}^n\}$ ▷ loudspeaker pattern coefficients

1: **for** $y = Y$ to 1 **do**
2: **for** $w = 1$ to W **do**
3: Compute the sound pressure at frequency f_y from a source at $\mathbf{x}_w^t$ to each of the M sampling points, and store the results in $\mathbf{p}^{\text{des}}$.
4: Compute the matrix $\mathbf{B}$ at frequency f_y for all remaining loudspeakers to all M sampling points, and store its columns as dictionary members in $\mathfrak{D}$.
5: Run Algorithm 6 with $\mathfrak{D}$ as the dictionary, $\mathbf{p}^{\text{des}}$ as the desired vector, and $n_{w,y}$ as the number of loudspeakers.
6: Store the resulting pattern coefficients $\mathbf{c}^n$'s for the selected loudspeakers.
7: Remove the loudspeakers selected at this iteration from further consideration.
8: **end for**
9: **end for**
10: **return** $\{\mathbf{c}^n\}$

Algorithm 6 is the basic building block for loudspeaker pattern design. It assumes that the location of the primary source is fixed and known in advance. This algorithm is generalized in the next section to account for the whole frequency range of the primary source and its various possible locations.

4.2.3 Multi-frequency pattern selection

In order to extend the loudspeaker pattern design algorithm to a more general setting, we assume that although the exact parameters of the primary source are not available, some side information is known about the primary source. Suppose that the frequency range and the possible locations of the primary source are known in advance. Using this information, the radiation patterns of loudspeakers can be designed during the design phase as static DoFs, and only their complex amplitudes are optimized based on the exact locations of the primary source.

In the extended algorithm, to account for various possible primary source locations, we distribute W points $\{\mathbf{x}_1^t, \mathbf{x}_2^t, ..., \mathbf{x}_W^t\}$ uniformly across the possible locations for the primary source, as representative locations. In addition, to account for the frequency band of interest, we distribute Y frequency points $f_1 < f_2 < ... < f_Y$ across this range. The idea is simple - select a different group of primary sources for each pair $(\mathbf{x}_w^t, f_y)$ and optimize their patterns for that position-frequency pair using Algorithm 6. In particular, let $n_{w,y}$ be the number of loudspeakers allocated to the pair $(\mathbf{x}_w^t, f_y)$, so that $\sum_{w=1}^{W} \sum_{y=1}^{Y} n_{w,y} = N$. Pattern design is then performed via Algorithm 7.

Note that the output of Step 5 in Algorithm 7 is not only the loudspeaker pattern coefficients, but also the indicies (or, equivalently, locations) of the selected loudspeakers. In the first iteration, $n_{1,Y}$ out of N loudspeakers are selected and their patterns designed; in the second iteration $n_{2,Y}$ out of the remaining $N - n_{1,Y}$ loudspeakers are selected and their patterns designed, and so on. In the last iteration, since only $n_{W,1}$ loudspeakers remain, no loudspeaker selection takes place and the only step is pattern design.

Therefore, in this algorithm, in contrast to Algorithm 6, the radiation patterns of the loudspeakers are designed ahead of time before system operation. After designing the radiation patterns, they will be placed in the loudspeaker array, and during the system operation only the complex amplitudes of loudspeakers change to minimize the reproduction error. Algorithm 6 takes the radiation patterns of loudspeakers as dynamic DoFs while Algorithm 7 regards them as static DoFs. It can be concluded that the performance of Algorithm 6 would be better than that of Algorithm 7 because the number of dynamic DoFs is much larger. However, the complexity of Algorithm 7 is much lower than that of Algorithm 6 which makes it practical to work in real-time scenarios.

The radiation patterns designed in Algorithm 7 can be implemented by combining simple (and sufficiently small) monopole loudspeakers in a specific configuration array [47] (chapter 6, page 198). The implementation can be considered a microelectronics design challenge, where printed loudspeakers and their array weights are integrated into a single device. Another way to implement higher order loudspeakers is to approximate the desired pattern by off-the-shelf loudspeakers in a specific configuration. However, the issue of implementation is beyond the scope of this thesis. These loudspeakers are built and tested for 2-D SFR in [83–85].

4.3 Joint optimization of placement and patterns

4.3.1 CMP-based joint optimization

Locations and radiation patterns of an SFR system were optimized in Sections 2.6.4 and 4.2.2 respectively using the CMP algorithm. In this section, a new algorithm is proposed to jointly optimize both locations and radiation patterns of the loudspeakers based on the CMP algorithm.

As in placement methods, $N_v \gg N$ locations are placed densely on the LR. For each candidate location, as in the pattern selection algorithm, a higher order loudspeaker is considered. Therefore, there are $(L+1)^2$ dictionary members for each loudspeaker locations, and the total dictionary members are:

$$\mathfrak{D} = \{\mathbf{b}_{1,1}, \mathbf{b}_{2,1}, ..., \mathbf{b}_{(L+1)^2,1}, \mathbf{b}_{1,2}, \mathbf{b}_{2,2}, ..., \mathbf{b}_{(L+1)^2,2}, ..., \mathbf{b}_{1,N_v}, \mathbf{b}_{2,N_v}, ..., \mathbf{b}_{(L+1)^2,N_v}\} \quad (4.7)$$

Table 4.1: Number of dictionary members.

(N, N_v, L)	Placement	Pattern	Joint
$(25, 400, 2)$	400	225	3600
$(25, 100, 5)$	100	900	3600
$(100, 900, 5)$	900	3600	32400
$(25, 400, 5)$	400	900	14400

where L is the loudspeaker order, $\mathbf{b}_{j,n_v}$ is the j-th column of $\mathbf{B}_{n_v}$, and $\mathbf{B}_{n_v}$ is formed based on Eq. (4.5) considering that the location of the loudspeaker is at the n_v-th candidate location. In this algorithm, the total number of dictionary members is $N_v(L+1)^2$, which can be large number. For example for $N_v = 900$ candidate locations and $L = 5$-th order loudspeakers, the number of dictionary members is 32400, which is much larger than the number of sampling points M (or size of vectors). From [64], a larger dictionary results in better representation. It means that the performance of the proposed method is expected to be better in comparison with the pattern selection algorithm with $N(L+1)^2$ and placement methods with N_v dictionary members. The number of dictionary members for different values of N, N_v, and L is given in Table 4.1 for three proposed algorithms.

It should be noted that, although the number of dictionary members increases by increasing N_v for the placement algorithm, the performance will not improve considerably. The reason is that the dictionary members are the ATFs of the candidate locations. For a fixed LR, by increasing the number of candidate locations (dictionary members) the distance between candidate locations decreases, and the correlation between the ATFs of the neighboring location increases. It implies that at a specific frequency, increasing the number of candidate locations does not add independent members to the dictionary anymore. However, increasing the number of dictionary members in the pattern and joint optimization algorithms by increasing L improves the system performance, because increasing the order of the loudspeakers increases the number of spherical harmonics which are spatially orthogonal. Hence, the system performance enhances since the number of independent dictionary members increases.

The details of the joint optimization algorithm are given in Algorithm 8.

Algorithm 8 Loudspeaker pattern and location optimization

Input: $\mathfrak{D}$ ▷ dictionary

Input: $\mathbf{p}^{\text{des}}$ ▷ desired vector

Input: $\mathbf{x}_j^v$ ▷ coordinate of j-th candidate location

Input: N ▷ number of loudspeakers

Output: A ▷ selected loudspeaker locations

Output: $\{\mathbf{c}^n\}$ ▷ inner coefficients of loudspeakers

Set $R^1\mathbf{p}^{\text{des}} = \mathbf{p}^{\text{des}}$

Set $A^{joint} = \{\varnothing\}$

for $n = 1$ to N **do**

  Find $\mathbf{d} \in \mathfrak{D}$ that is most correlated with $R^n\left(\mathbf{p}^{\text{des}}\right)$.

  Find the loudspeaker location that $\mathbf{d}$ corresponds to.

  Set $A^{joint} = A^{joint} \cup \{\mathbf{x^v}_n\}$, where $\mathbf{x^v}_n$ is the n-th selected location.

  Extract all members of $\mathfrak{D}$ that correspond to this location and place them into set $\mathfrak{B}$.

  Apply the CMP-based algorithm with $R^n(\mathbf{p}^{\text{des}})$ as the desired vector, $\mathfrak{B}$ as the dictionary, maximum power equal to 1, and $(L+1)^2$ as the number of iterations.

  The output of this algorithm will be the coefficient vector $\mathbf{c}^n = [c_1^n, c_2^n, ..., c_{(L+1)^2}^n]$, containing the harmonic coefficients of the n-th loudspeaker, and selected vectors $\{\mathbf{b}^{(j)}\}$ from $\mathfrak{B}$.

  Set $\widehat{\mathbf{p}}^n = \sum_{j=1}^{(L+1)^2} c_j^n \mathbf{b}^{(j)}$ as the ATF of the n-th selected loudspeaker.

  Set $\mathfrak{D} = \mathfrak{D} \setminus \mathfrak{B}$.

  To calculate the new error vector, find the corresponding coefficient to $\mathbf{b}^{(j)}$ as $\alpha^{(j)}$, by solving Eq. 2.42, where $p_n = p_{\max}/N$.

  Compute the new error vector:

$$R^{n+1}(\mathbf{p}^{\text{des}}) = R^n(\mathbf{p}^{\text{des}}) - \alpha^{(j)}\widehat{\mathbf{p}}^n.$$

end for

return $\{\mathbf{c}^1, .., \mathbf{c}^N\}$.

return A^{joint}.

In this algorithm, first the most correlated vector with the desired field is found. This vector corresponds to the location of a loudspeaker. Second, in order to find the expansion coefficients of the selected loudspeaker, the CMP based method is applied with the dictionary members corresponding to the selected loudspeaker (location) and the desired field as inputs. Third, the expansion coefficients of the selected loudspeaker are calculated via the CMP algorithm and its corresponding ATF is computed. Fourth, the updated error vector is calculated, and the selected location and its corresponding dictionary members are removed from the dictionary. This process continues until the N locations out of N_v candidate locations are found.

4.3.2 Multi-frequency joint optimization algorithm

It should be noted that Algorithm 8 optimizes the radiation patterns and locations of the loudspeakers when the frequency and location of the primary source are fixed. The joint optimization consists of finding, at each iteration, the term $h_l(kr){Y'}^{l}_{m_d}(\theta,\phi)$ (across all remaining candidate locations and all patterns) that is most correlated with the current approximation error, and running CMP over the corresponding sub-dictionary.

This algorithm can also be extended through Algorithm 7 to optimize the patterns and locations based on the features of the primary source (its frequency range and possible locations) before the system operation as static DoFs and update only the complex amplitudes of loudspeakers during the system operation as dynamic DoFs. For this purpose, in Step 5 of Algorithm 7, Algorithm 8 should be employed, and the locations selected by the joint optimization algorithm should be returned as output as well.

4.4 Experimental results

The error performance of the configuration shown in Fig. 3.1 is examined for the following four systems: *System 1 (S1)* is a benchmark system with N omni-directional loudspeakers placed uniformly on the LR. *System 2 (S2)* consists of N omni-directional loudspeakers whose locations on the LR are determined using Algorithm 5. *System 3 (S3)* consists of N higher-order loudspeakers placed uniformly on the LR, where only the radiation patterns are designed by Algorithm 6. *System 4* constis of N loudspeakers whose radiation patterns and locations are optimized by Algorithm 8.

First, the numerical simulations are performed assuming that the location and frequency of the primary source are fixed in Section 4.4.1. Then the system performance is tested for a primary source with variable frequency and location of the primary source in Section 4.4.2.

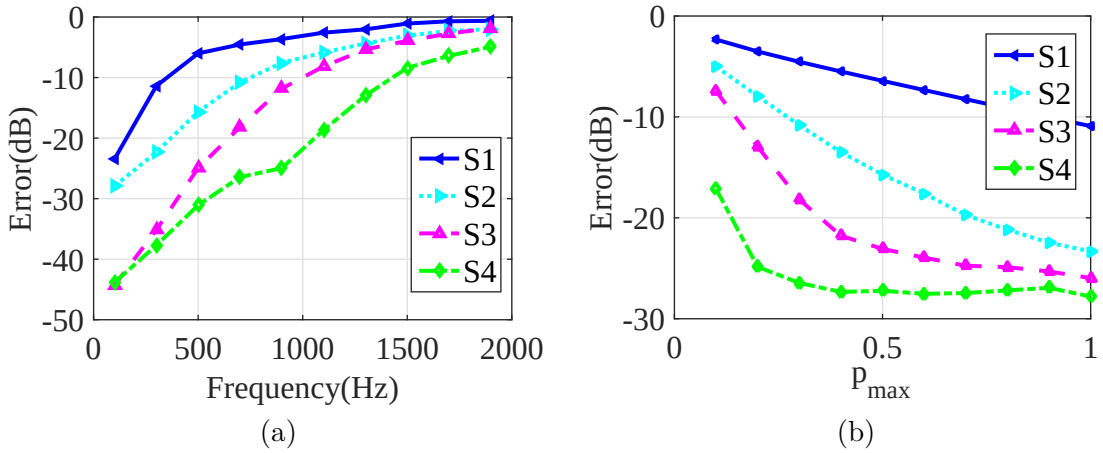


Figure 4.2: Comparison of the error performance among the proposed algorithms for (a) $p_{\max} = 0.3$ over the frequency range, (b) $f = 700$ Hz by changing the maximum normalized power.

4.4.1 Single tone frequency

In these experiments the frequency range of the primary source is between 200 Hz and 2000 Hz. In the experiments of this chapter (in contrast to the previous chapter), the sampling points are arranged around the cubic listening area. The reason is that based on the KH integral, reconstruction of a sound field on the surface of an enclosed volume results in sound field reproduction inside that volume. The distance between the neighboring sampling points is 25 cm, and the total number of sampling points is 98. In this situation, the sound field is under-sampled for frequencies greater than 686 Hz. However, the error is evaluated on 8000 points inside the cubic listening area which are distributed uniformly with a distance of 5 cm, so the results are reliable for frequencies less than 3 KHz since the error field is over-sampled. Unless otherwise stated, the number of candidate locations is 100, and the order of loudspeakers is $L = 5$.

The error performance as a function of frequency and the maximum normalized power is shown in Fig. 4.2. Fig. 4.2(a) compares the performance of the four systems for $p_{\max} = 0.3$ over the frequency range, and Fig. 4.2(b) compares the error at $f = 700$ Hz by varying the maximum normalized power. According to Fig. 4.2(a), at $f = 900$ Hz the performance of the joint optimization algorithm (System 4) is better than System 1 by 20 dB, System 2 by 17 dB, and System 3 by 13 dB. In addition from Fig. 4.2(b), the performance of System 4 (joint optimization algorithm) for $p_{\max} = 0.2$ is the same as the System 3 (pattern algorithm) for $p_{\max} = 0.8$, System 2 (placement algorithm) for $p_{\max} = 1$, and System 1 (benchmark) for $p_{\max}$ greater than 1. It means that the joint optimization method results in the same performance in the listening area while it reduces power wastage outside of the listening area.

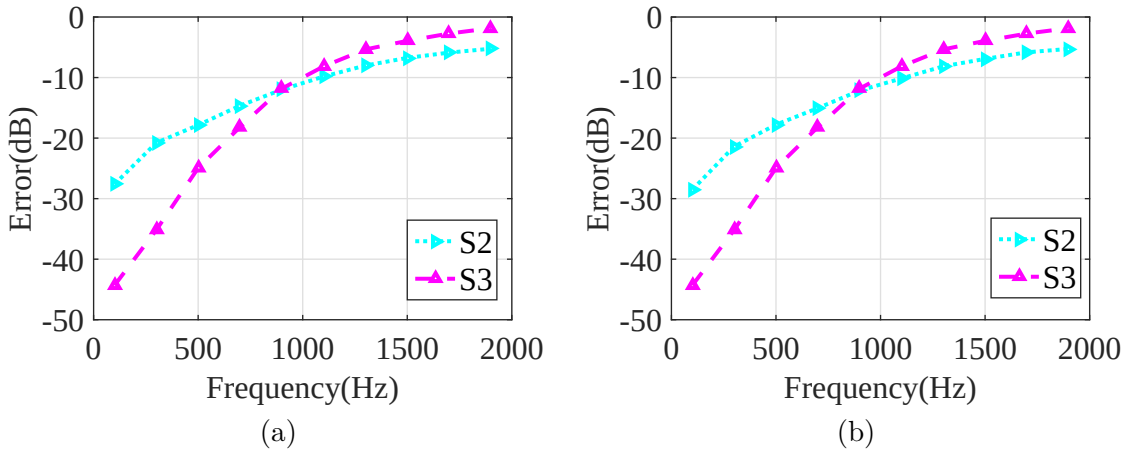


Figure 4.3: Error performance of the placement and pattern algorithms for (a) $N_v = 400$, and $N_v = 900$.

Furthermore, from Fig. 4.2, the performance of the pattern selection algorithm is better than the placement-only algorithm because the number of independent dictionary members in the pattern selection algorithm is greater than those of the placement-only method. Increasing the number of candidate locations increases the number of independent dictionary members at higher frequencies, so the performance of the placement method would be better at higher frequencies. As explained earlier, if the number of candidate locations increases further, the performance of the placement algorithm does not improve any more. To show this effect, the error performance of the placement algorithm is shown in Fig. 4.3 for $N_v = 400$ and $N_v = 900$, and it is compared against the error performance of the pattern selection algorithm. Comparing this figure with Fig. 4.2(a) reveals that the performance of the placement-only algorithm improves at higher frequencies when the number of candidate locations changes from 100 from 400, while it does not get better when the number of candidate locations increases from 400 to 900.

In the next experiment the error performance is shown in terms of the length of the cubic listening area and frequency. In this test, the number of sampling points is kept fixed by increasing the length of the listening area. Therefore, the distance between the sampling points increases, and the frequency for which aliasing occurs decreases. Therefore, the system performance degrades by increasing the length of the cubic listening area. To compare the error performance, these graphs are shown in Fig. 4.4 for frequency 1500 Hz versus the length of the cubic listening area, and versus frequency when the length of listening area is 0.5 m. As shown in Fig. 4.4(b), the error performance of System 3 and 4 are close together at lower frequencies since the Nyquist distance between loudspeakers are large (in meter) at lower frequencies. It implies that the locations obtained by System 4 are close to that of System 3 in which the distance between the loudspeakers is large.

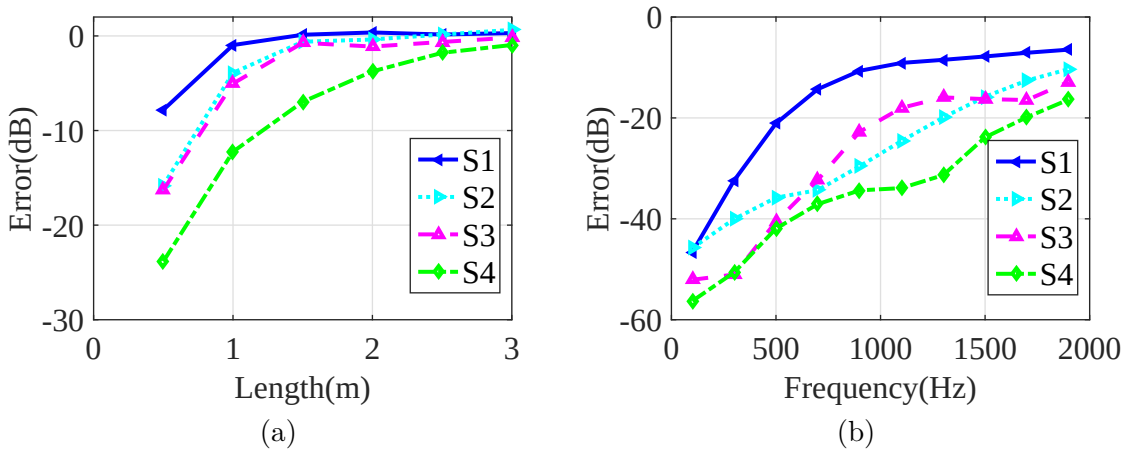
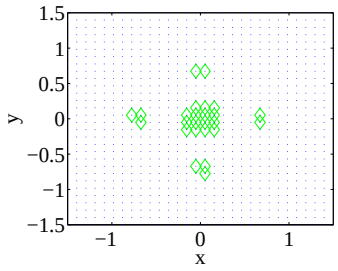
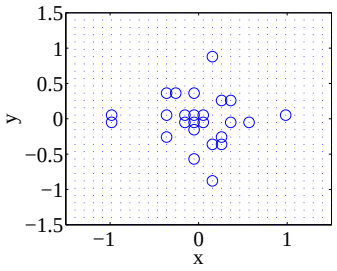


Figure 4.4: Comparison among the proposed algorithms versus (a) Length of the listening area at $f = 1500$ Hz and $p_{\max} = 0.5$, and (b) Frequency for 0.5 m×0.5 m×0.5 m cubic listening area at $p_{\max} = 0.5$

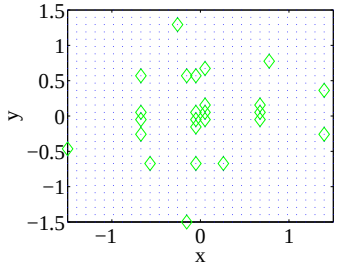
The locations of the loudspeakers selected by System 2 (placement-only method) and 4 (joint optimization method) are shown for $N_v = 900$, frequencies 600 Hz and 1600, and maximum normalized power 0.1 and 1 in Fig. 4.5. It can be concluded that, at lower frequencies, the distribution obtained by the joint optimization algorithm is more dispersed than that of the placement-only method. The reason is that in the joint optimization algorithm after selecting the n-th location, the CMP algorithm is applied to find the radiation patterns of the selected loudspeaker. The ATF of the selected loudspeaker is a linear combination of a set of dictionary members and it can better represent the error vector in comparison with employing only omni-directional loudspeakers. Hence, the selected ATF can explain the error vector better, so the selected ATF at the next iteration will be less similar to the previous one which results in increasing distance between the selected loudspeakers. At higher frequencies, both methods cluster the loudspeakers around the ray-cut, because the neighboring ATFs are not similar at these frequencies, and the optimum locations are clustered around the ray-cut as discussed in Section 3.4.



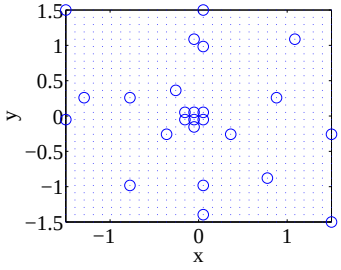
(a) $(600, 0.1)$



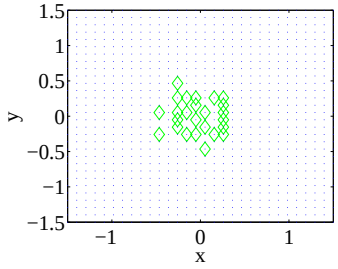
(b) $(600, 0.1)$



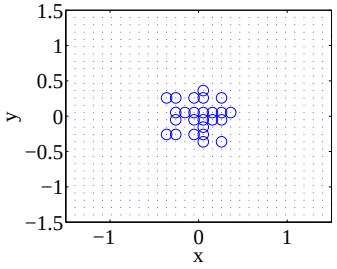
(c) $(600, 1)$



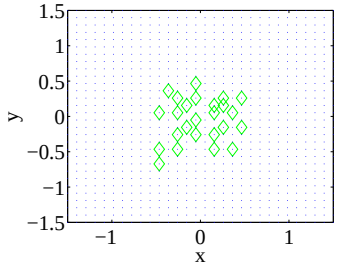
(d) $(600, 1)$



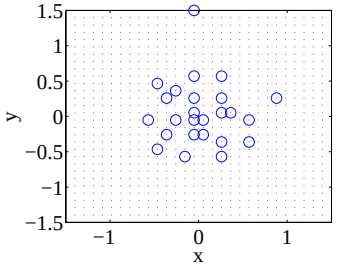
(e) $(1600, 0.1)$



(f) $(1600, 0.1)$



(g) $(1600, 1)$



(h) $(1600, 1)$

Figure 4.5: Loudspeaker placement produced by System 2 (left column) and System 4 (right column) for $f \in \{600, 1600\}$ Hz and $p_{\max} \in \{0.1, 1\}$.

The performance of Systems 3 and 4 for different loudspeaker orders is evaluated in Fig. 4.6 for $p_{\max} = 0.1$. According to this figure, the higher order results in better performance because the number of orthogonal dictionary members increases. The higher order loudspeakers, on the other hand, are more complicated for implementation. The perfor-

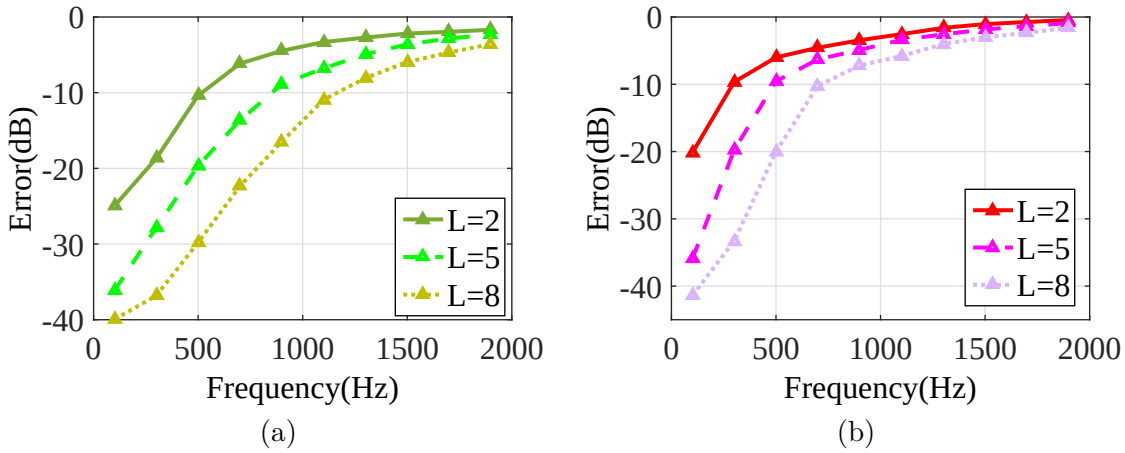


Figure 4.6: Error performance versus the order of loudspeakers in (a) System 3 and (b) System 4, for $p_{\max} = 0.1$

mance of Systems 3 and 4 for different loudspeakers orders is evaluated in Fig. 4.6 for $p_{\max} = 0.1$. According to this figure, the higher order results in better performance because the number of orthogonal dictionary members increases. The higher order loudspeakers, on the other hand, are more complicated for implementation.

The optimized far-field radiation patterns of 4 loudspeakers located at $(0, 0)$, $(1.5, 1.5)$, $(0, 1.5)$, and $(1.5, 0)$ are shown in Figs. 4.7(a), 4.7(b), 4.7(c), 4.7(d) respectively for $z > 0$. These radiation patterns are obtained for $f = 800$ and $p_{\max} = 0.1$ using the pattern selection algorithm (System 3). According to these figures, the main lobes of the optimized loudspeakers are in the direction of the listening area which results in good error performance under constrained power since the most portion of the power is focused in the listening area.

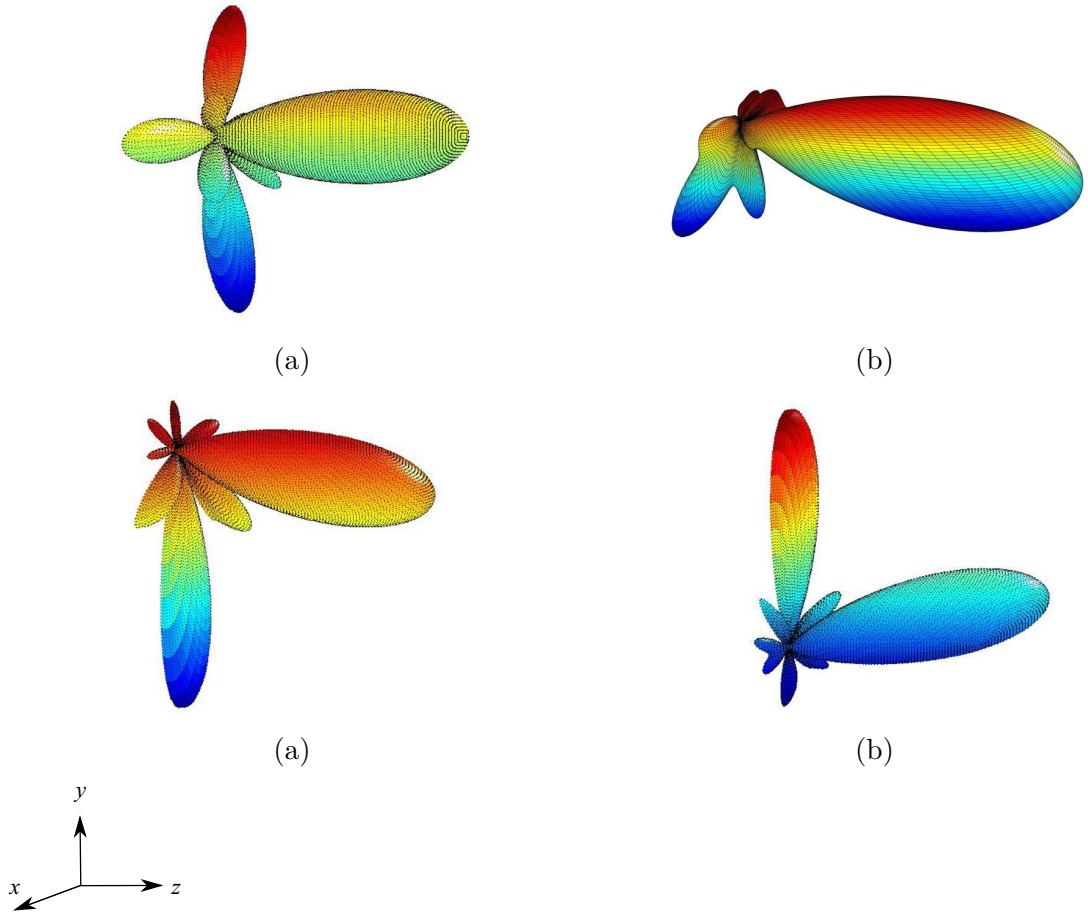


Figure 4.7: Optimized radiation patterns of loudspeakers located at (a) $(0, 0)$, (b) $(1.5, 1.5)$, (c) $(0, 1.5)$, and (d) $(1.5, 0)$ in System 3.

The radiation patterns of the loudspeaker array for the four systems are given in Fig. 4.8 for $f = 600$ Hz, $p_{\max} = 0.5$, $L = 5$, and $N_v = 900$. These patterns can explain why these systems have different performance under power limitation. According to this figure, half of the power flows into half-space $z < 0$ for Systems 1 and 2 because of using omni-directional loudspeakers which radiate equal power in all directions. However, for Systems 3 and 4, where the radiation patterns are optimized, a significant part of the power is concentrated in half-space $z > 0$ in which the listening area is located.

On the other hand, comparison between the radiation patterns of Systems 1 and 2 and between Systems 3 and 4 show the benefit of the placement method. In the benchmark configuration, although the main lobe is toward the listening area, there exists strong side-lobes in other directions. The number of side lobes and their corresponding power significantly decreases in the placement method. Also, System 3 (pattern selection algorithm) has a main lobe in the direction of the listening area with a large half power beam-width, which results in power wastage around the listening area. However, the half power beam width of System 4 is smaller than that of System 3.

(a) (b)

(a) (b)

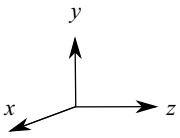


Figure 4.8: Radiation patterns of (a) System 1, (b) System 2, (c) System 3, (d) System 4 for $f = 600$ Hz and $p_{\max} = 0.5$.

In the last experiment, the reproduction error is tabulated in Table 4.2 for different locations of the primary source for $f = 1000$ Hz and $p_{\max} = 0.5$. According to this table for all cases, the joint optimization algorithm outperforms the other systems by more than 12 dB.

4.4.2 Multi-frequency primary source

The performance of the four systems when the location and frequency of the primary source are unknown is investigated in this section by applying Algorithm 7. In this algorithm, the locations and patterns of the loudspeakers are taken as static DoFs, and they are optimized before system operation. In our experiments we assume that the frequency range of the primary source is less than 2000 Hz.

Table 4.2: Reproduction error for different locations of primary source

Location	System 1	System 2	System 3	System 4
$(1.94, 0, -7.76))$	-1.85	-8.16	-9.95	-24.22
$(0, -2.8, -7.49)$	-2.39	-8.11	-12.54	-24.86
$(2.73, 1.82, -7.2)$	-1.35	-8.27	-9.98	-24.83
$(3.26, 3.26, -6.53)$	-2.97	-8.28	-12.57	-24.12
$(4.11, -4.11, -5.48)$	-3.63	-8.60	-13.53	-25.49

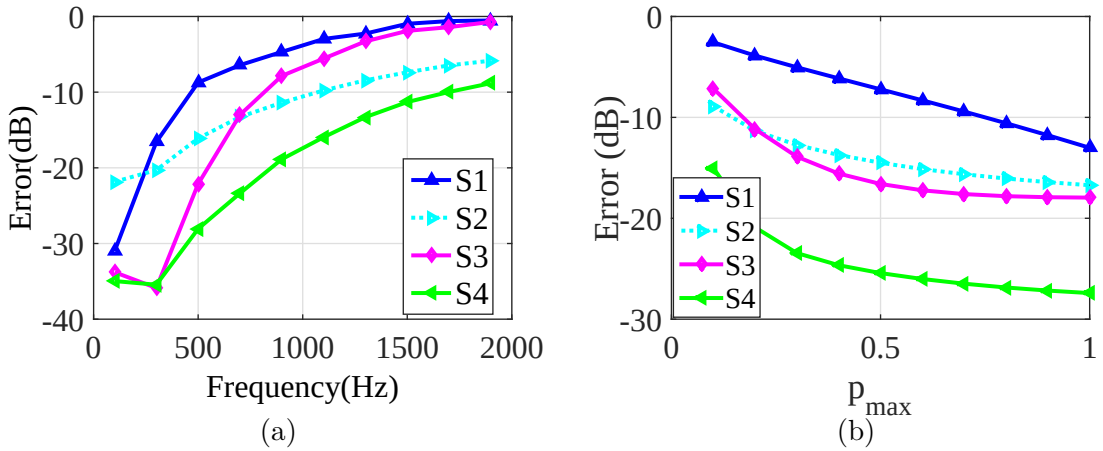


Figure 4.9: Reproduction error in terms of (a) frequency for $p_{\max} = 0.5$, (b) maximum normalized power at $f = 600$ Hz, when the location of primary source is known while its frequency is unknown in the design phase.

First, the reproduction error is examined at different frequencies with $p_{\max} = 0.3$. The location of the primary source is fixed at $(0, 0, -8)$, and the frequency for which the loudspeakers are optimized is sampled at $\{100, 300, 600, 1000, 1500\}$ Hz. Therefore, the patterns and (or) locations of the 5 loudspeakers are optimized for each frequency. After the design phase, the error performance is evaluated in Fig. 4.9(a) when $p_{\max} = 0.5$ and the frequency of the primary source is changed between 100 and 2000 Hz, and in Fig. 4.9(b) for $f = 600$ Hz and $p_{\max}$ between 0.1 and 1. According to this figure, the joint optimization algorithm outperforms the other systems across the whole frequency range, with error performance gains between 3–10 dB relative to the next best system at mid and higher frequencies.

In the next experiment, the locations of the primary source range between $(-4, \pm 0.5, -8)$ and $(+4, \pm 0.5, -8)$. In this experiment, the frequency for which the loudspeakers are optimized is sampled only at $\{100, 1500\}$ and the possible locations of the primary source are $(0, 0, -8)$, $(0, 0, -4)$, $(0, 0, +4)$, $(0, 0, -2)$, and $(0, 0, +2)$. Therefore two or three loudspeakers are designed for each frequency bin and sampled location. Again, after the design phase, the frequency and location of the primary source are changed and the reproduction error

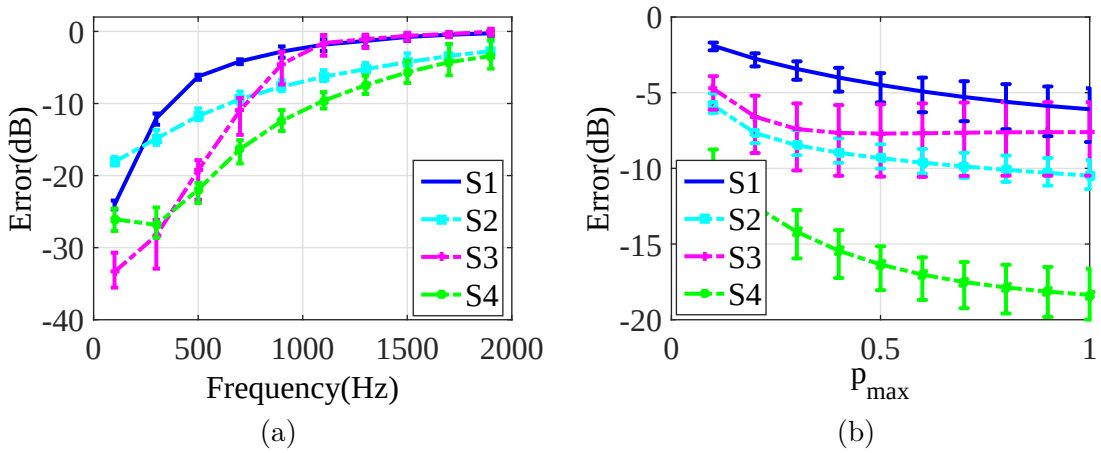


Figure 4.10: Error performance for (a) $p_{\max} = 0.3$ across the frequency (b) $f = 800$ in terms of $p_{\max}$, when the exact location and frequency of the primary source is not given in the design phase.

is shown in Fig. 4.10(a). In this figure, each curve shows the average reproduction error at the corresponding frequency when the location of the primary source is changing across 30 positions between $(-4, \pm 0.5, -8)$ and $(+4, \pm 0.5, -8)$, and the error bars show the maximum and minimum errors at each frequency for $p_{\max} = 0.3$. As shown in this figure, the reproduction error is less than -10 dB for frequencies less than 1000 Hz, 600 Hz, 600 Hz, 300 Hz in the System 4, System 2, System 3, and System 1 respectively. Fig. 4.10(b) shows the error performance versus the maximum normalized power at $f = 800$ Hz. According to the figure, the error of System 4 for $p_{\max} = 0.1$ is equal to that of System 2 for $p_{\max} = 0.9$, System 3 and System 1 for $p_{\max} > 1$. Hence, for the same reproduction error, controlling the sound field outside the listening cube is much easier in System 4 compared to the other three systems.

The reproduction error of the proposed algorithms is shown at $f = 800$ Hz and $p_{\max} = 0.3$ versus the locations of the primary source in Fig. 4.11 when the locations of the primary source is changing between $(-4, 0.25, -8)$ and $(+4, 0.25, -8)$. According to this figure and Fig. 4.10, the performance of the placement-only algorithm is less sensitive to the location of the primary source. The reason is that the radiation patterns of the loudspeakers are composed of one omni-directional pattern for the placement-only method while they are higher order patterns in the pattern selection and joint optimization algorithms.

4.5 Conclusion

A new method to optimize the radiation patterns of the loudspeakers was introduced in this chapter using higher order loudspeakers. The expansion coefficients of the loudspeakers were found by the Constrained Matching Pursuit algorithm. The patterns designed by the

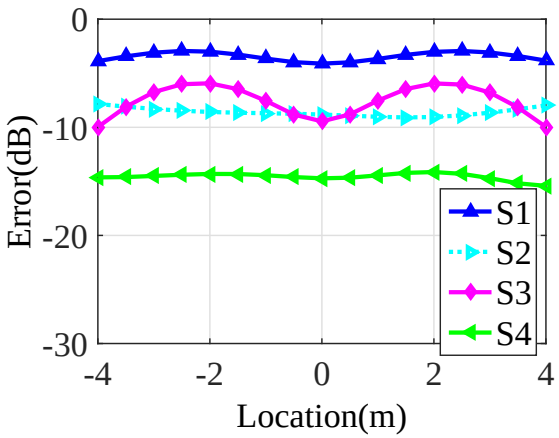


Figure 4.11: Reproduction error at $f = 800$ Hz and $p_{\max} = 0.3$ versus the location of the primary source.

proposed algorithm depend on the features of the primary source. In this chapter, another method was introduced to optimize the patterns (or locations) of the loudspeakers before the system operation when the exact location and frequency of the primary source are not known. With this method, the patterns (locations) of the secondary sources are fixed during the system operation, so the complexity of the run-time optimization decreases.

In addition, a joint placement and pattern optimization algorithm was proposed, which leads to improved system performance. The proposed method optimizes both locations and patterns of the loudspeakers by employing two stages of the CMP algorithm. The error performance of the benchmark configuration, placement, pattern selection and joint optimization algorithms were compared on the 3-D SFR configuration presented in the previous chapter.

According to our simulation results, the performance of the joint optimization method was better than those of the other algorithms. Depending on the frequency and the number of dictionary members, the performance of the placement-only and pattern selection algorithms were different. At lower frequencies, the performance of the pattern selection algorithm was better than that of the placement-only method. However, at higher frequencies, with sufficient number of dictionary members (candidate locations), the performance of the placement-only algorithm was better. In addition, since at lower frequencies, the optimum placement approaches that of the benchmark, the performance of the pattern selection and joint optimization algorithm was similar.

The error performance of the proposed methods were compared when the locations of the primary source and its frequency were not exactly known in advance. In these simulations, the number of dynamic DoFs is fixed for all methods while the number of static DoFs is different. Therefore, all methods were compared under the same system complexity. The

results of this comparison confirmed that the joint optimization algorithm outperforms other methods under different conditions since the number of static DoFs is larger in this method.

In the next chapter, the proposed placement and pattern algorithms will be applied to an SFR system for immersive communication. For this application, the psycho-acoustic features of the human ear are considered in sound field reproduction.

Chapter 5

SFR for Immersive Communication

5.1 Introduction

Immersive communication systems promise a greatly improved user experience through the use of advanced technologies tailored to various human senses, such as sight, hearing, and touch. This chapter focuses on 3-D audio for immersive communication. At the transmitting end, the incident sound field is captured via a microphone array and active talkers are detected through a novel algorithm operating in the frequency domain. The detected information is transmitted to the receiving end, where a 3-D sound field from virtual sources corresponding to active talkers is synthesized around listeners' heads using a direct least squares approximation approach.

In the audio system introduced in this chapter, the microphone and loudspeaker arrays are located on the perimeters of the concentric squares. The goal of this immersive communication system is to remove the perception of the physical distance between the two communicating ends. The system overview along with the system model for audio immersive communication is given in Section 5.2.

At the transmitting end, the sound field is captured using an array of microphones. The recorded sound field is processed using our proposed method in conjunction with a monitoring system in order to detect the locations of the talkers. The proposed algorithm works based on the Least Square algorithm, and it can be implemented in real-time. The details and error analysis of this algorithm are presented in Section 5.3.

The SFR is an important part of immersive communication. As 3-D video provides visual localization of sound sources in the (virtual) 3-D space, the role of the SFR is to complement the visual information by providing the correct auditory localization of sound sources. Sound field synthesis is accomplished via a loudspeaker array at the receiving end. The higher order loudspeakers are employed in the SFR system, and their radiation patterns are optimized by the pattern selection algorithm. To reproduce a sound field naturally, the

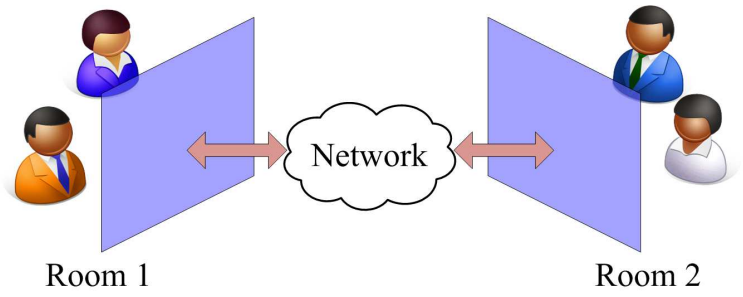


Figure 5.1: Illustration of an immersive communication system.

human' psycho-acoustic features are considered in determining the loudspeakers excitations. Section 5.4 describes the SFR method.

The performance of the proposed algorithm is examined in Section 5.5 using objective and subjective tests. The detection error rate in the active talker detection algorithm and reproduction error along with Interaural Level Difference (ILD) and Interaural Phase Difference (IPD) errors in the SFR algorithm are the quantitative variables for the objective test. Since the best way for quality judgment in an immersive communication system is obtained by subjective testing, the reproduced sound field at the receiving end is recorded and judged by the subjects.

The contributions of this chapter are:

- Employing a new loudspeaker and microphone configuration in sound field capture and rendering
- Introducing a new algorithm to localize the active talkers
- Error analysis of the proposed algorithm and finding an error bound to reduce the detection rate in this algorithm
- Designing the radiation patterns of the loudspeaker before system operation by the multi-frequency pattern selection algorithm, and reproducing the sound field based on the psycho-acoustic phenomena, *i.e.* the Head related Transfer Function of the listeners

5.2 System overview

Fig. 5.1 depicts an immersive communication system. Two rooms are equipped with the necessary hardware, such as 3-D displays, texture and depth cameras, microphone and loudspeaker arrays, and connected to each other via a link with a sufficiently high rate and low latency to support the two-way real-time transfer of the necessary information. There are several participants in each room.

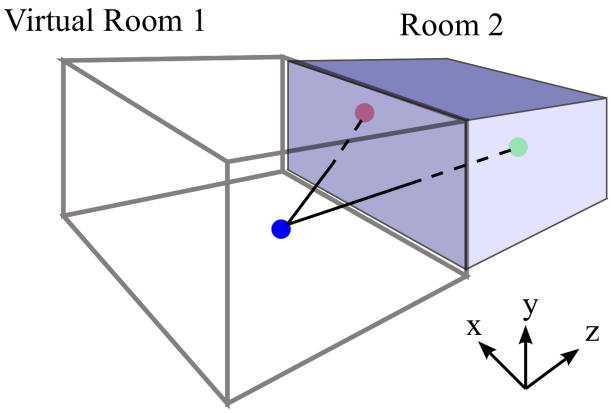


Figure 5.2: The concept of a virtual extension for immersive communication: Virtual Room 1 becomes a virtual extension of Room 2.

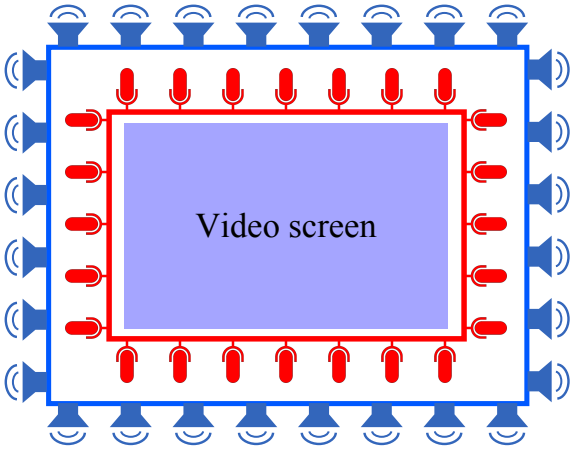


Figure 5.3: Microphone (red) and speaker (blue) arrays for SFR.

The goal of the immersive communication system is to make the participants feel as if the physical distance between the rooms has vanished and the two screens, shown in Fig. 5.1, have merged. To the participants, it should appear as if the screen that they are watching is an open window to the other room. Conceptually, each room should be seen as a virtual extension of the other room, as illustrated in Fig. 5.2. In this figure, one "talker" in room 1 is addressing two listeners in room 2. The immersive communication system makes it appear to the two listeners, whose head positions are indicated by the right-most dots, that the talker's head is located at the position indicated by left-most dot in the virtual room 1. The sound field generated around the heads of the two listeners should be the same as the one generated by a source at the left-most dot. For concreteness, the microphones and loudspeakers are arranged in concentric rectangular arrays around the screen, as shown in Fig. 5.3, and other setups can be deployed without changing the methodology.

The target application of this immersive communication system is video conferencing, so the following assumptions are considered in our simulations: (1) The signals received

by the microphones are either speech or ambient noise, and (2) The two conference rooms are quiet, and there is no outside noise source. (3) The talkers are simply modeled as omni-directional sound sources, which is a reasonable approximation for face-to-face conversation [86]. We will assume that through the use of texture and depth cameras, the 3-D positions of the talkers' and listeners' heads can be estimated via existing algorithms for face and lip detection [87, 88]. The effect of lip localization errors on the sound capture portion of the system is studied in Section 5.3, and the error tolerance is derived.

It is important to realize that transmitting only the pressures detected by the microphones is insufficient to achieve an immersive effect as illustrated in Fig. 5.2(a), since this would not create the virtual source at the required 3-D position in the virtual room. The actual position(s) of active talkers must be transmitted as well. Since all participants are being tracked by face and lip detection, what remains to be determined is which of them are active talkers at any given time. This is discussed in Section 5.3.

The detected locations of the active talkers are transmitted to the other end, where they are used to synthesize the sound field around the listeners' heads. In practice, the sound and position information must be compressed prior to transmission. There are a number of methods for multichannel audio compression [89], while 3-D positions of the talkers' heads can be encoded as point clouds [90], or even transmitted losslessly if the number of participants is small. However, since compression, transmission, echo cancellation (e.g. [91]) and their details are beyond the scope of the present work, we will simply assume that this information is available error- and echo-free at the receiving end, where it is used to synthesize the sound field.

5.3 Active Talker Detection

Let N_1 and N_2 be the number of participants in rooms 1 and 2, respectively. Each room has N loudspeakers and $M > \max(N_1, N_2)$ microphones. The participants in room 1 are the "talkers," and those in room 2 are the "listeners." There exists many methods for voice activity detection [92–95], finding the direction of arrival [96–98], and extracting the signals corresponding to the active talkers by beamforming methods [99,100]. In [101], these processes are combined for the purpose of active talker detection. The system described in [101] can be deployed in our scenario as well. However, in videoconferencing application, cameras are able to provide additional (and usually much more accurate) information about the locations of the talkers. Using this information, the signals corresponding to the active talkers can be detected more precisely compared to the approach in [101], as will be demonstrated later in Section 5.5.1.

5.3.1 Finding complex amplitudes

For reference, the physical link equations for room 1 and room 2 are explained in Appendix C. Let $p_{m,n}(t)$ in [Pa] be the pressure caused by the n-th talker at the m-th microphone. The signal is sampled (8 kHz for conventional telephone-quality speech, or 16 to 44.1 kHz for wideband speech) and processed using the short-time Fourier transform (STFT) with frame length of 1024 and frame shift of 256. The resulting frequency-domain signal, $p_{m,n}(f)$ in [Pa], is:

$$\begin{aligned} p_{m,n}(f, \mathbf{x}_n, \mathbf{y}_m) = a_n(f) \cdot \mathcal{T}_n(f, \theta_m, \phi_m) \cdot g(f; \|\mathbf{x}_n - \mathbf{y}_m\|_2) \\ \cdot \mathcal{M}_m(f, \theta_n, \phi_n), \end{aligned} \tag{5.1}$$

where $a_n(f)$ is the complex amplitude (the pressure caused by the talker) in [Pa·m], $g_{m,n}(.)$ is the free space Green's function between m-th microphone and n-th loudspeaker in [m^{-1}], $\mathcal{T}_n(f, \theta_m, \phi_m)$ is the dimensionless radiation pattern of the talker, and $\mathcal{M}_m(f, \theta_n, \phi_n)$ is the dimensionless receiving pattern of the microphone. In this equation, (θ_n, ϕ_n) are the elevation and azimuth angles of the n-th talker with respect to the m-th microphone, and (θ_m, ϕ_m) are the corresponding angles of the m-th microphone with respect to the n-th talker. The free space Green's function between $\mathbf{x}_n$ and $\mathbf{y}_m$ is given by Eq. (1.4).

Each talker is modeled by an omni-directional sound source, so $\mathcal{T}_n(f, \theta_m, \phi_m) = 1$. We consider the cardioid microphones in the simulations with $\mathcal{M}_m(f, \theta_r, \phi_r) = 1/2(1 + \cos(\theta_r))$ [102] assuming that all microphones are matched in their patterns. It should be noted that the sensitivity analysis on the microphone patterns is not part of our simulations.

The pressures sensed by all microphones are collected in $\mathbf{p}(f) = [p_1(f), p_2(f), ..., p_M(f)]^T$, where $p_m(f) = \sum_n p_{m,n}(f)$ is the total pressure from all sources sensed by the m-th microphone. The relationship between sound signals and microphone readings at f is $\mathbf{p}(f) = \mathbf{G}(f)\mathbf{a}(f)$, where $\mathbf{a}(f) = [a_1(f), a_2(f), ..., a_{N_1}(f)]^T$ and the (m, n)-th element of the $\mathbf{G}(f)$ is $g_{m,n} \cdot T_n \cdot M_m$. Dropping f from the notation, as in [38]:

$$\mathbf{p} = \mathbf{G}\mathbf{a}. \tag{5.2}$$

It is worth mentioning that in order to calculate the pressure caused by a sound source at an arbitrary point in space, the Green's function ($g_{m,n}$) should be multiplied only by the radiation pattern of the sound source ($\mathcal{T}_n$), while in order to calculate the pressure sensed by a microphone, the receiving pattern of the microphone ($\mathcal{M}_m$) should be considered as well.

Ideally, the complex amplitudes of participants' speech signals would be computed by pseudo-inverting Eq. (5.2):

$$\mathbf{a} = (\mathbf{G}^H\mathbf{G})^{-1}\mathbf{G}^H\mathbf{p}. \tag{5.3}$$

Since $M \gg N_1$, Eq. 5.3 leads to a unique solution for $\mathbf{a}$. That is, this equation gives the complex amplitudes of the talkers speech signals at each time frame, across all frequency components.

In developing the active talker detection algorithm, we use the following facts: (1) most of the energy of the speech signal is concentrated at frequencies below 4 kHz [103], and (2) the average Sound Pressure Level (SPL) at a distance of 1 m (which corresponds to $\|\mathbf{x}_n - \mathbf{y}_m\|_2 = 1$ in Eq. (1.4)) over the frequency range of the speech signal is between 40 dB to 60 dB. The complex amplitudes $\mathbf{a}$ in our system correspond to the sound pressure at 1 m. Hence, when a participant is talking, there is a high probability that the resulting pressure at 1 m (which is proportional to the magnitude of the complex amplitude) in a number of low frequency bins exceeds 40 dB SPL at 1 m. Therefore, at the core of active talker detection is the comparison of the components of $\mathbf{a}$ with a suitably chosen threshold T for frequencies below 4 kHz. If at least for K' frequency components, the coefficients (components of $\mathbf{a}$) corresponding to a talker are greater than the threshold, the talker is detected as active talker. After detection, the corresponding location to the active talker along with its complex amplitude across the whole frequency range is sent to room 2 for SFR. In order to reduce the effects of noise and increase the detection rate, the signal is preprocessed by a smoothing filter as in [93, 94]. Specifically, the signal is smoothed by a Hanning window, and the number of neighbors in the smoothing process for time and frequency components is found by the method proposed in [93].

The detection method outlined above works well if there is no error in the talkers' locations provided by the face and lip tracking system. However, face and lip position estimates in practice are not perfectly accurate, so errors may be introduced in $\mathbf{x}_n$ in Eq. (1.4). Such errors will not influence the denominator of Eq. (1.4) very much, because they are usually much smaller than $\|\mathbf{x}_n - \mathbf{y}_m\|_2$, the distance between the talker and the microphone. However, the phase in the numerator is sensitive: a 2 cm error in the location of talker's lip centroid estimate would result in a 20 degree phase error in $g_{m,n}$ at 1000 Hz, and a 71 degree phase error at 3400 Hz. Such errors mean that $\mathbf{a}$ from Eq. (5.3) will contain errors as well. Since active talker detection relies on thresholding of $\mathbf{a}$, errors in $\mathbf{a}$ may result in erroneous detection. In the following subsection we analyze the effect of errors in $\mathbf{G}$ on $\mathbf{a}$, and we find a bound for the maximum error in the lip detection that does not cause error in the active talker detection process. The resulting new active talker detection algorithm is also presented.

5.3.2 Error analysis

Let $\|\mathbf{x}_n - \mathbf{y}_m\|_2 = r_{m,n}$ be the distance of the n-th active talker from the m-th microphone, and let $\widetilde{\mathbf{x}_n} = \mathbf{x}_n + \boldsymbol{\epsilon}$ be the lip position estimated by the camera system, where $\mathbf{x}_n$ is the true position and $\boldsymbol{\epsilon}$ is the position estimate error. Assuming omni-directional sources and

microphones, the "noisy" element of the ATF matrix is:

$$\widetilde{g}_{m,n} = \frac{e^{ik(r_{m,n}+\eta_{m,n})}}{4\pi(r_{m,n}+\eta_{m,n})} \approx \frac{e^{ikr_{m,n}} \cdot e^{ik\eta_{m,n}}}{4\pi r_{m,n}} = g_{m,n}n_{m,n} \tag{5.4}$$

where $\eta_{m,n}$ is the error in $r_{m,n}$ induced by $\boldsymbol{\epsilon}$, which introduces a multiplicative noise term $n_{m,n}$ in the ATF matrix. Let $\widetilde{\mathbf{G}} = \mathbf{G} \odot \mathbf{N}$ be the "noisy" ATF matrix formed based on the participants' lip position estimates, where $\mathbf{G}$ is the true ATF matrix (involving true lip positions), $\mathbf{N}$ is the noise matrix containing multiplicative terms due to lip position errors, and $\odot$ is the element-wise multiplication. Evaluating Eq. (5.3) with $\mathbf{G}$ replaced by $\widetilde{\mathbf{G}}$ leads to $\widetilde{\mathbf{a}} = \mathbf{a} + \mathbf{e}$, where $\mathbf{a} = (\mathbf{G}^H\mathbf{G})^{-1}\mathbf{G}^H\mathbf{p}$ and $\mathbf{e} = [e_1, e_2, ..., e_{N_1}]^T$ represents the errors in complex amplitudes caused by incorrect lip position estimates.

As mentioned above, finding active talkers and their complex amplitudes amounts to comparing the components of $\widetilde{\mathbf{a}}$ with a threshold value T. If $|\widetilde{a}_j| > T$ for one frequency component, the j-th participant is recognized as an active talker, otherwise she is assumed to be silent and $\widetilde{a}_j$ set to zero. The question we aim to answer here is how much the error in $\widetilde{\mathbf{G}}$ affects the detection of active talkers.

For the moment, assume the first participant is the only active talker, so $\mathbf{a} = [a_1, 0, 0, ..., 0]^T$ and $\widetilde{\mathbf{a}} = [a_1 + e_1, e_2, ..., e_{N_1}]^T$. Note that Eq. (5.3) is also the solution of the following Least Squares (LS) problem:

$$\widetilde{\mathbf{a}} = \arg\min_{a^o} \|\widetilde{\mathbf{G}}\mathbf{a}^o - \mathbf{p}\|_2^2. \tag{5.5}$$

In this equation, $\mathbf{a}^o$ is a dummy vector (argument of the error function), and $\widetilde{\mathbf{a}}$ is the vector that minimizes the error function. Therefore, Eq. (5.3) minimizes the ℓ_2-norm of the difference between $\mathbf{p}$ and $\widetilde{\mathbf{G}}\widetilde{\mathbf{a}}$.

Let $\mathbf{g}_j$, $\widetilde{\mathbf{g}}_j$, and $\mathbf{n}_j$ be the j-th columns of $\mathbf{G}$, $\widetilde{\mathbf{G}}$, and $\mathbf{N}$, respectively, and let $\mathbf{v}$ be the microphone noise converted to pressure (i.e., multiplication of microphone electrical noise, which is in volts, by the inverse transfer function of microphone, see Appendix C). Then the total pressure sensed by the microphones is :

$$\mathbf{p} = \mathbf{G}\mathbf{a} + \mathbf{v} = a_1\mathbf{g}_1 + \mathbf{v}.$$

Let $\widetilde{\mathbf{p}} = \widetilde{\mathbf{G}}\widetilde{\mathbf{a}}$. It can be rewritten as:

$$\widetilde{\mathbf{p}} = \widetilde{\mathbf{G}}\widetilde{\mathbf{a}} = (a_1 + e_1)(\widetilde{\mathbf{g}}_1) + e_2\widetilde{\mathbf{g}}_2 + ... + e_{N_1}\widetilde{\mathbf{g}}_{N_1} = a_1\widetilde{\mathbf{g}}_1 + \widetilde{\mathbf{G}}\mathbf{e} \tag{5.6}$$

The above equation can be expressed as:

$$\widetilde{\mathbf{p}} = a_1\mathbf{g}_1 + a_1\mathbf{n}'_1 + \widetilde{\mathbf{G}}\mathbf{e} = \mathbf{p} - \mathbf{v} + a_1\mathbf{n}'_1 + \widetilde{\mathbf{G}}\mathbf{e} \tag{5.7}$$

where $\mathbf{n'}_1 = \mathbf{g}_1 \odot (\mathbf{1} - \mathbf{n}_1)$. Solving for $\mathbf{p}$ and substituting into $\|\widetilde{\mathbf{G}}\mathbf{a}^o - \mathbf{p}\|_2^2$ gives

$$\begin{aligned}\|\widetilde{\mathbf{G}}\mathbf{a}^o - \mathbf{p}\|_2^2 = \|\widetilde{\mathbf{G}}\mathbf{a}^o - \widetilde{\mathbf{p}} + a_1\mathbf{n'}_1 - \mathbf{v} + \widetilde{\mathbf{G}}\mathbf{e}\|_2^2 &= \|\widetilde{\mathbf{G}}\left(\mathbf{a}^o - \widetilde{\mathbf{a}} + \mathbf{e}\right) + a_1\mathbf{n'}_1 - \mathbf{v}|_2^2 \\ = \|\widetilde{\mathbf{G}}\left(\mathbf{a}^o - \mathbf{a}\right) + a_1\mathbf{n'}_1 - \mathbf{v}\|_2^2 &= \|\widetilde{\mathbf{G}}\mathbf{e}^o + a_1\mathbf{n'}_1 - \mathbf{v}\|_2^2.\end{aligned} \tag{5.8}$$

Hence, LS minimization in Eq. (5.5) results in the solution $\widetilde{\mathbf{a}} = \mathbf{a} + \mathbf{e}$ such that the error vector $\mathbf{e}$ minimizes the ℓ_2-norm of the difference between $\widetilde{\mathbf{G}}\mathbf{e}$ and $-a_1\mathbf{n'}_1 + \mathbf{v}$. This means that the resulting error vector can be written as:

$$\mathbf{e} = (\widetilde{\mathbf{G}}^H\widetilde{\mathbf{G}})^{-1}\widetilde{\mathbf{G}}^H(-a_1\mathbf{n'}_1 + \mathbf{v}). \tag{5.9}$$

Let $\widetilde{\mathbf{G}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$, where $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}$ are the SVD matrices of $\widetilde{\mathbf{G}}$. Based on [39], and using the relationship in (5.9), the ℓ_2 norm of the error vector is given by:

$$\|\mathbf{e}\|_2^2 = \sum_{j=1}^{N_1} \frac{1}{(\sigma_j)^2}|c_j|^2, \tag{5.10}$$

where σ_j is the j-th singular value of $\widetilde{\mathbf{G}}$, and c_j is the projection of $-a_1\mathbf{n'}_1 + \mathbf{v}$ onto the j-th column of $\mathbf{U}$, denoted $\mathbf{u}_j$. Since $\mathbf{U}$ is a unitary matrix, $|c_j|^2 \leq \|\mathbf{u}_j\|_2^2 \cdot \|(-a_1\mathbf{n'}_1) + \mathbf{v}\|_2^2 \leq |a_1|^2 \cdot \|\mathbf{n'}_1\|_2^2 + \|\mathbf{v}\|_2^2$, so we obtain the following bound:

$$\|\mathbf{e}\|_2^2 \leq \sum_{j=1}^{N_1} \frac{1}{(\sigma_j)^2}\left(|a_1|^2 \cdot \|\mathbf{n'}_1\|_2^2 + \|\mathbf{v}\|_2^2\right) = (|a_1|^2 \cdot \|\mathbf{n'}_1\|_2^2 + \|\mathbf{v}\|_2^2) \cdot \|\widetilde{\mathbf{G}}^+\|_F^2, \tag{5.11}$$

where $\widetilde{\mathbf{G}}^+ = (\widetilde{\mathbf{G}}^H\widetilde{\mathbf{G}})^{-1}\widetilde{\mathbf{G}}^H$ is the pseudo-inverse of $\widetilde{\mathbf{G}}$, and $\|\cdot\|_F$ is the Frobenius norm.

Recall that our working assumption is that only the first participant is an active talker. Hence, for accurate active talker detection, we must have $|a_1 + e_1| > T$ and $|e_j| \leq T$ for $j \geq 2$. Assuming $|a_1| \geq |e_1|$ (signal stronger than "noise"), the first condition can be rewritten as $|a_1 + e_1| \geq |a_1| - |e_1| > T$, or $|a_1| > T + |e_1|$. Hence, if $|e_1|$ satisfies the same condition as other e_j's, that is $|e_1| \leq T$, then accurate detection will be achieved if $|a_1| > 2T$. From the above discussion we see that the necessary and sufficient conditions for active talker detection are that the errors in lip detection ("noise") be bounded ($|e_j| \leq T$ for all j) and the signal be strong enough ($|a_1| > 2T$). For our simulations, this threshold is found as a function of the low end of the typical sound pressure level of speech at a distance of 1 m, which is 40 dB SPL [104]. From Eq. (5.2), the pressure amplitude at a distance of 1 m from the talker is equal to the magnitude of the complex amplitude divided by 4π. Hence, the threshold is set to $T = 1/2 \cdot 4\pi \cdot 20 \cdot 10^{-6} \cdot 10^{40/20}$, where $20 \cdot 10^{-6} \cdot 10^{40/20} = 40$ dB SPL is the low end of the typical sound pressure level of speech.

Using the condition $|e_j| \leq T$ for all j, we can work backwards and find out the maximum error in the lip detection algorithm that would still allow accurate active talker detection.

Since $|e_j| \leq T$ for all j, we have that $\|\mathbf{e}\|_2^2 \leq N_1 T^2$. Based on Eq. (5.11), this will hold if

$$\|\mathbf{n'}_1\|_2^2 < \frac{N_1 T^2}{|a_1|^2 \cdot \|\widetilde{\mathbf{G}}^+\|_F^2} - \frac{\|\mathbf{v}\|_2^2}{|a_1|^2}. \tag{5.12}$$

Let v_m be the noise at microphone m. If the noise at different microphones is independent and identically distributed (iid) with zero mean and variance σ_v^2, then by the law of large numbers

$$\|\mathbf{v}\|_2^2 = \sum_{m=1}^{M} (v_m)^2 \approx \sigma_v^2 \cdot M.$$

The squared magnitude of the received signal at m-th microphone is equal to $|p_m|^2 = |a_1|^2/(4\pi r_{m,1})^2$ where $r_{m,1}$ is the distance between the m-th microphone and the first participant, which is the only talker. Therefore, $|a_1|^2 = |p_m|^2 (4\pi r_{m,1})^2$. Substituting into Eq. (5.12) results in:

$$\|\mathbf{n'}_1\|_2^2 < \frac{N_1 T^2}{|a_1|^2 \cdot \|\widetilde{\mathbf{G}}^+\|_F^2} - \frac{\sigma_v^2 \cdot M}{|p_m|^2 (4\pi r_{m,1})^2} < \frac{N_1 T^2}{|a_1|^2 \cdot \|\widetilde{\mathbf{G}}^+\|_F^2} - \frac{M}{2SNR_{mic}(4\pi r_{\min})^2}. \tag{5.13}$$

where $SNR_{mic} = (0.5|p_m|^2)/\sigma_v^2$ is the microphone's signal to noise ration ($0.5|p_m|^2$ is the received signal energy), and $r_{\min} = \min_{m,n} \{r_{m,n}\}$ is the minimum distance between microphones and participants. Hence, a sufficient condition for accurate detection is that the ℓ_2-norm of the "noise" component on the ATF vector of the active talker ($\mathbf{n}_1$) satisfies the above inequality.

Let $\epsilon = \|\boldsymbol{\epsilon}\|_2$ be the norm of the lip position error vector. The maximum value of ϵ, call it $\epsilon_{\max}$, that satisfies the above inequality, gives the maximum inaccuracy in lip position estimate that would still allow accurate active talker detection. This value can be computed numerically from Eq. (5.13). In Section 5.5 we will examine how $\epsilon_{\max}$ obtained from Eq. (5.13) compares with experimentally obtained values via Monte Carlo simulations. Now that we know how accurately lip positions must be detected in order to achieve accurate active talker detection, we develop an algorithm that performs active talker detection. Let R be the maximum lip position error from the lip detection algorithm. Since lips are part of the face, this error can be upper-bounded by the accuracy of face detection. We used $R = 3$ cm in our simulations, unless otherwise stated. If $R \leq \epsilon_{\max}$, then simply thresholding $\widetilde{\mathbf{a}}$ is enough to achieve accurate active talker detection. In the more challenging case, $R > \epsilon_{\max}$, we can subdivide a sphere of radius R, centered at the detected location, into smaller sub-spheres of radii $\epsilon_{\max}$, and test the possibility that the lip centroid is in each of the smaller sub-spheres. This idea is illustrated in Fig. 5.4 and formalized in Algorithm 9, where $\widetilde{\mathbf{x}}_j$ is the (possibly inaccurate) initial estimate of the lip centroid of the j-th participant obtained from the lip detection algorithm.

Algorithm 9 Active talker detection

Input: $\mathbf{p}$ ▷ pressure measured by the microphones
Input: $\widetilde{\mathbf{x}}_j$ ▷ detected lips centroids
Input: R ▷ max. lip position error
Input: $\epsilon_{\max}$ ▷ max. tolerable position error
Output: A ▷ set of active talkers
Output: $\widehat{\mathbf{a}}$ ▷ complex amplitudes of active talkers

For each frame:
Set $A = \emptyset$.
Consider a sphere of radius R centered at each $\widetilde{\mathbf{x}}_j$.
Pack smaller sub-spheres of radius $\epsilon_{\max}$ inside the spheres of radius R. Let N_s be the total number of sub-spheres around all $\widetilde{\mathbf{x}}_j$'s.
Using Eq. (1.4), construct the $M \times N_s$ ATF matrix $\mathbf{G_s}$ that describes sound pressure transfer from each of the N_s sub-sphere centers to each of the M microphones.
Compute $\mathbf{a}_s = (\mathbf{G}_s^H \mathbf{G}_s + \gamma \mathbf{I})^{-1} \mathbf{G}_s^H \mathbf{p}$. This is the combination of complex amplitudes at sub-sphere centers that would have caused measured pressures $\mathbf{p}$ at the microphones, and where γ is a regularization parameter and $\mathbf{I}$ is the identity matrix. In our simulations, $\gamma = 0.001$.
for $j = 1$ to N_1 **do**
    Let $\mathbf{a}_j = [a_{j_1}, a_{j_2}, .., a_{j_{N_c}}]^T$ be the components of $\mathbf{a}_s$ that correspond to the sub-spheres around $\widetilde{\mathbf{x}}_j$, where N_c is the number of sub-spheres per talker.
    if at least for K' frequency components, $|\sum_{j_k=1}^{N_c} a_{j_k}| > T$ **then** (In our simulations, $K' = 30$.)
        The j-th participant is declared an active talker.
        Set $A = A \cup \{j\}$.
    else
        The j-th participant is not an active talker.
    end if
end for
For all frequency components in the current frame, using Eq. (1.4), construct the $M \times (|A| \cdot N_c)$ ATF matrix $\mathbf{G}_c$ that describes sound pressure transfer from all of the selected sub-sphere centers of active talkers to each of the M microphones.
Compute $\mathbf{b} = (\mathbf{G}_c^H \mathbf{G}_c + \gamma \mathbf{I})^{-1} \mathbf{G}_c^H \mathbf{p}$.
Calculate the complex amplitude of each active talker as $\widehat{a_j} = \sum_{n=1}^{N_c} b_n^j$, where b_n^j's are the elements of $\mathbf{b}$ that correspond to the j-th active talker.
Put the complex amplitudes of active talkers in vector $\widehat{\mathbf{a}}$.
return A, $\widehat{\mathbf{a}}$.

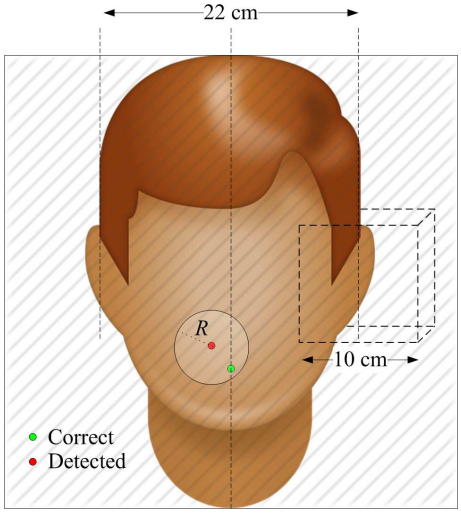


Figure 5.4: Illustration of the initially detected (red) and the correct lip centroid. A 10 cm cube around an ear is also illustrated.

5.4 Sound field reproduction

After finding the locations and complex amplitudes of the active talkers, they are sent to room 2 for sound field reproduction. This section treats loudspeaker radiation patterns as static DoFs, similar to [15, 38, 39] which are found by the pattern selection algorithm proposed in Chapter 4. Algorithm 6 is the basic building block for loudspeaker pattern design. It assumes that the locations of the speaker and the listeners are fixed and known in advance. Such an assumption is reasonable, for example, when the rooms are outfitted with fixed seats, such as media rooms or some conference rooms, which constrains the possible positions of participants. More generally, however, there will be a certain volume of space in the rooms that provides comfortable viewing of the 3-D display, and this entire volume should be taken into account when designing loudspeaker patterns. An example is given in Fig. 5.5, where the volume of interest is illustrated as a red rectangular parallelepiped. In addition, Algorithm 6 optimizes the patterns for one frequency of the desired field. While such patterns should be close to optimal at neighboring frequencies, they deviate from optimal as the frequency of interest becomes significantly different from the one used in the design. To avoid it, Algorithm 7 is deployed here for more general setting as follows. First, to account for all possible listeners' positions in the volume of interest, we distribute the virtual sampling points across the entire volume of interest (red parallelepiped in Fig. 5.5): $\{\mathbf{y}_1^l, \mathbf{y}_2^l, ..., \mathbf{y}_K^l\}$. Note that K, the number of sampling points during the design phase, *i.e.* before the system operation, the system operation, will be different from and usually much larger than the number of sampling points during the system operation. This will increase the dimension of the desired field vector $\mathbf{p}^{\text{des}}$. Next, as in Section 4.2.3, to account for various possible active talker locations, we distribute W points uniformly across the volume

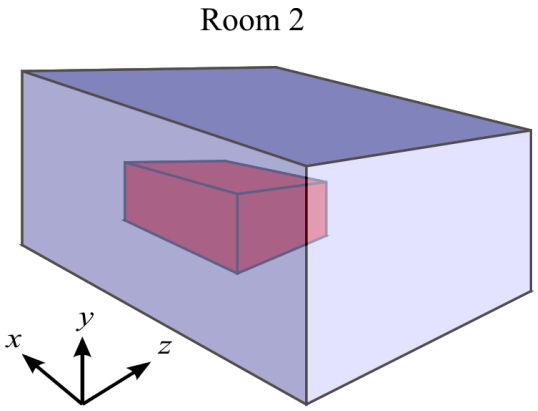


Figure 5.5: Possible locations of listeners' heads in room 2.

of interest, as representative locations and to account for the frequency band of interest (which is typically taken as 300 Hz to 3400 Hz for human speech [104]), we distribute Y frequency points across this range. It may also be possible to extend the design to a region exterior to the red parallelepiped in Fig. 5.5 and force the sound field to be reduced in the exterior in order to diminish undesired reverberation, as discussed for 2-D case in [105].

The radiation patterns of the loudspeakers are optimized before the system operation. During the system operation, in order to reproduce the sound field around the listener's heads, the complex amplitudes of the loudspeakers are optimized as dynamic DoFs. In the proposed model, the listeners' locations are detected by the monitoring system, and two cubic regions around the listeners' heads are considered as listening areas, see Fig. 5.4. M_v sampling points are distributed uniformly in each cubic region. Hence, the total number of sampling points is $M_s = 2M_v \cdot N_2$. In the previous chapters, the complex amplitudes of the loudspeakers were determined such that the reproduction error is minimized at the sampling points. In this chapter, the complex amplitudes of loudspeakers are optimized to minimize the error perceived by the listeners. The perceived error is defined as the perceived (by the listener) difference between the sensed desired field and the sensed reproduced field. To calculate these fields, the Head Related Transfer Function (HRTF) of the listeners should be taken into account in calculation of the complex amplitudes of the loudspeakers. In the following, first, the HRTF is described briefly, and then the complex amplitudes of the loudspeakers are optimized for SFR.

HRTF is defined as:

$$\mathcal{H}(f) = \frac{F_e(f)}{F_n(f)} \tag{5.14}$$

where f is the frequency, F_e is the Fourier transform of the received signal at the eardrum, and F_n is the Fourier transform of the signal at the same point without considering head. In [106] the human head is considered as a rigid sphere, and HRTF is found by solving wave

equation of a rigid sphere as follows:

$$\mathcal{H}(\rho,\mu,\theta) = -\frac{\rho}{\mu}e^{-i\rho\mu}\sum_{m_h=0}^{\infty}(2m_h+1)P_{m_h}(\cos(\theta))\frac{h_{m_h}(\mu\rho)}{h'_{m_h}(\mu)} \tag{5.15}$$

where $\rho = r/a$ is the normalized distance to the source, a is the radius of the rigid sphere (head), $\mu = ka = 2\pi fa/C$ is the normalized frequency, and θ is the angle of incidence with the corresponding ear. In our simulations, we employed the HRTF from Eq. (5.15).

At low frequencies whose wavelength is comparable with the size of a human head the magnitude of the HRTF does not change significantly at the two ear canals. In this situation, the phase difference between the two ears localizes the sound source. This cue is called the Interaural Phase Difference (IPD), and the corresponding time difference in receiving signals by the two ears is referred to as Interaural Time Difference (ITD). For a single tone source, the ITD can be measured by the phase difference between the sounds at the two ears. However, when the source frequency increases and the phase difference reaches 180°, the source location based on this phase shift becomes ambiguous [107]. This angle can be calculated in terms of the frequency and azimuth angle of the source relative to the listener's head. Let ϕ be the relative elevation angle of a point source to the listener's head. For a source which is located far enough from the human head, the distance between the two ears at the azimuth angle ϕ is:

$$\Delta d = a(\sin(\phi) + \phi), \tag{5.16}$$

The corresponding phase shift is $k\Delta d$, so the frequency above which the phase difference is ambiguous is:

$$f^{amb} = \frac{C}{2a(\phi + \sin(\phi))}. \tag{5.17}$$

This parameter is calculated for the numerical experiment in Section 5.5.

On the other hand, at higher frequencies whose wavelength is smaller than the size of a human head, the rigid sphere (human head) scatters the wave, and the contralateral ear, would be in a sound shadow. Therefore, at these frequencies the magnitude of HRTF is larger in the ipsilateral ear in comparison to that of the contralateral one. This phenomenon is called Interaural Level Difference (ILD) which is the most important clue in sound localization at higher frequencies. Here, the contralateral ear refers to the one that is further away from the source, and the ipsilateral erar is the one which is closer to the sound source.

In SFR for immersive communication, the sound field should be created such that the listeners feel that they are in the same environment with the talkers. Hence, the listeners should be able to localize the talker. It means that the ITD and ILD of the reproduced field should be as close as possible to those of the desired field. Therefore, in finding the complex amplitudes of the loudspeakers, the elements of the desired field and the ATF matrix at

the sampling points should be modified by the receiving pattern of the human ear. For this purpose, the perceived ATF matrix $\mathbf{G}_2$ is calculated form the loudspeaker array to the listening areas. The (m_s, n)-th element of this matrix is the ATF from the n-th loudspeaker to the m_s-th virtual sampling point multiplied by the receiving pattern of the Human head, Head-Related Transfer Function (HRTF) [106]. In addition, the elements of the perceived desired vector $\mathbf{p}_2^{\text{des}}$ are modified as:

$$p_2^{\text{des}}(m) = p^{\text{des}}(m)\mathcal{H}_m(\rho, \mu, \theta). \tag{5.18}$$

where $p^{\text{des}}(m)$ and $p_2^{\text{des}}(m)$ are the desired field and the desired field sensed by the listener at the m-th sampling point, and $\mathcal{H}_m(.)$ is the HRTF of the m-th listener when the sound source is placed at the active talker's location (for link equation, see Appendix C). The optimal excitation vector $\mathbf{s}^{\text{opt}}$ is then found by the Least Squares solution as:

$$\mathbf{s}^{\text{opt}} = (\mathbf{G}_2^H\mathbf{G}_2 + \lambda\mathbf{I})^{-1}\mathbf{G}_2^H\mathbf{p}_2^{\text{des}}, \tag{5.19}$$

Therefore, instead of reproducing the sound field at the sampling points, the sound field sensed by the listeners is reproduced in the listening areas. This method results in preservation of the localization cues and reproduction a more natural sound field, which is the main requirement of an immersive communication system. The HRTF-based reproduction error along with ILD and IPD errors are examined in the next section for the proposed configuration.

5.5 Numerical experiments

In our experiments, the (width × height × depth) of room 1 is (5m × 3m × 3m) and for room 2 it is (6.4m × 3m × 5m). The video screen is assumed to fit within a 2 m×2 m frame, so $M = 300$ cardioid microphones are distributed uniformly in a 2m × 2m peripheral array with SNR_{mic} of 30 dB, and $N = 48$ loudspeakers are uniformly distributed on a larger 2.5m × 2.5m peripheral array, as illustrated in Fig. 5.3. The order of loudspeakers is $L = 5$ in all experiments unless otherwise is stated.

The screen and the two arrays are placed on the x-y wall at a distance of 0.1m from the wall. The origin is in the center of the x-y wall, so $z > 0$ represents points in room 1 while $z < 0$ represents points from room 2 mapped to the virtual extension of room 1, as shown in Fig. 5.2. In the experiments where listeners' locations are assumed fixed and known, the listening volume consists of two 10cm × 10cm × 10cm cubes located around the ears of each listener, as shown in Fig. 5.4. Unless otherwise stated, all simulations are performed for reverberant rooms, and the reverberation is modeled by the image source model [108]. In our configuration, the microphone and loudspeaker arrays are installed on the wall, so that wall should be considered as a rough surface. This implies that the reflections from

Table 5.1: Computational complexity and execution times of the proposed algorithms.

	Complexity	Time (sec)
Algorithm 9	$O(N_1 N_s^3 + M N_1 N_s^2)$	5×10^{-5} to 2.5×10^{-2}
Algorithm 6	$O(M_s L^4 N + M_s L^2 N^2)$	0.4
SFR operation	$O(N^2 M_s + N^3)$	2.5×10^{-4}

that wall are not coherent, so these reflections are modeled by an incoherent image source method proposed in [109]. All image sources whose power is greater than or equal to 1% of the power of the actual source are retained. As an example, for the reflection coefficient of 0.7, the number of retained image sources is 6 in each direction, or 2197 in total. The reflection coefficients are assumed to be independent of angle of incidence and frequency.

The computational complexities of Algorithms 9 and 6, as well as the SFR operation-time complexity (solving Eq. (5.19)) are listed in Table 5.1. The corresponding execution times per frequency component of an un-optimized MATLAB implementation on a 3-GHz Intel Core 2 Quad Q9650 processor are also shown. Algorithm 9 and SFR should be performed in real time, while Algorithm 6 is the offline algorithm for loudspeaker pattern design. The execution time of Algorithm 9 is between 5×10^{-5} seconds for lower frequencies ($N_s = 8$ per user) and 2.5×10^{-2} seconds for higher frequencies ($N_s = 216$ per user) which is calculated for $M_v = 27$, $N_2 = 2$, $M_s = 108$, $N_1 = 2$, $L = 5$, $N = 48$. In Table 5.1, the complexity and execution time is shown per frequency component. If the speech signal is sampled at 8 kHz and the frame length of STFT is 1024, there are 1024 frequency components for each 0.128-second segment. Hence, (un-optimized) MATLAB implementation of Algorithm 9 and SFR would not work in real time on a current commodity processor such as a single 3-GHz Intel Core 2 Quad Q9650 processor, but real time operation would be possible on parallel processors. Optimized, embedded implementations would be more efficient.

Section 5.5.1 presents the objective evaluation of the proposed algorithms for active talker detection and SFR, and Section 5.5.2 compares our SFR system with three other SFR systems [42, 60, 61].

5.5.1 Objective evaluation

First, we examine the active talker detection method from Section 5.3, especially the maximum tolerable error $\epsilon_{\max}$ from Eq. (5.13). To do so, we assume the reflection coefficients of all the walls in room 1 are equal, and there are two participants in this room ($N_1 = 2$), whose true lip centroid coordinates in meters are $(1, 0, -1)$ and $(0.5, 0, -1)$, respectively. In the first test, one of these participants is actively talking and the other is silent. At each frequency, the talker's complex amplitude is generated randomly in phase and amplitude such that the resulting pressure at 1 m is in the range of 40 dB SPL to 60 dB SPL. At the

same time, her lip centroid position is perturbed by ϵ meters. Active talker detection is performed by thresholding $\widetilde{\mathbf{a}}$, as discussed in Section 5.3.2. An error event occurs when either: 1) the silent participant is detected as an active talker, 2) both participants are detected as active talkers, 3) neither participant is detected as an active talker. We increase ϵ until the first error event occurs, at which point the corresponding ϵ is recorded (See Fig. 5.6(a)).

The simulation was repeated $10,000$ times at each frequency. The smallest ϵ at which an error was detected is denoted by ϵ_1. The experiment is repeated with both participants simulated as active talkers, by randomly choosing their amplitudes as phases as described above. In this case, an error event occurs when either: 1) only one participant is detected as an active talker, or 2) neither participant is detected as an active talker. Again, the smallest ϵ at which an error event occurs is denoted by ϵ_2 at each frequency. The minimum of ϵ_1 and ϵ_2 at each frequency bin is shown in Fig. 5.6(a) for the reflection coefficients of 0.1, 0.5 and 0.9 in room 1. $\epsilon_{\max}$ from Eq. (5.13), for free space propagation, is shown as the blue solid curve. Since $\epsilon_{\max}$ is the largest theoretically predicted ϵ that would not compromise active talker detection, no error events should be recorded for $\epsilon < \epsilon_{\max}$, and the curves in Fig. 5.6(a) confirm this. $\epsilon_{\max}$ provides a useful parameter for system design. If the lip detection algorithm is highly accurate and its error in finding the lip centroids is less than $\epsilon_{\max}$, simple thresholding of vector $\widetilde{\mathbf{a}}$, as discussed in Section 5.3.2, is sufficient to detect active talkers and there is no need to run Algorithm 9.

The second experiment examines the performance of Algorithm 9 assuming that the lip position error in the z-direction is much smaller than the error in the x-y plane. The talkers' locations and complex amplitudes are the same as in the previous experiment. A crucial parameter of the algorithm is set in Step 4, where the radius of sub-spheres is specified. By default, this is set to $\epsilon_{\max}$ to ensure accurate detection. Here we examine what happens as this radius changes. In this experiment, lip detection accuracy is $R = 3$ cm. A random radial perturbation of up to 3 cm is added to the true location of the lip centroid 100 times, and for each perturbed location, we run Algorithm 9 for various values of the sub-sphere radius between 0.2 cm and 2 cm. For each value of the sub-sphere radius and each perturbed location, the experiment is performed 100 times, giving a total of $10,000$ runs for each sub-sphere radius. The rate of detection error events is measured at 1500 Hz and 3000 Hz in a reverberant room with a reflection coefficient of 0.9. The results are shown in Fig. 5.6(b).

As seen in the figure, the error rate goes to zero as the radius of sub-spheres decreases. In fact, the error rate approaches zero even for sub-sphere radii larger than $\epsilon_{\max}$, as could have been expected from the results in Fig. 5.6(a). This means that $\epsilon_{\max}$ (from Eq. 5.13) is a safe choice for the sub-sphere radius. The two horizontal lines in the figure indicate the error rates obtained at the corresponding frequencies by simply thresholding vector $\widetilde{\mathbf{a}}$ directly, without using Algorithm 9. Hence, without the help of Algorithm 9, active talker detection would face an error floor.

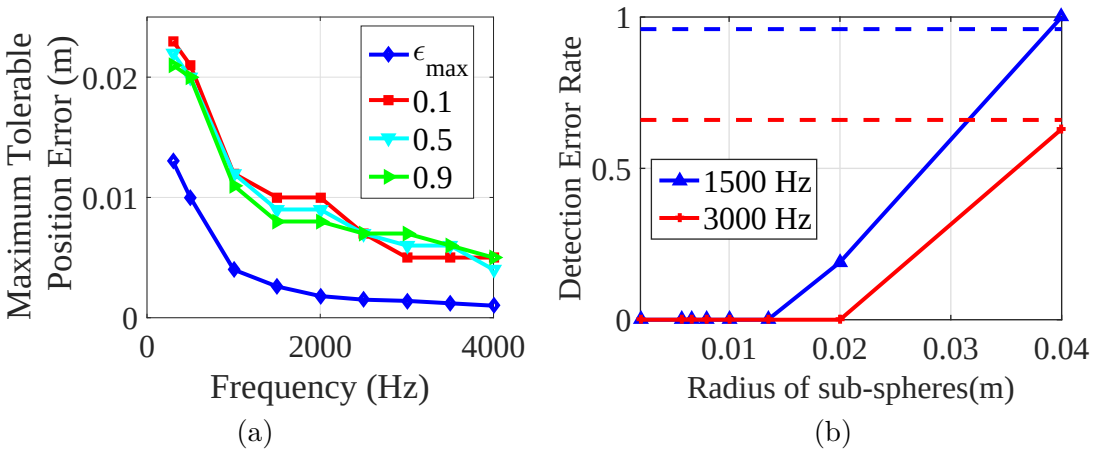


Figure 5.6: (a) Maximum tolerable error versus frequency for reverberant room, (b) Detection error rate of Algorithm 9 versus radius of sub-spheres.

In the third experiment, with the same parameters as above, it is assumed that the first talker is active. Then, 10 recorded speech files[1] sampled at 44100 Hz with 3 sec length are considered as the signals corresponding to the first talker. In the simulation, these are played out from the position of the first talker's lips ($(1, 0, -1)$ m, as above) and sensed by the microphone array. Assuming the lip position error is $R = 3$ cm, the output SNR is calculated in three ways: 1) by the beamforming method from [101]; 2) from Eq. (5.3); and 3) by Algorithm 9. The $R = 3$ cm lip position error in this simulation is equal to the 3% error in time delay of propagation associated with the direct path in the formulation of [101]. In [101], the direct-path signal is considered the desired signal while the reverberations are considered interference. Table 5.2 shows the results. As seen in the table, knowing the system geometry and thus the ATF even approximately (with an error of $R = 3$ cm), gives an average 3 dB advantage to Eq. (5.3) compared to the approach of [101]. A more systematic search for possible lip centroid locations in Algorithm 9 further improves the output SNR dramatically.

To assess our SFR system performance quantitatively, the HRTF-based error (in dB) is calculated as follows:

$$\text{Error (dB)} = 10 \log_{10} \frac{||\mathbf{p}_2^{\text{des}} - \mathbf{G}_2 \mathbf{s}^{\text{opt}}||_2^2}{||\mathbf{p}_2^{\text{des}}||_2^2}. \tag{5.20}$$

In addition, in order to provide a quantitative measure for the sense of immersion, we also investigate how well the listeners are able to localize the virtual sources of reproduced sound fields by computing two important parameters, the Interaural Time Difference (ITD) and the Interaural Level Difference (ILD). For frequencies less than 1500 Hz, the ITD plays a

[1] http://www.voxforge.org/

Table 5.2: Output SNR (dB) for the third experiment.

Speech sample	[101]	Eq. (5.3)	Algorithm 9
1	15.02	17.64	46.15
2	14.84	17.21	45.51
3	18.03	21.40	49.13
4	17.41	19.88	47.71
5	18.92	22.79	50.19
6	17.71	20.30	48.01
7	16.58	19.15	47.17
8	14.35	16.43	44.81
9	20.55	24.43	51.00
10	18.68	22.15	49.47
Average	17.21	20.14	47.91

more important role in sound localization, while at higher frequencies, the ILD and the ITD of the signal envelope are more important [110, 111].

In our simulations, the normalized ILD (IPD) error is calculated as follows. The desired ILD (IPD) is calculated at all pairs of the virtual sampling points (each pair corresponds to the two ears of one listener) and arranged in a $(M_s/2 = 2M_vN_2/2) \times 1$ vector ILD^{des}. Then, the ILD (IPD) after sound field reproduction is calculated at the same pairs of sampling points and arranged in another vector ILD^{rep}. Finally, the ILD error (in dB) is calculated as:

$$\text{ILD (IPD) Error (dB)} = 10 \log_{10} \frac{||\text{ILD}^{\text{rep}} - \text{ILD}^{\text{des}}||_2^2}{||\text{ILD}^{\text{des}}||_2^2}. \tag{5.21}$$

In the next set of experiments, the performance of the SFR system will be investigated in two different scenarios. In these scenarios the radiation patterns are designed ahead of time assuming free space conditions, but they are employed for reverberant rooms. Therefore, the derived radiation patterns are independent of the room size and the material used in the room (reflection coefficients), and they only depend on the relative possible locations of talkers and listeners. The reflection coefficients of all walls are considered to be equal.

In Scenario 1, two listeners are located at $(-1.5, 0, 2)$ and $(1.5, 0, 2)$, and two talkers are located at $(1, 0, -0.5)$ and $(-1, 0, -0.5)$. Both the listeners' and talkers' positions are assumed to be known in advance. In each cube around the listeners' ears, the number of virtual sampling points is $M_v = 27$. The loudspeaker radiation patterns are optimized for the two possible locations of the talkers ($W = 2$) and for 12 frequency bins ($Y = 12$), specifically f_y is changing between 200 Hz and 4000 Hz in steps of 345 Hz. The number of loudspeakers assigned to each position-frequency pair is $n_{w,y} = 2$. The designed radiation patterns are examined in five different reverberant rooms with reflection coefficients of 0, 0.2, 0.4, 0.6, 0.8. The average, minimum, and maximum errors across all five rooms are shown in Fig. 5.7 when the second talker is active. The average error is shown as the

curve, while minimum and maximum errors are shown as error bars. In this experiment $p_{\max} = 10^{-4}$.

According to Fig. 5.7(a), the optimized system outperforms the benchmark by 20 to 3 dB for frequencies less than 4 kHz for which the expansion coefficients of higher-order loudspeakers are optimized, in terms of the HRTF-based error. As seen from Fig. 5.7(b), the ILD error is smaller in the optimized system, and in both systems the ILD error is less than 10 dB for frequencies less than 2500 Hz.

It should be noted that, based on the relative locations of the listeners and talkers, the IPD is ambiguous for frequencies above 1102 Hz, but it represents the true ITD for lower frequencies. As shown in Fig. 5.7(c), the IPD error of the optimized system is improved between 3 dB to 40 dB for this frequency range in comparison with the benchmark.

Note that both systems perform better at lower frequencies. The reason is as follows. There are $M_s = 2N_2M_v = 2 \cdot 2 \cdot 27 = 108$ virtual sampling points (size of $\mathbf{p}^{\text{des}}$), and only $N = 48$ loudspeakers (size of $\mathbf{s}^{\text{opt}}$). Hence, the system in Eq. (5.19) is under-determined. However, at low frequencies, the pressure values do not differ much at neighboring virtual sampling points, which leads to linearly (almost) dependent equations in Eq. (5.19) and consequently makes the number of linearly independent equations closer to the number of unknowns. At higher frequencies this is no longer the case, so the performance suffers.

In Scenario 2, the listeners' locations are not known exactly in advance. They are presumed to be somewhere in the red parallelepiped in Fig. 5.5, whose volume is delimited by $(\pm 1.6, \pm 0.05, 2 \pm 0.05)$. In this case, $K = 200$ virtual sampling points are distributed across the volume of possible listeners' locations (50 samples uniformly in x-direction and 2 samples uniformly in y and z-direction) to optimize the expansion coefficients of loudspeakers using Algorithm 7. In this scenario several possibilities are also considered for the talkers locations. Specifically, 6 possibilities ($W = 6$) are considered for the talkers' locations: $\mathbf{x}_w^t \in \{(1,0,0),(1,0,-1),(1,0,-2),(-1,0,0),(-1,0,-1),\ \ (-1,0,-2)\}$, eight frequencies of interest ($Y = 8$) are considered in the loudspeaker pattern design ($f_y \in \{500, 1000, 1500, 2000, 2500, 3000, 3500, 4000\}$ Hz), and $n_{w,y} = 1$ loudspeaker is assigned to each location-frequency pair.

Fig. 5.8 shows the results. To generate the results, two listeners are placed at $(-1.5, 0, 2)$ and $(1.5, 0, 2)$, an active talker is at $(-1, 0, -1/2)$, and $p_{\max} = 10^{-4}$. As in the previous experiment, the results are presented for five reverberant rooms with reflection coefficients between 0 and 0.8. Again, the optimized system achieves better performance compared to the benchmark at lower frequencies because the expansion coefficients are optimized at these frequencies. For this arrangement, again the IPD is ambiguous for frequencies greater than 1102 Hz, and the IPD error of the optimized method outperforms the benchmark in this frequency range (Fig. 5.8(c)) .

As mentioned earlier, the ITD is important for sound localization at frequencies less than 1500 Hz. In the next experiment, the sound field is recreated using the loudspeaker patterns

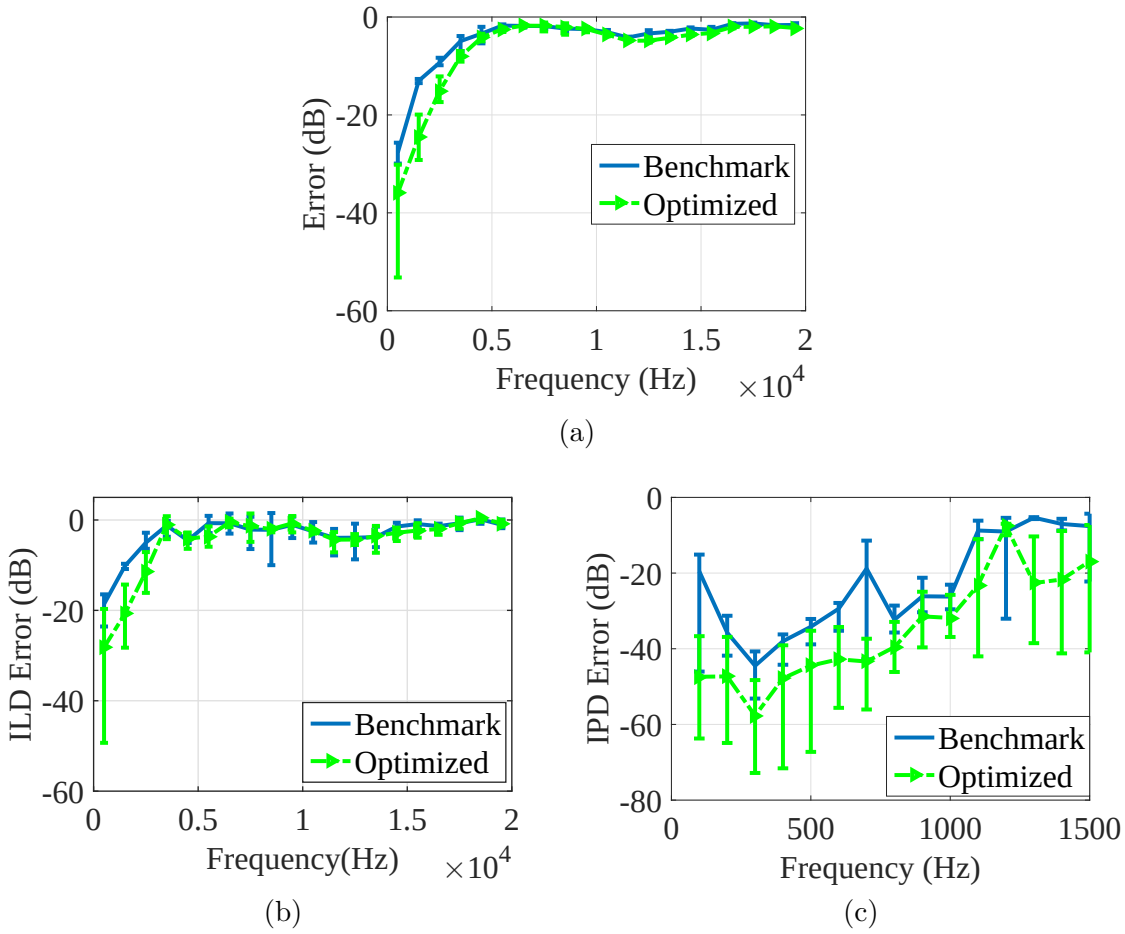


Figure 5.7: (a) HRTF-based reproduction error, (b) ILD error, and (c) IPD error for Scenario 1.

from Scenario 2, with an active talker at $(-1, 0, -0.5)$ and two listeners at $(-1.5, 0, 2)$ and $(+1.5, 0, 2)$ in a reverberant room with reflection coefficients of 0.65. For each listener, the ITD is calculated by finding the time index that maximizes the Interaural Cross Correlation (IACC) coefficient between the signals at the two ears [112]. The results of this experiment are shown in Fig. 5.9 for $p_{\max} = 10^{-4}$. According to this figure, the optimized system allows the synthesized field to match the desired ITD for both listeners, while the benchmark configuration does not perform as well as the optimized system.

The next experiment shows the reproduction error and the ILD error of the optimized system in Fig. 5.10 in terms of the length of the cubic listening area across the frequency range when the reflection coefficients of walls are $R = 0.5$ and $p_{\max} = 10^{-4}$. In this test, the radiation patterns designed for the second scenario are used, and the number of sampling points is fixed for all sizes of the cubic listening area. The results of this test show that the precise approximation of the ear locations results in less reproduction and ILD errors across the whole audio range. For example, when the length of listening area is 2 cm the

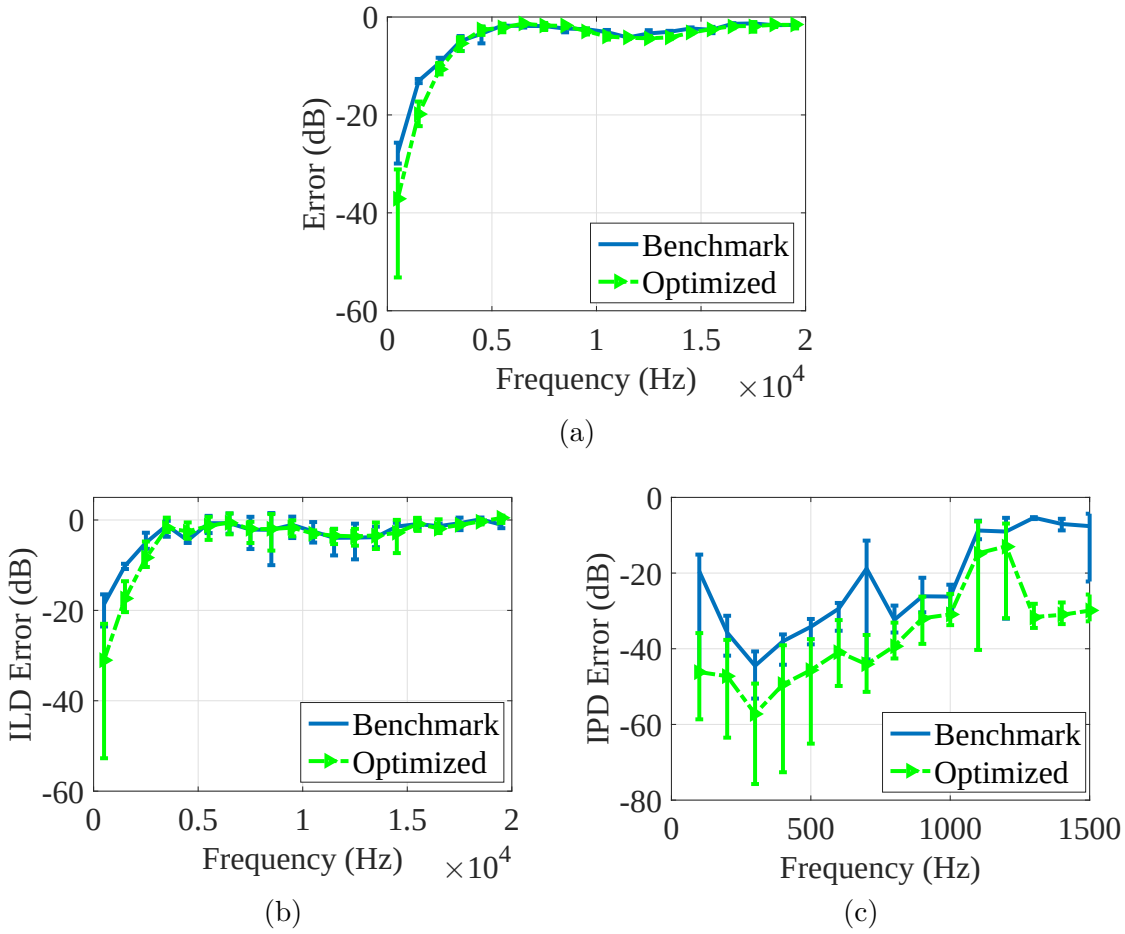


Figure 5.8: (a) HRTF-based reproduction error and (b) ILD error, and (c) IPD error for Scenario 2.

reproduction error is less than -20 dB and the ILD error is less than -10 dB for frequencies less than 10 KHz.

5.5.2 Comparison with other SFR systems

In this section, first, the performance of our SFR system with the optimized radiation patterns will be compared against the SFR methods in [60,61]. These two methods work based on the Kirchhoff Helmholtz (KH) integral to design the radiation patterns of loudspeakers. In these methods, monopole and radial dipole loudspeakers are located all around the listening area on a surface of a sphere. In [60], as in our proposed method, the radiation patterns of loudspeakers are designed in advance, and during the system operation only their excitation (weights) vectors change based on Higher Order Ambisonics (HOA). Therefore, in this method there is one static DoFs per loudspeaker, and the number of dynamic DoFs is equal to the number of loudspeakers. In [61], the radiation patterns of all loudspeakers

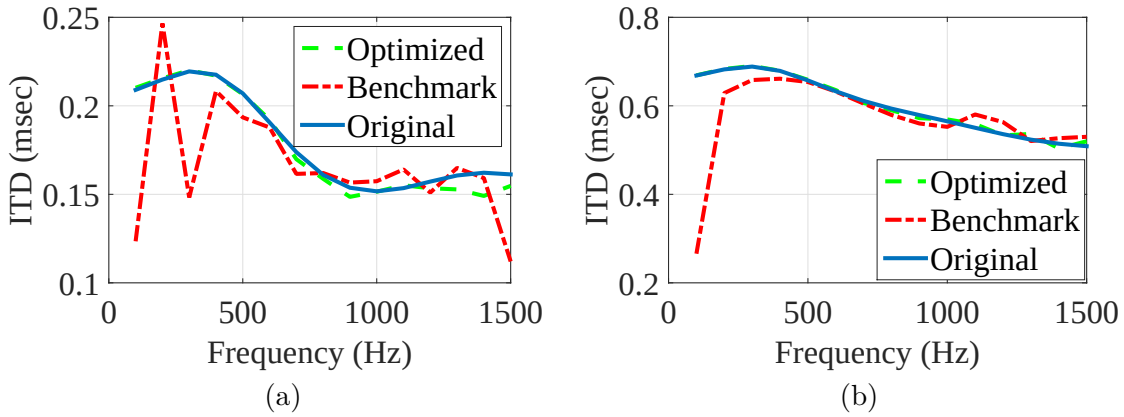


Figure 5.9: ITD for the listener located at (a) $(-1.5, 0, 2)$, (b) $(+1.5, 0, 2)$ across the frequency range.

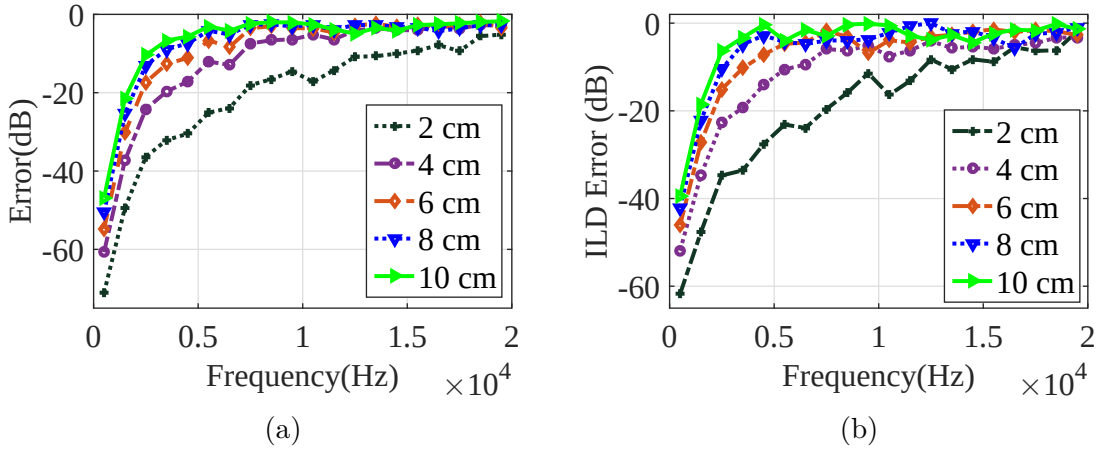


Figure 5.10: (a) HRTF-based error and (b) ILD error of the proposed method versus the length of the cubic listening area.

change during the system operation. It implies that this method has no static DoFs, but the number of dynamic DoFs is twice the number of loudspeakers. To find the radiation patterns (or excitation vector), the HOA technique is employed in this paper as well.

For comparison, we assume that the listening area is a sphere whose center is at $(0, 0, 2.25)$ with a radius of 30 cm, and the primary source is at $(-1, 0, -0.5)$. The number of loudspeakers in [60,61] is 144, and they are located on a concentric sphere with the listening area with a radius of 1.5 m. For our method, loudspeakers are arranged in a rectangular array (Fig. 5.3) on the x-y plane, with the center at the origin, so its distance to the center of the spheres is 2.25 m. To find $\mathbf{s}^{\text{opt}}$ in Eq. (5.19), the number of virtual sampling points (dimension of $\mathbf{p}_2^{\text{des}}$) in the spherical listening area is 125 in case (a), and 27 in case (b) and case (c). However, once the field is synthesized, the error is calculated at 1000 points in the listening area. To be consistent with the methods in [60,61], the simulations are performed

in free-space conditions, the HRTF is not considered for sound field reproduction, and no power constraint is taken into account. Fig. 5.11 shows the results of a comparison of three cases.

Case (a): Fig. 5.11(a) is obtained using the same default parameters employed in the simulations in each paper. For the methods from [60,61], the number of loudspeakers is 144, and the truncation order in HOA is 10. In our method, the number of loudspeakers is $N = 48$, their order is $L = 5$, and the number of sampling points to specify the desired field is 125. The radiation patterns for our system are the ones designed in Scenario 2 (Section 5.5.1). Based on Fig. 5.11(a), the methods from [60, 61] perform better than our method, however at the expense of higher complexity and using 3 times as many loudspeakers (144 vs. 48).

Case (b): Since the comparison in case (a) can be argued to be unfair due to different run-time complexities of the systems involved, for the simulations in Fig. 5.11(b), the number of dynamic DoFs in the three systems is made approximately equal. Specifically, the number of dynamic DoFs remains 48 in our method, as in case (a), with 48 loudspeakers. The number of loudspeakers for the system from [60] is reduced to 49 to make its number of dynamic DoFs equal to 49. Finally, since the system in [61] has two dynamic DoFs per loudspeaker, we reduce its number of dynamic DoFs to 25, making the total number of its dynamic DoFs equal to 50. Further, to (approximately) match the size of matrices involved in the computation of driving signals for the loudspeakers (Eqs. (5.19) in our case, (29) in [60], (23) in [61]), we set the number of sampling points to specify the desired field to 27 in our case, and set the truncation order of HOA to 4 for [60, 61]. This makes the run-time complexity of the three systems almost the same. The reproduction error for this case is shown in Fig. 5.11(b). According to this figure, our system outperforms the other two, especially at low frequencies.

Case (c): In this case, we attempt to match the overall (not just run-time) complexity of the systems involved by matching their total number of DoFs, static plus dynamic. For our system we set the number of loudspeakers to $N = 48$ and the loudspeakers are composed of one monopole and one dipole (same as in [60, 61]), with the dipoles aligned with the z-direction. The gain coefficients of the monopoles and dipoles are found through Algorithm 6 ahead of time as in Scenario 2 (Section 5.5.1). Therefore, there is 1 static DoF per loudspeaker, and the total number of static DoFs is 48. The number of dynamic DoFs is also equal to 48 (the number of loudspeakers), so the total number of DoFs is $48 + 48 = 96$. For the system from [60], the number of loudspeakers is set to 49. As mentioned above, for [60], both the number of static DoFs and the number of dynamic DoFs is 1 per loudspeaker, so the total number of DoFs is $49 + 49 = 98$. In the system from [61], the number of static DoFs is zero, but the number of dynamic DoFs is 2 per loudspeaker. Hence, setting the number of loudspeakers to 49 makes the total number of DoFs equal to 98 in this system as well. With these parameters, the total number of DoFs (static plus dynamic) is approximately the same in all systems. However, the run-time

complexity of [61] is higher, since it has (approximately) twice the number of dynamic DoFs as the other two systems.

As in case (b), we set the number of sampling points to specify the desired field to 27 in our system, and set the truncation order of HOA to 4 for [60, 61]. The results are shown in Fig. 5.11(c), from which we see that our system again outperforms [60, 61] across the frequency range, but with a smaller margin than in case (b). The performance of all three systems is now closer to each other, which is not surprising, considering that the total number of DoFs have been approximately matched.

From this comparison, it can be concluded that the methods in [60, 61] outperform our approach when they employ a larger number of loudspeakers (static and/or dynamic DoFs) at lower frequencies. However, when the number of dynamic DoFs is matched, our approach works better. When the total complexity (static+dynamic DoFs) is matched among the three systems, their performances become more similar, but our method still has some advantage at lower frequencies. In addition, one practical advantage of our SFR approach is that it utilizes a rectangular array of loudspeakers, as shown in Fig. 5.3, which is easier to install than the spherical arrays utilized in Fig. [60, 61].

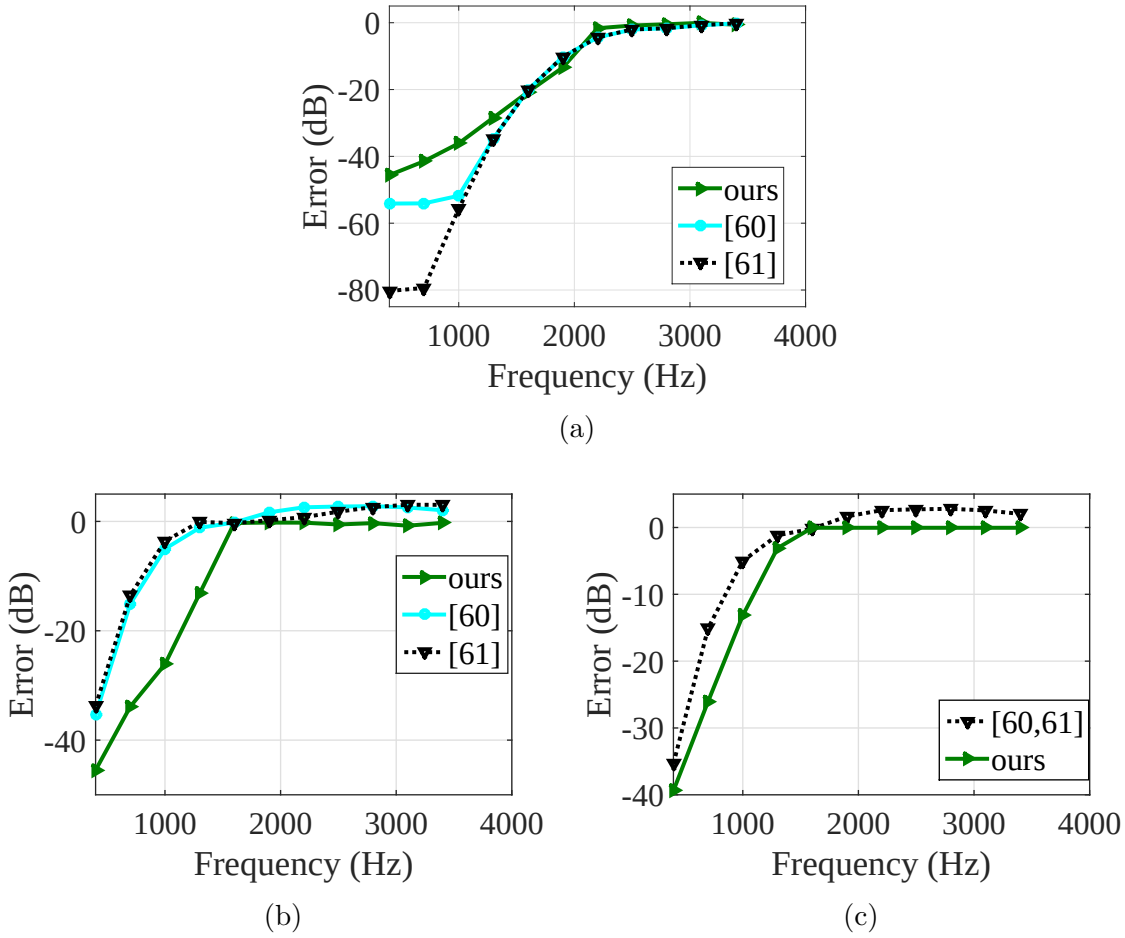


Figure 5.11: Performance comparison in terms of the reproduction error of our SFR method and the methods in [60] and [61] when (a) default parameters from each paper are used, (b) the number of dynamic DoFs is matched, (c) the total number of DoFs is matched.

In the next experiment, the performance of our proposed structure is compared against a linear array of loudspeakers employing Wave Field Synthesis (WFS) to find loudspeaker driving functions [42]. In the first case, a linear array of 48 dipole loudspeakers is located between $(-1.25, 0, 0)$ and $(+1.25, 0, 0)$. It is assumed that two listeners are located at $(0, 0, 2)$ and $(0, 0.5, 3)$, and the active talker is at $(+1, 0, -1/2)$. For our structure, 48 loudspeakers with the order of $L = 5$ (proposed method) and $L = 0$ (benchmark) are located around the screen, and the radiation patterns are the ones selected in Scenario 2. Hence, the number of dynamic DoFs in the three methods (linear array, proposed and benchmark) are equal. Fig. 5.12(a) shows the results of this comparison in free space, without taking power limitation and HRTF into account. Based on this figure, the proposed method and the benchmark outperform the linear array across a range of frequencies because the two listeners are not in the same plane, hence the field produced by the linear array does not provide a good approximation for the desired sound field.

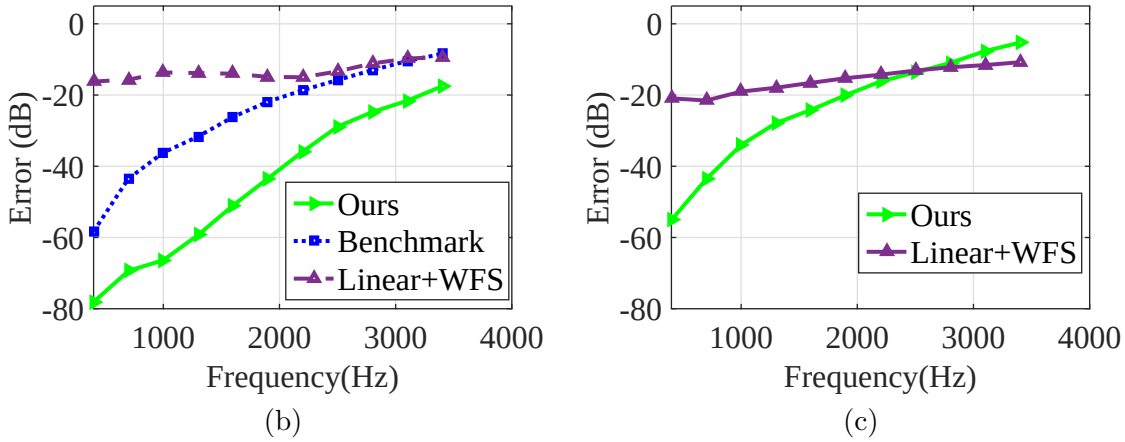


Figure 5.12: Performance comparison in terms of the reproduction error of our SFR method against a linear array + WFS when (a) the number of dynamic DoFs is matched, (b) the number of dynamic DoFs in WFS method is matched with the total number of DoFs in our method.

In the second case, a linear array of 200 loudspeakers is located between $(-3, 0, 0)$ and $(+3, 0, 0)$ and again WFS is used to find the driving functions. The number of dynamic DoFs for this array is therefore 200. For the proposed structure, 20 loudspeakers of order of 2 are used. The number of static DoFs is 9 per loudspeaker (180 for the whole array) and the number of dynamic DoFs is 20. Hence, while the number of dynamic DoFs in the linear array is equal to the total number of DoFs (static and dynamic) in our structure, it is 10 times higher than the number of dynamic DoFs in our structure, resulting in 10 times the run-time complexity. The results of this comparison are shown in Fig. 5.12(b). According to this figure, the proposed structure has better performance at frequencies up to about 2700 Hz. Again, the reason is that the linear array of loudspeakers cannot provide a good approximation for the 3-D sound field.

Finally, the effects of changing the number of static and dynamic DoFs in our method are investigated. First, the HRTF-based reproduction error in Scenario 2 is shown in Fig. 5.13(a) when there are $N = 48$ loudspeakers in the array, and their orders are $L = 2$ and $L = 5$. Therefore, there is a factor of 4 difference in the number of static DoFs ($(L+1)^2$) between these two cases. In Fig. 5.13(b), the order is $L = 5$, while the number of loudspeakers is $N = 12$ and $N = 48$, so in this case there is a factor of 4 difference in the number of dynamic DoFs. These results show that increasing the number of dynamic DoFs (number of loudspeakers, N) has a larger effect on improving the system performance than increasing the number of static DoFs (loudspeaker order, L). Intuitively this makes sense, because the static DoFs are assigned based on the global parameters such as the range of locations of the listeners and talkers, and the range of frequencies of interest, while dynamic DoFs are updated based on the exact desired sound field.

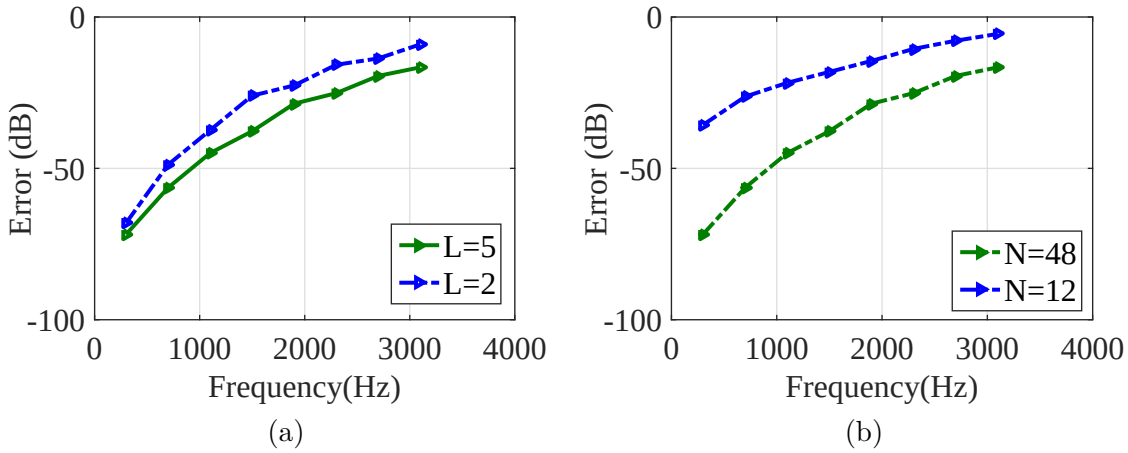


Figure 5.13: Performance comparison in terms of HRTF-based reproduction error of our SFR method in free space for (a) the same number of dynamic DoFs ($N = 48$), different number of static DoFs ($L = 2$ and $L = 5$), (b) the same number of static DoFs ($L = 5$), different number of dynamic DoFs ($N = 12$ and $N = 48$).

5.5.3 Subjective testing

In this section we describe two subjective tests conducted to assess the quality of the sound field produced by the proposed SFR method. These tests involved 12 participants (7 males, 5 females) aged between 22 and 36, all with normal hearing. All participants were trained before the test, and they passed the pre-screening and post-screening phase [113, 114]. In the first test, 10 audio excerpts (5 male voice recordings and 5 female voice recordings), each 5 seconds long, were played to the participants using stereo headphones (Sony MDR-NC7). The participants were asked to assess the quality and the sense of the direction with a grade between 1 (bad quality) and 100 (excellent quality) in comparison with the reference signal. A schematic of the user interface along with the grading scale is shown in Fig. 5.14. The four audio files include the reference (original), anchor (low pass filtered version of the reference clip [113]), one sound file produced by the radiation patterns designed in Scenario 2 (Section 5.5.1), and the sound file created by the benchmark system.

It is assumed that the active talker is at $(-1, 0, -0.5)$, the listener is located at $(1.5, 0, 2)$, and the reflection coefficient of all walls is 0.7. For six of the 10 audio excerpts the maximum power was set to $p_{\max} = 10^{-5}$ and for the other four $p_{\max} = 10^{-3}$. To create the reference signal, the original audio clip is played at the talker's location in virtual room 1, and the sound field is sampled at positions of the listener's ears in room 2. Although in our objective evaluation the HRTF of a human head was obtained by solving a wave equation on the rigid sphere, it is not truly correct to employ that function in the subjective test because HRTF varies from one person to another.

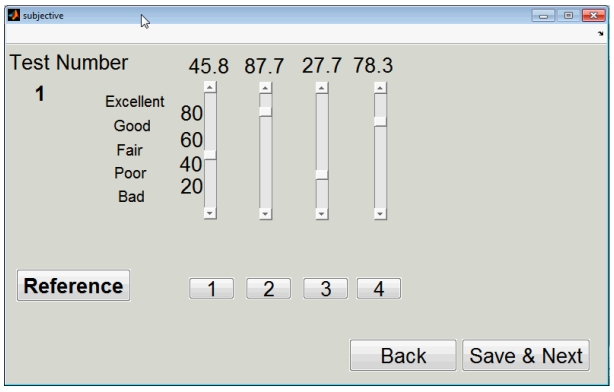


Figure 5.14: User interface employed in the subjective test.

In our test, the individualized HRTF for each participant is derived by the method described in [115] which finds the closest HRTF to each subject. This method uses the CIPIC database [116] which is composed of 40 HRTFs from 40 different subjects (left and right ears). The method proposed in [115], first, measures the anthropometric information of the subjects, such as pinna height, pinna width, and cavum concha height. Then this information is arranged in a vector, and the closest HRTF to the subject is found by the classification method in [115].

After finding the individualized HRTF, the two sampled sound signals are recorded and played to the participants via stereo headphones by applying the individualized HRTF in Eq. (5.19). The sound files for our SFR system and the benchmark system are created by sampling the sound field produced by the loudspeaker array (with higher order sources in our method, and omni-directional sources in the benchmark) at the same two locations (listener's ears) as the reference. Again, the sampled signals are recorded and played out via stereo headphones.

Following the recommendations in [113], for each of the 10 excerpts, the participants were asked to grade the quality and sense of direction for all four sound files by moving the slide bars shown in Fig. 5.14. They were able to play the audio files as many times as they wanted and in any order, until they reached a final decision. They knew that the reference signal was included among the six test files, but were not told which one was the reference signal.

Fig. 5.15 shows the average score across all audio sequences, subjects, and excerpts , as well as the 95% confidence interval for each score [113]. These scores are given in Appendix B. As seen in this figure, on average, the participants evaluated the quality of our proposed method "Excellent" while the average grade was "Good" for the benchmark configuration.

In the second test one single tone audio file with a frequency of 500 Hz is considered as the reference file. Six audio files were played for subjects for comparison with the reference file. The amplitude of the single tone frequency is equal in all audio files while their

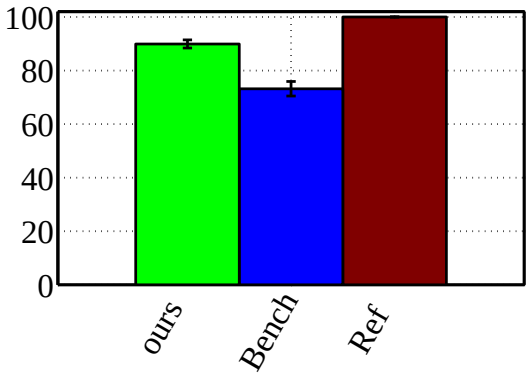


Figure 5.15: The average score and 95% confidence interval of the subjective test results.

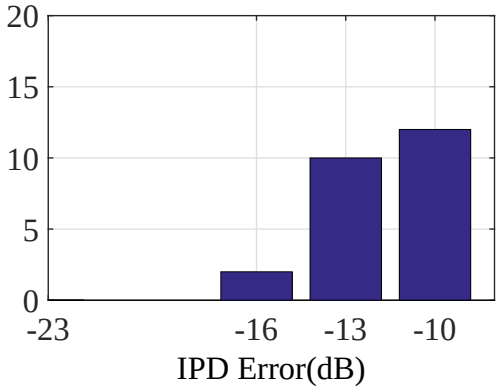


Figure 5.16: The number of subjects who reported the direction of arrival has changed versus the IPD error.

phases at the two ears are replaced by those recreated by our proposed structure when the order of loudspeakers is 0. To create these six files, the number of loudspeakers was $N \in \{4, 8, 12, 20, 24, 48\}$ with IPD errors in the range of -10 dB to -46 dB. The subjects were asked whether or not they felt the direction of sound had changed in comparison with the reference signal. The number of subjects reporting the change in the direction of the sound versus the IPD error is shown in Fig. 5.16. These results suggest that the IPD error of -10 dB is highly audible, whereas IPD errors below -16 dB seem to be mostly inaudible. Relating this back to the IPD error results in Figs. 5.7(c) and 5.8(c), we can conclude that the optimized system extends the range of frequencies over which the IPD error is inaudible by up to 1000 Hz compared to the benchmark SFR system.

5.6 Conclusion

The performance of the audio layer of an audiovisual immersive communication system was studied through numerical and subjective experiments under a variety of conditions.

At the transmitting end, active talker detection was analyzed. A bound on the maximum tolerable error in lip detection that would allow accurate active talker detection was derived analytically, and subsequently verified via simulation.

At the receiving end, the sound field from active talker(s) was synthesized to match the virtual positions of the talker(s) in the 3-D visual scene. The radiation patterns were designed by the pattern selection algorithm presented in Chapter 4. In order to preserve the ILD and ITD after sound field reproduction, the excitations of loudspeakers were derived by applying the HRTF to the ATF matrix and the desired vector. The fidelity of the SFR system is measured by the HRTF-based reproduction error, the IPD and ILD errors. The simulation results show that the fidelity of the sound field produced by the optimized loudspeakers is better than that produced by a benchmark system employing omni-directional loudspeakers.

In contrast to the simulations of the previous chapters, all simulations were performed for reverberant rooms. Error reduction in the range of 2-20 dB was observed for the reverberant rooms. In addition, the quality of the sound produced by the optimized system was evaluated as "Excellent," while the benchmark received a grade of "Good", in the subjective tests.

The next chapter concludes this thesis. The deficiencies of the proposed methods, and their possible solutions will be introduced as possible future work.

Chapter 6

Conclusions and Future Work

6.1 Summary and conclusion

The goal of this thesis was to improve the quality of sound field reproduction systems. The approach was to optimize the static degrees of freedom (locations and patterns of loudspeakers) rather than choosing them intuitively. For this purpose, in Chapter 1, various methods of the SFR systems were reviewed and notable papers in this field were summarized. Wave Field Synthesis, Higher Order Ambisonics, and Direct Approximation are three main methods in finding the excitation of loudspeakers in SFR. The Direct Approximation method is used in this thesis since it is applicable to any geometry. Loudspeaker placement and pattern optimization are two other approaches for SFR that minimize the sound field reproduction error. Motivating applications of SFR systems were also presented with a special focus on immersive communication.

In Chapter 2, selecting the appropriate locations of the loudspeakers on the loudspeaker region was investigated. The main issue in finding the optimum locations is that the reproduction error is not convex in terms of the locations of the loudspeakers, so an algorithm is required to find good, albeit sub-optimal, loudspeaker locations. The existing placement methods, i.e., Gram-Schmidt-based and Lasso-based placement methods, were summarized in this chapter. Two methods for loudspeaker placement were presented. The first method sought the locations of loudspeakers with the help of an "ideal" non-realizable ATF matrix obtained after singular value decomposition of the ATF of a benchmark loudspeaker configuration. The locations of loudspeakers were then selected such that their realizable ATF matrix approaches the ideal one. The second method was based on the Constrained Matching Pursuit (CMP) algorithm. This iterative method produces a sparse representation of the input vector in terms of the dictionary members under power constraint. Here, the ATF of the candidate locations were considered as dictionary members, and the desired field was considered as the input vector. In contrast to the existing placement methods, these placement methods considered power constraint in selecting the locations of loudspeakers.

In Chapter 3, the performances of the four placement algorithms were compared. First, the role of power limitation in reproducing the sound field inside and outside the listening area was studied. A higher power limit leads to a more accurate sound field inside the listening area, and excess field outside of the listening area. Second, the optimum locations of the loudspeakers in terms of the frequency and maximum normalized power were studied. These depend not only on the frequency but also on the maximum normalized power. At lower power, the optimum placement was more clustered around the ray-cut, however, at higher power, the placement was more dispersed. Then, the selected locations by each of the four algorithms were presented for a specific and useful scenario, and the reasons of their behavior were given based on the mathematical process in each algorithm. Finally, the error performance of the placement methods was evaluated for frequencies less than 2000 Hz for different values of the maximum normalized power and the size of the listening area. The results of this comparison were that the performance of the CMP-based method is better than other placement methods. This is because it takes the power constraint into account. Without power limitation, the Lasso outperformed other placement methods.

In Chapter 4, the radiation patterns of higher order loudspeakers were optimized to minimize the SFR error. The expansion coefficients of the loudspeakers were found based on the Constrained Matching Pursuit algorithm. Then, a multi-frequency version of the pattern selection algorithm was presented which worked based on the frequency range and possible locations of the primary source. Joint optimization of the placement and pattern algorithm was also introduced. This algorithm was based on two CMP-based algorithms for optimizing the pattern and locations of loudspeakers. The outer CMP algorithm finds the locations of the loudspeakers, then in order to find the expansion coefficients of the higher order loudspeaker, the selected loudspeaker was fed into the inner CMP algorithm. Finally, the performance of the placement-only, pattern selection, and joint optimization methods were compared for a single tone and for multi-frequency sources. For a single tone primary source, the quality of the joint optimization method was better than other methods because the number of dynamic DoFs in this method was larger than that of the other methods. For a multi-frequency scenario, the locations and expansion coefficients of loudspeakers were fixed during system operation, so the dynamic DoFs of all methods were equal. In this case, although the complexity of all methods was the same, the joint optimization method outperforms other algorithms because it had a larger number of static DoFs, i.e., DoFs in the design phase.

In Chapter 5, the pattern selection algorithm was employed to design the radiation patterns of the loudspeakers for SFR application in immersive communication. To reproduce the sound field, in one room the sound field was captured by an array of microphones, and the talkers' locations were detected with the help of a video-based monitoring system. The effect of the detection error of the lip centroid on the discrimination algorithm was studied, and an error bound was derived. A new algorithm was proposed to find the talkers' exact

locations and to separate their corresponding audio signals based on the calculated error bound. After this step, the audio signals of the talkers along with their corresponding locations were transmitted to the receiving room for sound field reproduction. At the receiving end, the radiation patterns of loudspeakers were optimized ahead of time based on the possible fixed locations of the talkers and listeners and the speech frequency range. During system operation, the sound field was reproduced around the listeners' heads by considering the human head as a rigid sphere and applying its transfer function, which includes the receiving pattern, on finding the complex amplitudes of the loudspeakers. The performance of this SFR method was compared with sound field reproduction with the HOA and WFS methods. In the HOA method, the secondary sources were located on the surface of a sphere, while in the WFS method a linear array of loudspeakers was used for SFR. The comparison showed, for a larger number of loudspeakers, the HOA method outperformed our presented method. However, under the same conditions, the performance of our method was better than the HOA method. In addition, the performance of our method was much better than that of the WFS methods when the listeners were at different heights, as expected. In contrast to the simulations of the previous chapters, all simulations were performed for a reverberant room which was modeled using the image method. Since the SFR method was used for immersive communication, the performance of the SFR system was evaluated not only by measurement of the reproduction error but also by the ILD and IPD errors which model how well a listener can localize a sound source. The quality of the system was assessed through subjective testing as well. For this purpose the sound field was recreated based on the personalized HRTF of each subject, and all 12 subjects evaluated the quality of the reproduced sound field as "Excellent".

6.2 Future directions

As future work, the following problems may be considered:

1- Improving the performance of the multi-frequency algorithm: The multi-frequency algorithm proposed in Chapter 4 optimizes the radiation patterns (and locations) of a group of loudspeakers for a given primary source location at one frequency. For this purpose, the frequency range and possible locations of the primary sources were sampled uniformly. However, these points can be selected based on the probability density function of the primary source locations and the frequency spectrum of the primary source. This selection may reduce the average error.

2- Introducing a new multi-frequency algorithm: In the multi-frequency algorithm presented in Chapter 4, the patterns and locations of a group of a secondary sources were optimized for a predetermined location of the primary source at one frequency. It means that this group of secondary sources lead to good performance for the frequency and location for which they were optimized. In other words, they were not necessarily the best

for the other locations and frequencies. A contribution in this direction is an algorithm in which each secondary source is optimized for all possible locations and frequencies of the primary source. One idea is using the dictionary learning approaches [117, 118] to find the best vectors for approximating all desired fields. By this approach we can approximate the desired field and assign the most suitable radiation pattern to loudspeakers from the different members of the dictionary.

3- Implementation of higher order loudspeakers: The main issue with employing higher order loudspeakers is that the loudspeakers have complicated patterns which may be difficult for implementation in practice. Hence, working on this issue makes the usage of this algorithm in real world scenarios practical. To do so, one way is to simplify the obtained radiation patterns such that they can be implemented with existing radiation patterns. Another way to avoid using higher order loudspeakers, in the pattern selection algorithm is to modify this algorithm. For this purpose, a set of available loudspeaker patterns can be considered as dictionary in the pattern selection algorithm. With this method, each loudspeaker in the array can be replaced by one or a combination of dictionary members (available loudspeakers).

4- Multi objective optimization: To reproduce a sound field for immersive communication, one tries to reproduce the sound field such that the localization clues do not change. In the method presented in Chapter 5, the human head was treated as a rigid sphere, and the ATF matrix and the desired field were modified by considering the HRTF. With this method, the ILD and IPD of the reproduced field were close to those of the desired field. Another possibility is finding the complex amplitudes of loudspeakers such that the reproduction error, the ILD error, and the IPD error are all minimized. For this purpose, the cost functions are the reproduction error, the ILD and ITD errors (as defined in Chapter 5) at the sampling points. This problem can be solved by multi-objective optimization methods [119–122], and the results are expected to lead to less ILD and IPD errors.

5- Modeling a furnished room: The experiments of Chapter 5 were simulated for a reverberant room. However, the effect of furniture was not considered in our simulations. Introducing a model for a furnished room and applying it on our proposed methods would give us better intuition as to how the proposed algorithm will work in real world scenarios.

6- Coupling between loudspeakers: In the modeled SFR system in this thesis and most of SFR literature, it is assumed that all loudspeakers in the array work independently and they do not have any effect on each other. However, any mutual coupling between loudspeakers has influence on the SFR performance, and they should be considered in a modeled SFR system.

7- System implementation: All SFR methods in this thesis were modeled SFR, and they were not implemented in real world scenarios. The next important step of this project, is implementation of the placement and pattern selection algorithms and evaluation of these methods by subjective testing.

Bibliography

[1] W. B. Snow. Basic principles of stereophonic sound. IRE Trans. Audio, 3:42–53, Mar. 1955.

[2] J. G. Apostolopoulos, P. A. Chou, B. Culbertson, T. Kalker, M. D. Trott, and S. Wee. The road to immersive communication. Proceedings of the IEEE, 100(4):974–990, Apr. 2012.

[3] M. Zimmermann. The nervous system in the context of information theory. In Human physiology, pages 166–173. Springer, 1989.

[4] E. Steinbach, S. Hirche, M. Ernst, F. Brandi, R. Chaudhari, J. Kammerl, and I. Vittorias. Haptic communications. Proceedings of the IEEE, 100(4):937–956, 2012.

[5] D. Boothroyd. Touch, time and technics levinas and the ethics of haptic communications. Theory, Culture & Society, 26(2-3):330–345, 2009.

[6] M. Reiner. The role of haptics in immersive telecommunication environments. IEEE Trans. Circuits and Systems for Video Technology, 14(3):392–401, 2004.

[7] E. Steinbach, S. Hirche, J. Kammerl, I. Vittorias, and R. Chaudhari. Haptic data compression and communication. IEEE Signal Processing Magazine, 28(1):87–96, 2011.

[8] Y. Huang, J. Chen, and J. Benesty. Immersive audio schemes. IEEE Signal Processing Magazine, 28(1):20–32, Jan 2011.

[9] C. Kyriakakis. Fundamental and technological limitations of immersive audio systems. Proceedings of the IEEE, 86(5):941–951, May 1998.

[10] S. Enomoto, Y. Ikeda, S. Ise, and S. Nakamura. 3-d sound reproduction system for immersive environments based on the boundary surface control principle. In Virtual and Mixed Reality-New Trends, pages 174–184. Springer, 2011.

[11] H. Teutsch, S. Spors, W. Herbordt, W. Kellermann, and R. Rabenstein. An integrated real-time system for immersive audio applications. In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03), pages 67–70, Oct 2003.

[12] A. Sontacchi, M. Strauss, and R. Holdrich. Audio interface for immersive 3D-audio desktop applications. In IEEE Intl. Symp. Virtual Environments, Human-Computer Interfaces and Measurement Systems (VECIMS'03), pages 179–182, Jul. 2003.

[13] K. U. Doerr, H. Rademacher, S. Huesgen, and W. Kubbat. Evaluation of a low-cost 3D sound system for immersive virtual reality training systems. IEEE Trans. Visualization and Computer Graphics, 13(2):204–212, Mar. 2007.

[14] M. Camras. Approach to recreating a sound field. The Journal of the Acoustical Society of America, 43:1425–1431, Jun. 1968.

[15] G. N. Lilis, D. Angelosante, and G. B. Giannakis. Sound field reproduction using lasso. IEEE Trans. Audio, Speech, and Language Processing, 18:1902–1921, Nov. 2010.

[16] S. Spors, R. Rabenstein, and J. Ahrens. The theory of wave field synthesis revisited. In 124th AES Convention, Amsterdam, May. 2008.

[17] J. Ahrens and S. Spors. Sound field reproduction using planar and linear arrays of loudspeakers. IEEE Trans. Audio, Speech, and Language Processing, 18:2038–2050, Nov. 2010.

[18] J. Ahrens and S. Spors. A comparison of wave field synthesis and higher-order ambisonics with respect to physical properties and spatial sampling. In 125th Conv. of the AES, San Francisco, CA, Oct. 2008.

[19] A. J. Berkhout, D. De Vries, and P. Vogel. Acoustic control by wave field synthesis. The Journal of the Acoustical Society of America, 93:2764–2778, May 1993.

[20] P. Gauthier and A. Berry. Adaptive wave field synthesis for sound field reproduction: Theory, experiments, and future perspectives. In 123th Conv. of the AES, New York, Oct. 2007.

[21] S. Spors. Extension of an analytic secondary source selection criterion for wave field synthesis. In 123th AES Convention, 10 2007.

[22] M. Boone, E. Verheijen, and P. van Tol. Spatial sound-field reproduction by wave-field synthesis. Journal of the Audio Engineering Society, 43(12):1003–1012, 1995.

[23] D. de Vries, E. Start, and V. Valstar. The wave-field synthesis concept applied to sound reinforcement restriction and solutions. In 96th AES Convention, 2 1994.

[24] D. de Vries. Wave field synthesis: History, state-of-the-art and future. In IEEE, Second International Symposium on Universal Communication, Dec. 2008.

[25] S. Spors, H. Buchner, and R. Rabenstien. A novel approach to active listening room compensation for wave field synthesis using wave-domain adaptive filtering. In Proc. IEEE ICASSP'04, volume 4, pages 29– 32, May 2004.

[26] 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 IEEE Second International Symposium on Universal Communication (ISUC'08.), pages 36– 41, Dec. 2008.

[27] N. Kamado, H. Saruwatari, and K. Shikano. Robust sound field reproduction integrating multi-point sound field control and wave field synthesis. In Proc. IEEE ICASSP'11, pages 441 – 444, May 2011.

[28] M. Cobos, J. López, A. Gonzalez, and J. Escolano. Stereo to wave-field synthesis music up-mixing: An objective and subjective evaluation. In IEEE 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP), pages 1279–1284, March 2008.

[29] D. Ward and T. Abhayapala. Reproduction of a plane-wave sound field using an array of loudspeakers. IEEE Trans. Audio, Speech, and Language Processing, 9:697–707, Sep. 2001.

[30] A. Gupta and T. Abhayapala. Three-dimensional sound field reproduction using multiple circular loudspeaker arrays. IEEE Trans. Audio, Speech, and Language Processing, 19:1149–1159, July 2011.

[31] J. Daniel and S. Moreau. Further study of sound field coding with higher order ambisonics. In 116th AES Convention, Berlin, Germany, May 2004.

[32] J. Daniel, S. Moreau, and R. Nicol. Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging. In 114th AES Convention, 3 2003.

[33] J. Ahrens and S. Spors. Focusing of virtual sound sources in higher order ambisonics. In 124th AES Convention, 5 2008.

[34] S. Bertet, J. Daniel, and S. Moreau. 3D sound field recording with higher order ambisonics - objective measurements and validation of spherical microphone. In 120th AES Convention, 5 2006.

[35] J. Ahrens and S. Spors. An analytical approach to sound field reproduction with a movable sweet spot using circular distributions of loudspeakers. In Proc. IEEE ICASSP'09, pages 273–276, May 2009.

[36] J. Ahrens and S. Spors. Analytical driving functions for higher order ambisonics. In Proc. IEEE ICASSP'08, pages 373–376, May 2008.

[37] J. Ahrens and S. Spors. Reproduction of a plane-wave sound field using planar and linear arrays of loudspeakers. In IEEE 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP), pages 1486–1491, March 2008.

[38] P. Gauthier and A. Berry. Sound-field reproduction in-room using optimal control techniques: Simulations in the frequency domain. The Journal of the Acoustical Society of America, 2:662–678, Feb. 2005.

[39] T. Betlehem and C. Withers. Sound field reproduction with energy constraint on loudspeaker weights. IEEE Trans. Audio, Speech, and Language Processing, 19:2388–2392, Oct. 2012.

[40] N. Radmanesh and T. Burnett. Generation of isolated wideband sound fields using a combined two-stage lasso-ls algorithm. IEEE Trans. Audio, Speech, and Language Processing, 21:378 – 387, Feb. 2011.

[41] T. Betlehem and P. D. Teal. A constrained optimization approach for multi-zone surround sound. In Proc. IEEE ICASSP'11, pages 437–440, May 2011.

[42] E. N. G. Verheijen. Sound reproduction by wave field synthesis. PhD thesis, Delft University of Technology, 1998.

[43] J. Ahrens and S. Spors. An analytical approach to 2.5d sound field reproduction employing linear distributions of non-omnidirectional loudspeakers. In Proc. IEEE ICASSP 2010, Texas, USA, March 2010.

[44] J. Ahrens and S. Spors. An analytical approach to sound field reproduction using circular and spherical loudspeaker distributions. Acta Acustica utd. with Acustica, 94:988–999, Nov. 2008.

[45] J. Ahrens and S. Spors. Sound field reproduction employing non-omnidirectional loudspeakers. In 126th AES Convention. Audio Engineering Society, 2009.

[46] J. Ahrens and S. Spors. An analytical approach to 3d sound field reproduction employing spherical distributions of non-omnidirectional loudspeakers. In IEEE International Symposium, on Communications, Control and Signal Processing (ISCCSP), pages 1–5, 2010.

[47] E. G. Williams. Fourier Acoustics. Academic Press, 1999.

[48] M. Poletti. Three-dimensional surround sound systems based on spherical harmonics. Journal of the Audio Engineering Society, 53(11):1004–1025, 2005.

[49] Y. Wu and T. Abhayapala. Theory and design of soundfield reproduction using continuous loudspeaker concept. IEEE Trans. Audio, Speech, and Language Processing, 17(1):107–116, 2009.

[50] Y. Wu and T. Abhayapala. Spatial multizone soundfield reproduction: Theory and design. IEEE Trans. Audio, Speech, and Language Processing, 19(6):1711–1720, 2011.

[51] T. Ajdler. The plenacoustic function and its applications. PhD thesis, École Polytechnique Fédérale de Lausanne, 2006.

[52] O. Kirkeby and P. A. Nelson. Reproduction of plane wave sound fields. The Journal of the Acoustical Society of America, 94(5):2992–3000, 1993.

[53] T. Ajdler, L. Sbaiz, and M. Vetterli. The plenacoustic function and its sampling. IEEE Trans. Signal Processing, 54(10):3790–3804, 2006.

[54] D. P. Bertsekas. Constrained Optimization and Lagrange Multiplier Methods. Academic Press, 1996.

[55] J. Nocedal and S. J. Wright. Numerical Optimization. Springer, 2006.

[56] E. G. GolÊźshteÄŋn and N. V. TretÊźïÿăiïÿąakov. Modified Lagrangians and monotone maps in optimization. John Wiley, 1996.

[57] J. N. Hooker. Integrated Methods for Optimization. Springer, 2007.

[58] F. Asano, Y. Suzuki, and D. C. Swanson. Optimization of control source configuration in active control systems using gram-schmidt orthogonalization. IEEE Trans. Speech and Audio Processing, 7(2):213–220, 1999.

[59] N. Radmanesh and I. Burnett. Generation of isolated wideband sound fields using a combined two-stage lasso-ls algorithm. IEEE Trans. Audio, Speech, and Language Processing, 21(2):378–387, 2013.

[60] M. A. Poletti, F. M. Fazi, and P. A. Nelson. Sound-field reproduction systems using fixed-directivity loudspeakers. The Journal of the Acoustical Society of America, 127(6):3590–3601, 2010.

[61] M. Poletti, F. Fazi, and P. Nelson. Sound reproduction systems using variable-directivity loudspeakers. The Journal of the Acoustical Society of America, 129(3):1429–1438, 2011.

[62] M. A. Poletti and T. D. Abhayapala. Spatial sound reproduction systems using higher order loudspeakers. In Proc. IEEE ICASSP'11, pages 57–60, Prague, May 2011.

[63] P. N. Samarasinghe, M. A. Poletti, S. M. Salehin, T. Abhayapala, and F. M. Fazi. 3D soundfield reproduction using higher order loudspeakers. In Proc. IEEE ICASSP'13, pages 306–310. IEEE, 2013.

[64] S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Processing, 41:3397–3415, Dec. 1993.

[65] Heinrich Kuttruff. Room Acoustics. Taylor and Francis, 2009.

[66] T. Wu and K. Lange. Coordinate descent algorithms for lasso penalized regression. The Annals of Applied Statistics, pages 224–244, 2008.

[67] J. B. Anderson R. Vaughan. Channels, propagation and antennas for mobile communications. Number 50. IET, 2003.

[68] D. E. Manolakis. Efficient solution and performance analysis of 3-D position estimation by trilateration. IEEE Trans. Aerospace and Electronic Systems, 32(4):1239–1248, Oct. 1996.

[69] Q. Liu, Q. Wang, and L. Wu. Size of the dictionary in matching pursuit algorithm. IEEE Trans. Signal Processing, 52(12):3403–3408, Dec. 2004.

[70] R. Gribonval. Fast matching pursuit with a multiscale dictionary of gaussian chirps. IEEE Trans. Signal Processing, 49(5):994–1001, May 2001.

[71] R. Gribonval and E. Bacry. Harmonic decomposition of audio signals with matching pursuit. IEEE Trans. Signal Processing, 51(1):101–111, Jan. 2003.

[72] H. Huang and A. Makur. Backtracking-based matching pursuit method for sparse signal reconstruction. IEEE Signal Processing Letters, 18(7):391–394, July 2011.

[73] L. Rebollo-Neira and D. Lowe. Optimized orthogonal matching pursuit approach. IEEE Signal Processing Letters, 9(4):137–140, April 2002.

[74] L. Daudet. Sparse and structured decompositions of signals with the molecular matching pursuit. IEEE Trans. Audio, Speech, and Language Processing, 14(5):1808–1816, Sep. 2006.

[75] J. Blauert. Spatial hearing: the psychophysics of human sound localization. MIT press, 1997.

[76] S.M. Kuo and D. R. Morgan. Active noise control: a tutorial review. Proceedings of the IEEE, 87:943–973, June 1999.

[77] S. Junsupasen and N. Yodpiji. A survey of workplace noise in an electric power plant. In Proceedings of the Asia Pacific Industrial Engineering and Management Systems, 2012.

[78] P. A. Nelson. Active control of acoustic fields and the reproduction of sound. Journal of Sound and Vibration, 177(4):447–477, 1994.

[79] S. Koyama, K. Furuya, Y. Hiwasaki, and Y. Haneda. Analytical approach to wave field reconstruction filtering in spatio-temporal frequency domain. IEEE Trans. Audio, Speech, and Language Processing, 21(4):685–696, 2013.

[80] G. Strangt. Introduction to Linear Algebra. Wellesley - Cambridge Press, 4th edition, 2009.

[81] L. W. Johnson, R. D. Riess, and J. T. Arnold. Introduction to Linear Algebra. Addison-Wesley, 4th edition, 1998.

[82] J. L. Stratton. Electromagnetic Theory. McGraw-Hill, 1941.

[83] M. Poletti, T. Betlehem, and T. Abhayapala. Higher-order loudspeakers and active compensation for improved 2d sound field reproduction in rooms. Journal of the Audio Engineering Society, 63(1/2):31–45, 2015.

[84] M. Poletti and T. Betlehem. Design of a prototype variable directivity loudspeaker for improved surround sound reproduction in rooms. In Audio Engineering Society Conference: 52nd International Conference: Sound Field Control-Engineering and Perception. Audio Engineering Society, 2013.

[85] M. Poletti, T. Betlehem, and T. Abhayapala. Higher order loudspeakers for improved surround sound reproduction in rooms. In 133th AES Convention. Audio Engineering Society, 2012.

[86] W. T. Chu and A. Warnock. Detailed directivity of sound fields around human talkers. Technical Report IRC-RR-104, National Research Council Canada, Dec. 2002.

[87] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigun. Real-time face detection and motion analysis with application in liveness assessment. IEEE Trans. Information Forensics and Security, 2(3):548–558, Sep. 2007.

[88] D. Nguyen, D. Halupka, P. Aarabi, and A. Sheikholeslami. Real-time face detection and lip feature extraction using field-programmable gate arrays. IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics, 36(4):902–912, Aug. 2006.

[89] D. Yang. High Fidelity Multichannel Audio Compression. PhD thesis, University of Southern California, Aug. 2002.

[90] J. Kammerl, N. Blodow, R.B. Rusu, S. Gedikli, M. Beetz, and E. Steinbach. Real-time compression of point cloud streams. In Proc. IEEE ICRA'12, pages 778–785, May 2012.

[91] K. Murano, S. Unagami, and F. Amano. Echo cancellation and applications. IEEE Communications Magazine, 28(1):49–55, Jan. 1990.

[92] K. Sakhnov, E. Verteletskaya, and B. Simak. Approach for energy-based voice detector with adaptive scaling factor. IAENG International Journal of Computer Science, 36(4):394, 2009.

[93] T. Gerkmann, C. Breithaupt, and R. Martin. Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors. IEEE Trans. Audio, Speech, and Language Processing, 16(5):910–919, 2008.

[94] T. Gerkmann, M. Krawczyk, and R. Martin. Speech presence probability estimation based on temporal cepstrum smoothing. In Proc. IEEE ICASSP'10, pages 4254–4257. IEEE, 2010.

[95] B. Borgström and A. Alwa. Improved speech presence probabilities using hmm-based inference, with applications to speech enhancement and asr. IEEE Journal of Selected Topics in Signal Processing, 4(5):808–815, 2010.

[96] C. Liu and H. Hang. Direction of arrival estimation of speech signals using ica and music methods. In The 5-th IEEE Conference on Industrial Electronics and Applications (ICIEA), pages 1768–1773, June 2010.

[97] J. Dmochowski, J. Benesty, and S. Affes. Direction of arrival estimation using the parameterized spatial correlation matrix. IEEE Trans. Audio, Speech, and Language Processing, 15(4):1327–1339, 2007.

[98] M. Swartling, B. Sällberg, and N. Grbić. Source localization for multiple speech sources using low complexity non-parametric source separation and clustering. Signal Processing, 91(8):1781–1788, 2011.

[99] S. Timofeev, A. Bahai, and P. Varaiya. Wideband adaptive beamforming system for speech recording. In Proc. IEEE ICASSP'07, volume 2, pages II–989. IEEE, 2007.

[100] H. Cox, R. Zeskind, and M. Owen. Robust adaptive beamforming. IEEE Trans. Acoustics, Speech and Signal Processing, 35(10):1365–1376, 1987.

[101] D. Ba, D. Florêncio, and C. Zhang. Enhanced mvdr beamforming for arrays of directional microphones. In IEEE International Conference on Multimedia and Expo, pages 1307–1310. IEEE, 2007.

[102] G. Ballou. Handbook for sound engineers. Taylor & Francis, 2013.

[103] H. Reetz and A. Jongman. Phonetics: Transcription, production, acoustics, and perception, volume 34. John Wiley & Sons, 2011.

[104] J. Eargle. Handbook of recording engineering. Springer, 2005.

[105] M. A. Poletti, T. D. Abhayapala, and P. Samarasinghe. Interior and exterior sound field control using two dimensional higher-order variable-directivity sources. The Journal of the Acoustical Society of America, 131(5):3814–3823, 2012.

[106] R. O. Duda and W. L. Martens. Range dependence of the response of a spherical head model. The Journal of the Acoustical Society of America, 104(5):3048–3058, 1998.

[107] S. Carlile. Virtual auditory space: Generation and applications. Springer - R.G. Landes, 1996.

[108] J. Allen and D. Berkley. Image method for efficiently simulating small-room acoustics. The Journal of the Acoustical Society of America, 65:943–950, 1979.

[109] S. Siltanen, T. Lokki, S. Tervo, and L. Savioja. Modeling incoherent reflections from rough room surfaces with image sources. The Journal of the Acoustical Society of America, 131(6):4606–4614, 2012.

[110] J. Blauert. Spatial hearing: the psychophysics of human sound localization. MIT press, 1997.

[111] E Goldstein. Sensation and perception. Cengage Learning, 2013.

[112] I. Choi, B. G. Shinn-Cunningham, S. B. Chon, and K.-M. Sung. Objective measurement of perceived auditory quality in multichannel audio compression coding systems. Journal of the Audio Engineering Society, 56(1/2):3–17, 2008.

[113] Recommendation ITU-R BS.1534-1. Method for the subjective assessment of intermediate quality level of coding systems. 2003.

[114] Recommendation ITU-R BS.1116-1. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. 1997.

[115] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis. Hrtf personalization using anthropometric measurements. In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03), pages 157–160, 2003.

[116] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The cipic hrtf database. In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'01), pages 99–102, 2001.

[117] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Processing, 54(11):4311–4322, Nov. 2006.

[118] I. Tošić and P. Frossard. Dictionary learning. IEEE Signal Processing Magazine, 28(2):27–38, Mar. 2011.

[119] A. Abraham, L. Jain, and R. Goldberg. Evolutionary Multiobjective Optimization. Springer, 2005.

[120] J. Branke, K. Deb, K. Miettinen, and R. SÅĆowiÅĎski. Multiobjective Optimization, Interactive and Evolutionary Approaches. Springer, 2008.

[121] G. P. Liu, Jian-Bo Yang, and J. F. Whidborne. Multiobjective optimisation and control. Research Studies Press, 2003.

[122] Y. Collette and P. Siarry. Multiobjective optimization: principles and case studies. Springer, 2003.

[123] P. Morse and K. Ingard. Theoretical acoustics. Princeton University Press, 1968.

[124] S. M. Kuo and D. R. MorganR. Active noise control: a tutorial review. Proceedings of the IEEE, 87(6):943–973, 1999.

[125] E. Lehmann and A. Johansson. Diffuse reverberation model for efficient image-source simulation of room impulse responses. IEEE Trans. Audio, Speech, and Language Processing, 18(6):1429–1439, Aug. 2010.

[126] S. Jauk, T. GrubeSa, and H. Domitrovic. Virtual sources method for simulation of indoor acoustics. In IEEE 46th International Symposium Electronics in Marine, pages 283–288, June 2004.

[127] E. Lehmann, A. Johansson, and S. Nordholm. Reverberation-time prediction method for room impulse responses simulated with the image-source model. In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pages 159–162, Oct. 2007.

[128] H. Hashimoto, K. Terai, I. Kakuhari, Y. Nakamura, and H. Sano. Active control system for low frequency road noise combined with an audio system. In 111th AES Convention, 11 2001.

[129] P. Darlington and P. Guiu. Active noise reduction in personal audio delivery systems; assessment using loudness balance methods. In 128th AES Convention, 5 2010.

[130] K. Mayyas. Performance analysis of the deficient length lms adaptive algorithm. IEEE Trans. Signal Processing, 53:2727–2734, Dec. 2005.

[131] M. Chakraborty and H. Sakai. Convergence analysis of a complex lms algorithm with tonal reference signals. IEEE Trans. Speech and Audio Processing, 13:286–292, March 2005.

[132] S. Lane and R. L. Clark. Improving loudspeaker performance for active noise control applications. Journal of the Audio Engineering Society, 46(6):508–519, 1998.

[133] S. Goodman, K. Burlage, S. Dineen, S. Austin, and S. Wise. Using active noise control for recording studio hvac system silencing. In 93th AES Convention, 10 1992.

[134] V. Bartels. Headset with active noise-reduction system for mobile applications. Journal of the Audio Engineering Society, 40(4):277–281, 1992.

[135] A. Thorndal, J. Larsen, and L. G. Johansen. Active road noise reduction in audi a8 w12 with adaptive suspension: A feasibility study. In Audio Engineering Society Conference: 48th International Conference: Automotive Audio, 9 2012.

[136] T. Betlehem and T. Abhayapala. Theory and design of sound field reproduction in reverberant rooms. The Journal of the Acoustical Society of America, 4:2100–2111, April 2005.

[137] Z. Hussain, J. Shawe-Taylor, D. Hardoon, and C. Dhanjal. Design and generalization analysis of orthogonal matching pursuit algorithms. IEEE Trans. Information Theory, 57(8):5326–5341, Aug. 2011.

[138] Y. Luan and F. Jacobsen. A method of measuring the greenâĂŹs function in an enclosure. The Journal of the Acoustical Society of America, 123:40–44, 2008.

[139] T. Okamoto, D. Cabrera, M. Noisternig, B. Katz, Y. Iwaya, and Y. Suzuki. Improving sound field reproduction in a small room based on higher-order ambisonics with a 157-loudspeaker array. In Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics, 2010.

[140] T. Betlehem and T. Abhayapala. A modal approach to soundfield reproduction in reverberant rooms. In Proc. IEEE ICASSP'05, volume 3, pages 288 – 292, May 2005.

[141] F. Toole. Loudspeakers and rooms for sound reproductionâĂŤa scientific review. Journal of the Audio Engineering Society, 54(6):451–476, 2006.

[142] J. Nowak and M. StrauÃ§. Sound field reproduction analysis in a car cabin based on microphone array measurements. In Audio Engineering Society Conference: Automotive Audio, 9 2012.

[143] L. Fincham, A. Jones, and R. Small. The influence of room acoustics on reproduced sound, part 2: Design of wideband coincident-source loudspeakers. In 87th AES Convention, 10 1989.

[144] N. B. Nielsen and A. Celestinosn. Low frequency sound field enhancement system for rectangular rooms using multiple low frequency loudspeakers. In 120th AES Convention, 5 2006.

[145] H. Khalilian, I. V. Bajić, and R. G. Vaughan. Towards optimal loudspeaker placement for sound field reproduction. In Proc. IEEE ICASSP'13, pages 321–325, Vancouver, May 2013.

[146] http://www.gcaudio.com/resources/howtos/loudness.html.

[147] C. Hansen. Fundamentals of acoustics. GA (eds.) Occupational Exposure to Noise: Evaluation, Prevention and Control. World Health Organization, Geneva, 2001.

[148] H. D. Hristov. Fresnal Zones in Wireless Links, Zone Plate Lenses and Antennas. Artech House, Inc., 2000.

[149] H. Khalilian, I. V. Bajić, and R. G. Vaughan. Loudspeaker placement for sound field reproduction by constrained matching pursuit. In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'13), New Paltz, NY, Oct. 2013.

[150] H. Khalilian, I. V. Bajić, and R. G. Vaughan. 3D sound field reproduction using diverse loudspeaker patterns. In IEEE International Conference on Multimedia and Expo Workshops, San Jose, CA, Jul. 2013.

[151] X. Zhu, G. T. Beauregard, and L. Wyse. Real-time signal estimation from modified short-time fourier transform magnitude spectra. IEEE Trans. Audio, Speech, and Language Processing, 15(5):1645–1653, 2007.

[152] Recommendation ITU-R BS.1284-1. General methods for the subjective assessment of sound quality. 2003.

[153] M. P. Norton and D. G. Karczub. Fundamentals of noise and vibration analysis for engineers. Cambridge University Press, 2003.

[154] G. S. Kino. Acoustic waves: Devices, imaging, and analog signal processing. Prentice Hall, 1987.

[155] C. D. Austin, J. N. Ash, and R. L. Moses. Parameter estimation using sparse reconstruction with dynamic dictionaries. In Proc. IEEE ICASSP'11, pages 2852–2855. IEEE, 2011.

[156] Samuel R Atcherson, Clifford A Franklin, and Laura Smith-Olinde. Hearing assistive and access technology. Plural Publishing, 2015.

[157] E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. Communications on pure and applied mathematics, 59(8):1207–1223, 2006.

[158] C. D. Austin, R. L. Moses, J. N. Ash, and E. Ertin. On the relation between sparse reconstruction and parameter estimation with model order selection. IEEE Journal of Selected Topics in Signal Processing, 4(3):560–570, 2010.

[159] H. Wierstorf, C. Hohnerlein, S. Spors, and A. Raake. Coloration in wave field synthesis. In Audio Engineering Society Conference: 55th International Conference: Spatial Audio, pages 1–8. Audio Engineering Society, 2014.

Appendix A

Proofs

A.1 Proof of Lemma 2.5.1

Proof. (Lemma 2.5.1): Since $\mathbf{U}$ is a unitary matrix, any M dimensional vector can be written as a linear combination of its columns. Apply this to the i-th column of $\mathbf{U}^g$, $\mathbf{u}_i^g$:

$$\mathbf{u}_i^g = \sum_{k=1}^{M} a_k \mathbf{u}_k = \mathbf{U}\mathbf{a}, \tag{A.1}$$

where $\mathbf{a} = [a_1, a_2, ..., a_M]^T$. Since $\mathbf{u}_i^g$ is a unit vector, as are $\mathbf{u}_k$'s, the ℓ_2 norm of $\mathbf{a}$ is equal to 1, that is $\|\mathbf{a}\|_2^2 = 1$. In addition, the i-th column of $\mathbf{U}^g$ is the i-th eigenvector of $\mathbf{G}\mathbf{G}^H$, and its corresponding eigenvalue is $(\sigma_i^g)^2$, that is:

$$\mathbf{G}\mathbf{G}^H \mathbf{u}_i^g = (\sigma_i^g)^2 \mathbf{u}_i^g. \tag{A.2}$$

Now replace $\mathbf{G}$ with $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$ and $\mathbf{u}_i^g$ with $\mathbf{U}\mathbf{a}$ on the left hand side of (A.2) to obtain

$$\begin{aligned} \mathbf{G}\mathbf{G}^H \mathbf{u}_i^g &= \mathbf{G}\mathbf{G}^H \mathbf{U}\mathbf{a} \\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H \mathbf{V}\mathbf{\Sigma}^H \mathbf{U}^H \mathbf{U}\mathbf{a} = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^H \mathbf{a}, \end{aligned} \tag{A.3}$$

and also replace $\mathbf{u}_i^g$ with $\mathbf{U}\mathbf{a}$ on the right hand side of (A.2) to obtain

$$(\sigma_i^g)^2 \mathbf{u}_i^g = (\sigma_i^g)^2 \mathbf{U}\mathbf{a}. \tag{A.4}$$

The expressions in (A.3) and (A.4) must be equal, so

$$\mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^H \mathbf{a} = (\sigma_i^g)^2 \mathbf{U}\mathbf{a}. \tag{A.5}$$

Multiply both sides of equation (A.5) by $\mathbf{U}^H$ from the left:

$$\mathbf{U}^H \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^H \mathbf{a} = (\sigma_i^g)^2 \mathbf{U}^H \mathbf{U}\mathbf{a}. \tag{A.6}$$

Since $\mathbf{U}^H\mathbf{U} = \mathbf{I}$, this implies

$$\mathbf{\Sigma}\mathbf{\Sigma}^H \mathbf{a} = (\sigma_i^g)^2 \mathbf{a}, \tag{A.7}$$

which can be rewritten as:

$$\begin{bmatrix} {\sigma_1}^2 a_1 \\ {\sigma_2}^2 a_2 \\ \vdots \\ {\sigma_N}^2 a_N \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} (\sigma_i^g)^2 a_1 \\ (\sigma_i^g)^2 a_2 \\ \vdots \\ (\sigma_i^g)^2 a_N \\ (\sigma_i^g)^2 a_{N+1} \\ \vdots \\ (\sigma_i^g)^2 a_M \end{bmatrix}. \tag{A.8}$$

This equation holds if:

1. a_k's are zero for $k > N$.
2. ${\sigma_k}^2 a_k = (\sigma_i^g)^2 a_k$ for $k \leq N$, which holds under one of the following conditions:
   (a) All a_k's are equal to zero. Based on equation (A.1), it means that i-th column of $\mathbf{U}^g$, which is the i-th eigenvector of the $\mathbf{G}\mathbf{G}^H$, is equal to zero, which is not possible.
   (b) ${\sigma_k}^2 = (\sigma_i^g)^2$ for all $k = 1, 2, ..., N$. This equality is not true in general because σ_k's are selected to be monotonically non-increasing, but otherwise arbitrary.
   (c) There exist an index j for which ${\sigma_j}^2 = (\sigma_i^g)^2$, and $a_k = 0$ for $k \neq j$. According to equation (A.1), it means that $\mathbf{u}_i^g = a_j \mathbf{u}_j$, and since $\|\mathbf{a}\|_2^2 = 1$, a_j is equal to either 1 or -1.

In conclusion, equation (A.8) holds only if conditions 1) and 2c) hold. This means that $\mathbf{u}_i^g$, which is the i-th eigenvector of $\mathbf{G}\mathbf{G}^H$, is either parallel or anti-parallel with the j-th column of $\mathbf{U}$ (that is, $\mathbf{u}_i^g = \pm \mathbf{u}_j$) for $j \leq N$, and in addition, ${\sigma_i^g}^2 = {\sigma_j}^2$. Thus, since all eigenvectors are perpendicular to each other, there is a one to one correspondence between the eigenvectors of $\mathbf{G}\mathbf{G}^H$ and $\mathbf{u}_j$'s with $j \leq N$: for any fixed i, there is an index j such that $\mathbf{u}_j = \pm \mathbf{u}_i^g$ and $\sigma_j = (\sigma_i^g)^2$.

Now the question is what exactly is the correspondence between $\mathbf{u}_i^g$'s and $\mathbf{u}_j$'s? The answer lies in the correspondence between σ_i^g's and σ_j's. Since both of these sequences are arranged in non-increasing order, mapping in order is the only possibility. Therefore, $\mathbf{\Sigma} = \mathbf{\Sigma}^g$ and $\mathbf{u}_i = \pm \mathbf{u}_i^g$. □ □

A.2 Proof of Theorems 2.5.2 and 2.5.3

Proof. (Theorem 2.5.2): For the initial ATF matrix $\mathbf{G}$, since $\mathbf{U}^g$ is a $M \times M$ unitary matrix, every M dimensional vector can be expanded in terms of its columns. Hence, $\mathbf{p}^{\text{des}} = \sum_{i=1}^{M} c_i^g \mathbf{u}_i^g$, where $c_i^g = (\mathbf{u}_i^g)^H \mathbf{p}^{\text{des}}$, and consequently $\sum_{i=1}^{M} |c_i^g|^2 = \|\mathbf{p}^{\text{des}}\|_2^2$. In addition, since the singular values are arranged in non-increasing order and $\gamma > 0$, we have:

$$\frac{\gamma^2}{((\sigma_1^g)^2 + \gamma)^2} \leq \frac{\gamma^2}{((\sigma_2^g)^2 + \gamma)^2} \leq ... \leq \frac{\gamma^2}{((\sigma_N^g)^2 + \gamma)^2} < 1.$$

Using this in (2.17) implies:

$$\begin{aligned}\|\mathbf{p}^{\text{des}}-\mathbf{G}\mathbf{s}\|_2^2 &= \sum_{n=1}^{N}\frac{\gamma^2}{((\sigma_n^g)^2+\gamma)^2}|c_n^g|^2+\sum_{m=N+1}^{M}|c_m^g|^2\\ &\geq \sum_{n=1}^{N}\frac{\gamma^2}{((\sigma_1^g)^2+\gamma)^2}|c_n^g|^2+\sum_{m=N+1}^{M}|c_m^g|^2\\ &= \frac{\gamma^2}{((\sigma_1^g)^2+\gamma)^2}\sum_{n=1}^{N}|c_n^g|^2+\sum_{m=N+1}^{M}|c_m^g|^2.\end{aligned} \tag{A.9}$$

Now use the fact that $\sum_{n=1}^{N}|c_n^g|^2=\|\mathbf{p}^{\text{des}}\|_2^2-\sum_{m=N+1}^{M}|c_m^g|^2$ in (A.9) to obtain:

$$\begin{aligned}&\|\mathbf{p}^{\text{des}}-\mathbf{G}\mathbf{s}\|_2^2\geq\\ &\frac{\gamma^2}{((\sigma_1^g)^2+\gamma)^2}\left(\|\mathbf{p}^{\text{des}}\|_2^2-\sum_{m=N+1}^{M}|c_m^g|^2\right)+\sum_{m=N+1}^{M}|c_m^g|^2=\\ &\frac{\gamma^2}{((\sigma_1^g)^2+\gamma)^2}\|\mathbf{p}^{\text{des}}\|_2^2+\left(1-\frac{\gamma^2}{((\sigma_1^g)^2+\gamma)^2}\right)\sum_{m=N+1}^{M}|c_m^g|^2\\ &\geq\frac{\gamma^2}{((\sigma_1^g)^2+\gamma)^2}\|\mathbf{p}^{\text{des}}\|_2^2=\|\mathbf{p}^{\text{des}}-\mathbf{G}^{\text{ideal}}\mathbf{s}\|_2^2,\end{aligned} \tag{A.10}$$

where the second inequality follows due to $0\leq\frac{\gamma^2}{((\sigma_i^g)^2+\gamma)^2}\leq 1$, and the last equality is from equation (2.25). Hence, for the same excitation vector $\mathbf{s}$ (and therefore the same input power), the reproduction error of $\mathbf{G}^{\text{ideal}}$ is no larger than that of the initial ATF matrix $\mathbf{G}$. □ □

Proof. (Theorem 2.5.3): Let $e_g(\mathbf{s})=\|\mathbf{p}^{\text{des}}-\mathbf{G}\mathbf{s}\|_2^2$ and $e_{\text{ideal}}(\mathbf{s}')=\|\mathbf{p}^{\text{des}}-\mathbf{G}^{\text{ideal}}\mathbf{s}'\|_2^2$. Recall from (2.26) that for the optimal excitation vector of the ideal ATF matrix

$$\|\mathbf{s}'\|_2^2=\frac{(\sigma_1^g)^2}{((\sigma_1^g)^2+\gamma)^2}\|\mathbf{p}^{\text{des}}\|_2^2.$$

The normalized power in equation (2.16) can be rewritten as follows:

$$\begin{aligned}
\|\mathbf{s}\|_2^2 &= \sum_{n=1}^{N} \frac{(\sigma_n^g)^2}{((\sigma_n^g)^2+\gamma)^2}|c_n^g|^2 = \sum_{n=1}^{N} \frac{(\sigma_n^g)^2+\gamma-\gamma}{((\sigma_n^g)^2+\gamma)^2}|c_n^g|^2 \\
&= \sum_{n=1}^{N} \frac{1}{((\sigma_n^g)^2+\gamma)}|c_n^g|^2 - \sum_{n=1}^{N} \frac{\gamma}{((\sigma_n^g)^2+\gamma)^2}|(c_n^g)|^2 \\
&= \sum_{n=1}^{N} \frac{1}{((\sigma_n^g)^2+\gamma)}|c_n^g|^2 - \frac{1}{\gamma}\sum_{n=1}^{N} \frac{\gamma^2}{((\sigma_n^g)^2+\gamma)^2}|c_n^g|^2 \\
&+ \frac{1}{\gamma}\sum_{m=N+1}^{M}|c_m^g|^2 - \frac{1}{\gamma}\sum_{m=N+1}^{M}|c_m^g|^2 \\
&= \sum_{n=1}^{N} \frac{1}{((\sigma_n^g)^2+\gamma)}|c_n^g|^2 + \frac{1}{\gamma}\sum_{m=N+1}^{M}|c_m^g|^2 - \frac{e_g(\mathbf{s})}{\gamma}.
\end{aligned} \tag{A.11}$$

As mentioned before, $\sum_{n=1}^{N}|c_n^g|^2 = \|\mathbf{p}^{\text{des}}\|_2^2 - \sum_{m=N+1}^{M}|c_m^g|^2$ and the sequence $\frac{1}{(\sigma_n^g)^2+\gamma}$ is monotonically non-decreasing in n, so the first two terms of the above equation satisfy in the following inequalities:

$$\begin{aligned}
&\sum_{n=1}^{N} \frac{1}{(\sigma_n^g)^2+\gamma}|c_n^g|^2 + \sum_{m=N+1}^{M}|c_m^g|^2 \geq \sum_{n=1}^{N} \frac{1}{(\sigma_1^g)^2+\gamma}|c_n^g|^2 \\
&+ \sum_{m=N+1}^{M}|c_m^g|^2 = \frac{1}{(\sigma_1^g)^2+\gamma}\left(\|\mathbf{p}^{\text{des}}\|_2^2 - \sum_{m=N+1}^{M}|c_m^g|^2\right) \\
&+ \sum_{m=N+1}^{M}|c_m^g|^2 = \frac{1}{(\sigma_1^g)^2+\gamma}\|\mathbf{p}^{\text{des}}\|_2^2 \\
&+ \sum_{m=N+1}^{M}|c_m^g|^2\left(1 - \frac{1}{(\sigma_1^g)^2+\gamma}\right) \geq \frac{1}{(\sigma_1^g)^2+\gamma}\|\mathbf{p}^{\text{des}}\|_2^2.
\end{aligned} \tag{A.12}$$

Now let the reproduction error of $\mathbf{G}^{\text{ideal}}$ and $\mathbf{G}$ be equal, that is $e_g(\mathbf{s}) = e_{\text{ideal}}(\mathbf{s}') = \frac{\gamma^2}{(\sigma_1^2+\gamma)^2}\|\mathbf{p}^{\text{des}}\|_2^2$. Replacing this value of $e_g(\mathbf{s})$ in (A.11) and combining equations (A.11) and (A.12) results in:

$$\begin{aligned}
\|\mathbf{s}\|_2^2 &= \sum_{n=1}^{N} \frac{1}{((\sigma_n^g)^2+\gamma)}|c_n|^2 + \frac{1}{\gamma}\sum_{m=N+1}^{M}|c_m^g|^2 - \frac{e_g(\mathbf{s})}{\gamma} \\
&\geq \frac{1}{(\sigma_1^g)^2+\gamma}\|\mathbf{p}^{\text{des}}\|_2^2 - \frac{\gamma}{((\sigma_1^g)^2+\gamma)^2}\|\mathbf{p}^{\text{des}}\|_2^2 \\
&= \frac{(\sigma_1^g)^2}{((\sigma_1^g)^2+\gamma)^2}\|\mathbf{p}^{\text{des}}\|_2^2 = \|\mathbf{s}'\|_2^2.
\end{aligned} \tag{A.13}$$

Therefore, for the same reproduction error, the input power required by $\mathbf{G}^{\text{ideal}}$ is no larger than that of $\mathbf{G}$. □ □

A.3 Proof of Theorem 2.6.1

Proof. (Theorem 2.6.1): Let β_n be defined as

$$\beta_n = \frac{p_n}{|(\mathbf{b}^{(n)})^H R^n \mathbf{a}|^2}. \tag{A.14}$$

The approximation error vector at the n-th iteration is $R^{n+1}\mathbf{a} = R^n\mathbf{a} - \alpha_n \mathbf{b}^{(n)}$. Therefore the squared ℓ_2 norm of the error vector at the n-th iteration is:

$$\|R^{n+1}\mathbf{a}\|_2^2 = \|R^n\mathbf{a}\|_2^2 + |\alpha_n|^2 - 2\,\mathrm{Re}\left((\alpha_n \mathbf{b}^{(n)})^H R^n \mathbf{a}\right). \tag{A.15}$$

The value of α_n is given in (2.42). There are two cases: $\sqrt{p_n} \le |(\mathbf{b}^{(n)})^H R^n \mathbf{a}|$ and $\sqrt{p_n} > |(\mathbf{b}^{(n)})^H R^n \mathbf{a}|$. If $\sqrt{p_n} > |(\mathbf{b}^{(n)})^H R^n \mathbf{a}|$ for all n, then CMP reduces to MP in each iteration, so we can simply use the tighter upper bound in (2.38) to prove the theorem. Otherwise, there is at least one n such that $\sqrt{p_n} \le |(\mathbf{b}^{(n)})^H R^n \mathbf{a}|$. For all such n's, $\beta_n \le 1$, so we have

$$\beta_{\min} = \min_n\{\beta_n\} \le 1. \tag{A.16}$$

Now focus on any particular n. First, consider the case $\sqrt{p_n} \le |(\mathbf{b}^{(n)})^H R^n \mathbf{a}|$, and replace the corresponding value of α_n from (2.42) into (A.15). After algebraic manipulation, this leads to

$$\|R^{n+1}\mathbf{a}\|_2^2 = \|R^n\mathbf{a}\|_2^2 - \sqrt{p_n}\left(2|(\mathbf{b}^{(n)})^H R^n \mathbf{a}| - \sqrt{p_n}\right). \tag{A.17}$$

Since $\sqrt{p_n} \le |(\mathbf{b}^{(n)})^H R^n \mathbf{a}|$, the term in the bracket is at least as large as $\sqrt{p_n}$, which leads to the following inequality:

$$\begin{aligned} \|R^{n+1}\mathbf{a}\|_2^2 &= \|R^n\mathbf{a}\|_2^2 - \sqrt{p_n}\left(2|(\mathbf{b}^{(n)})^H R^n \mathbf{a}| - \sqrt{p_n}\right) \\ &\le \|R^n\mathbf{a}\|_2^2 - p_n. \end{aligned} \tag{A.18}$$

Note that I in (2.39) is bounded as $0 < I \le 1$, which means

$$p_n = \beta_n |(\mathbf{b}^{(n)})^H R^n \mathbf{a}|^2 \ge I^2 \beta_n \|R^n\mathbf{a}\|_2^2 \ge I^2 \beta_{\min} \|R^n\mathbf{a}\|_2^2, \tag{A.19}$$

and this in turn means that the right-hand side of (A.18) is bounded by:

$$\begin{aligned} \|R^{n+1}\mathbf{a}\|_2^2 \le \|R^n\mathbf{a}\|_2^2 - p_n &\le \|R^n\mathbf{a}\|_2^2\left(1 - \beta_{\min} I^2\right) \\ &\le \|R^{n-1}\mathbf{a}\|_2^2\left(1 - \beta_{\min} I^2\right)^2 \\ &\le \ldots \\ &\le \|R^1\mathbf{a}\|_2^2 (1 - \beta_{\min} I^2)^n \\ &= \|\mathbf{a}\|_2^2 (1 - \beta_{\min} I^2)^n. \end{aligned} \tag{A.20}$$

Next consider the case $\sqrt{p_n} > |(\mathbf{b}^{(n)})^H R^n \mathbf{a}|$. In this case, the CMP algorithm selects the same value of α_n as MP, and here the corresponding bound from MP [64] can be used:

$$\begin{aligned} \|R^{n+1}\mathbf{a}\|_2^2 &= \|R^n\mathbf{a}\|_2^2 - |(\mathbf{b}^{(n)})^H R^n \mathbf{a}|^2 \\ &\leq \|R^n\mathbf{a}\|_2^2 \Big(1 - I^2\Big) \\ &\leq \|R^n\mathbf{a}\|_2^2 \Big(1 - \beta_{\min} I^2\Big), \end{aligned} \tag{A.21}$$

where the last inequality is due to the fact that $\beta_{\min} \leq 1$. This is the same as the first inequality in (A.20). Following the same chain of inequalities, we again obtain

$$\|R^{n+1}\mathbf{a}\|_2^2 \leq \|\mathbf{a}\|_2^2 (1 - \beta_{\min} I^2)^n. \tag{A.22}$$

Therefore, in both cases, an upper bound on the error in CMP is given by (A.22). Taking the square root, we obtain

$$\|R^{n+1}\mathbf{a}\|_2 \leq \|\mathbf{a}\|_2 \Big(1 - \beta_{\min} I^2\Big)^{n/2}. \tag{A.23}$$

Now let $p_{\min} = \min_n\{p_n\}$ and note that $|(\mathbf{b}^{(n)})^H R^n \mathbf{a}|^2 \leq \|R^n\mathbf{a}\|_2^2 \leq \|\mathbf{a}\|_2^2$. Therefore, from (A.14), $\beta_{\min} \geq p_{\min}/\|\mathbf{a}\|_2^2$. Using this in (A.23) leads to

$$\|R^{n+1}\mathbf{a}\|_2 \leq \|\mathbf{a}\|_2 \left(1 - \frac{p_{\min}}{\|\mathbf{a}\|_2} I^2\right)^{n/2}, \tag{A.24}$$

which proves the theorem. □ □

Appendix B

Subjective test results

Table B.1 shows the results of the subjective test explained in Section 5.5.3. In this table, the scores of each row correspond to a subject for the benchmark configuration and the proposed method, and the scores of each column correspond to an audio excerpt.

Table B.1: Subjective test scores

Sample	1	2	3	4	5	6	7	8	9	10
S01-ours	91	93	85	84	86	90	94	96	100	100
S01-bench	75	73	80	63	63	64	67	70	82	86
S02-ours	93	89	88	79	80	80	80	81	100	95
S02-bench	85	89	50	60	60	82	79	81	90	81
S03-ours	88	87	99	68	76	79	81	100	97	88
S03-bench	40	49	60	61	62	66	68	89	83	80
S04-ours	70	97	71	72	74	75	78	99	95	79
S04-bench	51	77	57	58	60	52	60	86	88	55
S05-ours	100	87	90	87	75	79	100	83	92	94
S05-bench	85	77	40	40	53	49	80	60	77	84
S06-ours	85	89	95	70	71	73	81	86	72	74
S06-bench	86	87	48	45	84	41	56	86	41	39
S07-ours	94	93	91	77	77	79	82	85	82	85
S07-bench	87	83	85	45	47	57	62	47	71	83
S08-ours	89	100	92	96	54	100	100	55	81	82
S08-bench	29	81	59	55	29	80	89	32	32	45
S09-ours	96	93	96	66	68	68	81	85	73	74
S09-bench	82	85	89	50	50	62	62	73	66	69
S10-ours	45	94	86	100	48	48	92	56	98	47
S10-bench	24	81	47	83	29	37	81	42	47	44
S11-ours	90	86	87	100	100	91	79	62	71	76
S11-bench	76	68	66	89	83	82	37	28	42	49
S12-ours	57	59	87	95	73	63	58	63	62	91
S12-bench	35	37	81	83	48	38	37	48	40	83

Appendix C

Acoustic Link Equation

This appendix first explains the physical link equation between a loudspeaker and a sampling point. This link equation is used for the SFR systems in Chapters 2 to 4. Then, the physical link equation is explained for the immersive communication model explained in Chapter 5 between the talkers and microphones in room 1, and between the loudspeakers and listeners in room 2.

C.1 Link Equations between a loudspeaker and a sampling point:

Let $\mathbf{x}_n = (x_n, y_n, z_n)$ be the location of the n-th loudspeaker and $\mathbf{y}_m = (x_m, y_m, z_m)$ be the location of m-th sampling point. Let $V_n(f)$ be the input voltage of the loudspeaker in [V], $\mathcal{L}_n(f, \theta_m, \phi_m)$ be the dimensionless radiation pattern of the n-th loudspeaker, $E(f)$ in [Pa·m/V] be the transfer function of the loudspeaker (which converts voltage to pressure of the loudspeaker at 1 m), and $g_{m,n}(f; \|\mathbf{x}_n - \mathbf{y}_m\|_2)$ be the free space Green's function. The pressure at $\mathbf{y}_m$ is formulated by:

$$\begin{aligned} p_{m,n}(f, \mathbf{x}_n, \mathbf{y}_m) &= V_n^l(f) \cdot E(f) \cdot \mathcal{L}_n(f, \theta_m, \phi_m) \cdot g(f; \|\mathbf{x}_n - \mathbf{y}_m\|_2) \\ &= s_n(f) \cdot \mathcal{L}_n(f, \theta_m, \phi_m) \cdot g(f; \|\mathbf{x}_n - \mathbf{y}_m\|_2), \end{aligned} \tag{C.1}$$

In this equation, (θ_m, ϕ_m) are the spherical angles of the m-th sampling point with respect to the n-th loudspeaker. Following the convention of other literature, we consider the complex amplitude as the input to the loudspeaker, i.e., $s_n(f) = V_n(f) \cdot E(f)$ in [Pa·m] is the complex amplitude of the input to the loudspeaker in our formulation.

C.2 Link Equations for the immersive communication model in Chapter 5

This section explains the physical link equation between the talkers and microphones in room 1, and between the loudspeakers and listeners in room 2.

Room 1: Let $\widehat{\mathbf{x}_n} = (\widehat{x_n}, \widehat{y_n}, \widehat{z_n})$ be the location of the n-th talker in room 1 and $\widehat{\mathbf{y}_m} = (\widehat{x_m}, \widehat{y_m}, \widehat{z_m})$ be the location of the m-th microphone. Let $a_n(f)$ be the complex amplitude (the pressure caused by the n-th talker at 1 m) in [Pa·m], $\mathcal{T}_n(f, \theta_m, \phi_m)$ be the dimensionless radiation pattern of the talker, and $g_{m,n}(f; \|\widehat{\mathbf{x}_n} - \widehat{\mathbf{y}_m}\|_2)$ in [m^{-1}] be the free space Green's function. In this thesis, as in other SFR literature, we work with the pressure, and the pressure sensed by the m-th microphone due to n-th talker is:

$$p_{m,n}(f, \mathbf{x}_n, \mathbf{y}_m) = a_n(f) \cdot \mathcal{T}_n(f, \theta_m, \phi_m) \cdot g(f; \|\widehat{\mathbf{x}_n} - \widehat{\mathbf{y}_m}\|_2) \cdot \mathcal{M}_m(f, \theta_n, \phi_n), \tag{C.2}$$

where $\mathcal{M}_m(f, \theta_n, \phi_n)$ is the dimensionless receiving pattern of the microphone, (θ_n, ϕ_n) are the spherical angles of the n-th talker with respect to the m-th microphone, and $(\theta_m,\ \phi_m)$ are the spherical angles of the m-th microphone with respect to the m-th talker. The voltage produced by m-th microphone is:

$$\begin{aligned} V_{m,n}(f, \widehat{\mathbf{x}_n}, \widehat{\mathbf{y}_m}) &= p_{m,n}(f, \widehat{\mathbf{x}_n}, \widehat{\mathbf{y}_m}) \cdot W(f) \\ &= a_n(f) \cdot \mathcal{T}_n(f, \theta_m, \phi_m) \cdot g(f; \|\widehat{\mathbf{x}_n} - \widehat{\mathbf{y}_m}\|_2) \cdot \mathcal{M}_m(f, \theta_n, \phi_n) \cdot W(f), \end{aligned} \tag{C.3}$$

where $W(f)$ in [V/Pa] is the transfer function of the microphone that converts the pressure sensed by the microphone to the output voltage.

Room 2: Let $\widetilde{\mathbf{x}_n} = (\widetilde{x_n}, \widetilde{y_n}, \widetilde{z_n})$ be the location of the n-th loudspeaker in room 2 and $\widetilde{\mathbf{y}_m} = (\widetilde{x_m}, \widetilde{y_m}, \widetilde{z_m})$ be the location of the m-th sampling point around a listener's head. Let $s_n(f)$ be the complex amplitude of the n-th loudspeaker in [Pa.m], $\mathcal{L}_n(f, \theta_m, \phi_m)$ be the dimensionless radiation pattern of the loudspeaker ,and $g_{m,n}(f; \|\widetilde{\mathbf{x}_n} - \widetilde{\mathbf{y}_m}\|_2)$ be the free space Green's function. According to Eq.(C.1), the pressure at m-th sampling point, in absence of the Human Head, is equal to $s_n(f) \cdot \mathcal{L}_n(f, \theta_m, \phi_m) \cdot g(f; \|\widetilde{\mathbf{x}_n} - \widetilde{\mathbf{y}_m}\|_2)$. In presence of the Human head, the pressure at $\widetilde{\mathbf{y}_m}$ is formulated by:

$$p_{m,n}(f, \widetilde{\mathbf{x}_n}, \widetilde{\mathbf{y}_m}) = s_n(f) \cdot \mathcal{L}_n(f, \theta_m, \phi_m) \cdot g(f; \|\widetilde{\mathbf{x}_n} - \widetilde{\mathbf{y}_m}\|_2) \cdot \mathcal{H}_m(f, \theta_n, \phi_n), \tag{C.4}$$

where $\mathcal{H}_m(f, \theta_n, \phi_n)$ is the dimensionless receiving pattern of the human head (Head Related Transfer Function of the listener). In this equation (θ_n, ϕ_n) are the spherical angles of the loudspeaker with respect to the sampling point (human head), and (θ_m, ϕ_m) are the spherical angles of the sampling point with respect to the loudspeaker.