HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Electrical and Communications Engineering
Laboratory of Acoustics and Audio Signal Processing

Timo Haapsaari

Two-Way Acoustic Window using Wave Field Synthesis

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, May 28, 2007

Supervisor:	Professor Vesa Välimäki
Instructors:	D.Sc. Aki Härmä

HELSINKI UNIVERSITY OF TECHNOLOGY

ABSTRACT OF THE MASTER'S THESIS

Author:	Timo Haapsaari		
Name of the thesis:	Two-Way Acoustic Window using Wave Field Synthesis		
Date:	May 28, 2007	**Number of pages:**	75
Department:	Electrical and Communications Engineering		
Professorship:	S-89		
Supervisor:	Prof. Vesa Välimäki		
Instructors:	Aki Härmä, Dr.Sc.(Tech.)		

In this Master's Thesis a two-way multichannel audio communication system is introduced. The aim is to create a virtual acoustic window between two rooms, providing correct spatial localization of multiple audio sources on both sides.

Extending monophonic communication systems to feature multichannel sound capture and reproduction increases the intelligibility of speech and the accuracy of source localization achieved with the system. Adding multiple channels to the system also increases the complexity of the acoustic echo cancellation. Methods known from stereophonic systems extend to multichannel systems.

By using arrays of microphones and loudspeakers it becomes possible to try to recreate a part of the acoustic wave field existing in the recording space. A method for achieving this is wave field synthesis (WFS).

To solve the acoustic feedback problem, a 48 channel acoustic echo canceller was implemented. To maximize the achieved echo attenuation, a combination of adaptive and static filters were used. The implementation provided a stable solution that made normal conversation through the window possible.

To verify the quality of the system, a listening test was performed. In the test, WFS was compared against three other recording and reproduction methods on four different attributes of the perceived sound scape. The results show that WFS offers clear potential to be used in multichannel communication systems and in creation of the acoustic opening.

Keywords: acoustic signal processing, digital filters, sound reproduction, acoustic arrays

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Timo Haapsaari		
Työn nimi:	Two-Way Acoustic Window using Wave Field Synthesis		
Päivämäärä:	28.5.2007	**Sivuja:**	75
Osasto:	Sähkö- ja tietoliikennetekniikka		
Professuuri:	S-89		
Työn valvoja:	Prof. Vesa Välimäki		
Työn ohjaajat:	TkT Aki Härmä		

Tässä diplomityössä esitellään monikanavainen ja kaksisuuntainen audiokommunikaatiojärjestelmä. Sen tavoitteena on luoda kaksisuuntainen akustinen avanne kahden tilan välille ja mahdollistaa tarkka äänilähteiden paikantuminen molemmissa tiloissa.

Kun yksikanavainen kommunikaatiojärjestelmä laajennetaan monikanavaiseksi, on myös mahdollista parantaa puheen ymmärrettävyyttä. Toisaalta lisääntynyt kanavamäärä monimutkaistaa akustisen kierron poistamiseen käytettyjä tekniikoita. Tekniikat, jotka tunnetaan kaksikanavaisista järjestelmistä on mahdollista laajentaa myös monikanavaisiin järjestelmiin.

Käyttämällä kaiutin- ja mikrofonihiloja on osittain mahdollista äänittää äänikenttä toisaalla ja toistaa se samanlaisena toisessa tilassa. Tämä voidaan toteuttaa tässä työssä käytetyllä menetelmällä, jota kutsutaan äänikenttäsynteesiksi.

Akustisen kierron poistamiseksi toteutettiin 48-kanavainen järjestelmä, joka hyödynsi staattisten ja adaptiivisten suodinten yhdistelmää. Järjestelmä osoittautui stabiiliksi ja mahdollisti normaalin keskustelun rakennetun akustisen avanteen läpi.

Aaltokenttäsynteesiä verrattiin muihin äänentoisto- ja äänitysjärjestelmiin kuuntelukokeiden avulla. Tulokset osoittavat, että äänikenttäsynteesin ominaisuudet ovat riittävät korkealaatuisen ja monikanavaisen äänikommunikaatiojärjestelmän toteuttamiseksi.

Avainsanat: akustinen signaalinkäsittely, digitaaliset suotimet, äänentoisto, akustiset hilat

Acknowledgements

This Master's Thesis has been done for Philips Research Eindhoven. I would like to thank my instructor Aki Härmä and Werner de Bruijn for various helpful conversations and support during my research and writing process. Furthermore, I wish to thank my colleagues Okke Ouweltjes and Tobias May from Philips Research for their expertise and help in tasks related to signal processing and acoustic measurements.

My gratitude also goes to my fellow students Rubén Castillo Ochoa, Félicien Vallet, Stefan Gustafsson and Thomas Peeters and all the other people I met during my stay in Eindhoven for creating an enjoyable, open and humorous atmosphere. I will miss you all.

In addition, I wish to thank my parents for driving around the Europe for me. The journey would have been much more difficult without you.

Finally, I would like to thank you Tarja, for your endless love and patience. Soon, there will be three of us. I love you both.

Espoo, May 28, 2007

Timo Haapsaari

Contents

Abbreviations

DML	Distributed mode loudspeaker
DSP	Digital signal processor
FFT	Fast Fourier transform
GUI	Graphical user interface
ILD	Interaural level difference
ITD	Interaural time difference
LMS	Least mean squares method
NLMS	Normalized least mean squares method
PC	Personal computer
WFE	Wave field extrapolation
WFS	Wave field synthesis

Chapter 1

Introduction

Electronic communication is already a natural part of people's everyday lives. Telephony, mobile phones, instant messaging or computer based communication methods are typical technologies of communicating over distances. While it is natural to use these devices, the received communication experience is not necessarily natural, for example, in comparison to a real face-to-face conversation. A simple single-channel communication system inevitably can not convey the high amount of sound information present at the recording side, most importantly the locations of the sound sources. This becomes even more apparent in simultaneous communication with two or more people on the other side. Due to this fact it is also impossible to reproduce the reduced information with loudspeakers on the other side of the communication channel. Consequently, the separation of different sounds becomes harder.

An acoustic opening is a concept that tries to solve these problems and improve the spatial quality of the communication experience. It is a multichannel communication system with the aim to realize a virtual window, or in other words a virtual acoustic opening between the two remote communication rooms and create an impression of two adjacent rooms with a part of the wall between them acoustically removed. The idea is depicted in Figure 1.1.

By recording the sound field by multiple microphones and reproducing it with multiple loudspeakers on the other side we can in ideal conditions duplicate the sound information from the recording side. The idea was first introduced in 1934 [47] at Bell Laboratories while seeking a solution for recording orchestras in concert halls and reproducing the music afterwards to an audience somewhere else, while preserving the acoustics of the original concert hall. They captured sound sources at nine separate positions in an acoustically treated room by different configurations of two or three microphones placed on a horizontal line and played the signals through similarly located loudspeakers in another room in realtime. They used listening tests to verify if the sound sources could be heard from the

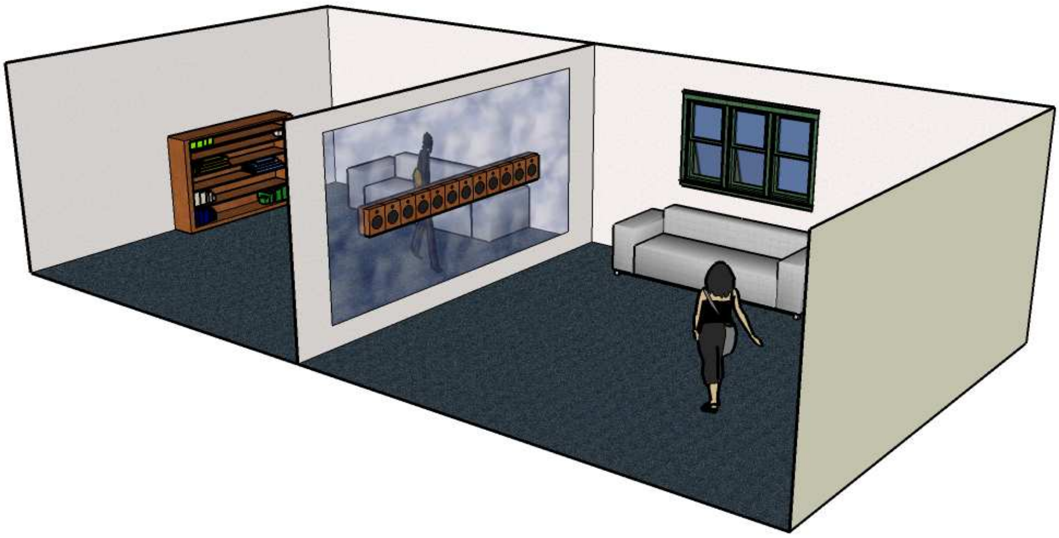


Figure 1.1: *The impression of two remote rooms becoming adjacent with an acoustic opening.*

correct locations. The results were promising and the best results were provided by connecting each microphone from the recording side directly to the corresponding loudspeaker on the reproduction side. With this method the perceived sound source location almost matches direct listening, in other words without the microphone-loudspeaker system between the listener and the sound source. The end result of the experiment was the first successful implementation of an acoustic opening between the recording and the listening room.

The setup used by Bell Laboratories was simple by today's standards. Huge advancements in loudspeaker, microphone and signal processing technologies have been made since their experiments, providing new possibilities for implementing the acoustic opening. Using tens or hundreds of small size microphones and loudspeakers offers a real improvement in accuracy of the recorded and reproduced sound field. Increasing the number of audio channels in the system also increases the total complexity of the system, but together with modern signal processing techniques gives an almost unlimited number of combinations and methods for recording and reproducing the sound. A system with tens of microphones and loudspeakers can be easily driven by a modern PC or a single DSP, making it feasible to use such a system also in a home environment.

The balance between the system's complexity and the achieved qualitative gains should be controlled. If increasing the number of loudspeakers does not produce advantages in the perceived audio quality, it is not obviously needed. The received audio quality within a large listening area has been investigated for setups with different numbers of loudspeakers in [30]. It was concluded that decreasing the spacing of the loudspeakers and therefore increasing the number of loudspeakers also increases significantly the size of the *sweet spot* i.e. the area where the perceived sound quality is good.

The increased number of audio channels additionally increases the amount of audio data to be transferred if the system is used as a communication device between two remote locations. Sending raw data from tens of microphones over the same connection that is normally used for only a couple of channels becomes impossible. Using the fact that all microphones are picking up almost the same signal with small differences, we can reduce the amount of data to be transferred. The topic has been investigated in [28] where a general framework and coding methods for the audio transfer were proposed. Instead of transferring all recorded channels it is possible to transfer a smaller number of channels and additional reconstruction filters to synthesize the loudspeaker signals at the receiving side. This enables the use of current communication channels with the acoustic opening.

The biggest problem with the acoustic opening arises when it is being used as a two-way communication system. On both sides there are multiple microphones recording the sound field in the room. This recorded information is then played back through the loudspeakers on the other side making communication possible, but also the sound from the loudspeakers and the reflections from the room are picked up by the microphones on that side. The system then renders the recorded audio back to the side that it originated from creating a closed acoustic loop between the two rooms. This produces an audible echo, and in the worst case the level of the echo increases on each loop, causing a loud and howling sound that destroys the ability to use the system for communication. This is discussed in detail for example in [46] or in [12].

The aim of this thesis is to review the properties of current monophonic communications systems and extend the discussion to stereophonic and finally to multichannel communication systems. The advantages of multichannel communication are discussed with the problems with the acoustic echo cancellation and the increasing complexity of the system. An additional aim of this thesis is to investigate a method for multichannel sound reproduction called Wave Field Synthesis (WFS, first introduced in [8]); to derive the requirements for hardware and signal processing and to build a two way acoustic opening according to these requirements. A multichannel acoustic echo canceller combining static filters and simple adaptation process is then implemented and its effectiveness is examined. In addition, multiple sound recording and reproduction techniques are compared by listening tests on four different attributes of the perceived sound scape. After the results a conclusion is derived for selecting the best methods for sound capture and reproduction in an acoustic opening and possibilities for future work and research are discussed.

Chapter 2

Multichannel communications

2.1 Introduction

Most current communication devices still use monophonic signal transmission. The sound is picked up by a single microphone and rendered at the other end of the communication channel with one loudspeaker. A common variation is using two loudspeakers playing the same monophonic signal. In this section components of a single-channel communication system are introduced and the system's disadvantages are reviewed. The system is then extended into stereophonic and the advantages and the difficulties encountered are discussed with possible solutions. Furthermore the communication system is extended to feature multichannel sound capture and reproduction and the problems with system complexity are investigated and possible solutions are provided. In addition, with each setup an acoustic echo canceller is needed to remove the feedback from the loudspeakers to the microphones. Effective implementations for the echo cancellers are discussed with each case.

2.2 Monophonic communications

An example of a monophonic communication system is shown in Figure 2.1. The user is using a hands-free telephone with one speaker and one microphone. The user's speech is picked up by the microphone but in addition the signal output from the loudspeaker is picked up. The acoustic path between the loudspeaker and the microphone can be represented with an impulse response h. The sound from the loudspeaker can be removed from the microphone signal if h is known. The loudspeaker signal is filtered using an estimate $\hat{h}$ of the acoustic path and the end result is subtracted from the microphone signal. Depending on the accuracy of the estimate, some amount of the echo still remains in the microphone signal. In conjunction with selective attenuation of the microphone signal and speech de-

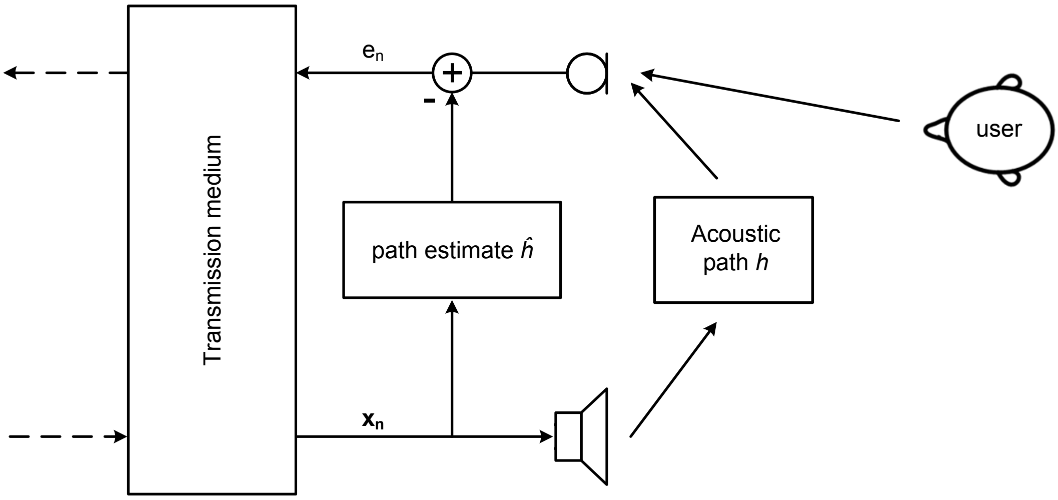


Figure 2.1: *One side of a monophonic communication system. Loudspeaker and microphone are directly coupled with an acoustic path h between them. An acoustic echo canceller estimates the acoustic path ($\hat{h}$) and removes the output of the loudspeaker from the microphone signal.*

tection this method is used in most current communication systems [29].

Acquiring the estimate $\hat{h}$ for the acoustic path can be performed in several ways. The simplest way is to measure the path off-line and use the (truncated) result for the echo canceller. This is not a very robust method due to the fact that the acoustic path is affected constantly by several aspects, for example the location of the user.

More effective solutions can be implemented by adapting the acoustic path estimate to take the changes into account by minimizing the residual echo that is left in the signal picked up by the microphone. Most commonly used algorithms are the least mean square (LMS) and normalized least mean square (NLMS) methods [24]. By adapting each coefficient according to (2.1) at each time moment n the estimate $\hat{h}$ converges towards the correct solution h.

$$\boldsymbol{\hat{h}_n} = \boldsymbol{\hat{h}_{n-1}} + 2\mu \boldsymbol{x_n} e_n \tag{2.1}$$

In the equation μ is called adaptation step size, $\boldsymbol{x_n}$ is the excitation signal and e_n is the residual error signal left in the microphone signal. To guarantee the convergence of the adaptation, the step size has to be limited according to (2.2), where L is the length of $\boldsymbol{\hat{h}}$ and σ_x^2 is the variance of $\boldsymbol{x_n}$.

$$0 < \mu < \frac{1}{L\sigma_x^2} \tag{2.2}$$

Convergence of the LMS algorithm is slow, mainly due to the use of a constant step size. It can be shown that it converges slower with highly colored excitation signals like speech. In addition in can be shown that to speed up the convergence, μ has to be selected to be in the middle of the range indicated by (2.2). Effectively this means that the step size has to be a function of the variance of the excitation signal. This leads to the normalized LMS where the coefficient update rule is given by

$$\hat{\boldsymbol{h}}_{\boldsymbol{n}} = \hat{\boldsymbol{h}}_{\boldsymbol{n-1}} + \frac{\alpha}{L\sigma_x^2}\boldsymbol{x}_{\boldsymbol{n}} e_n. \tag{2.3}$$

For convergence α has to be selected such that

$$0 < \alpha < 2. \tag{2.4}$$

Furthermore for non-stationary signals the variance σ_x^2 that is used has to be time-varying. A good way of implementing this is to use the dot product of the excitation frame [24] given by

$$\sigma_x^2 = \boldsymbol{x}_{\boldsymbol{n}}^{\boldsymbol{t}} \boldsymbol{x}_{\boldsymbol{n}}. \tag{2.5}$$

2.3 Stereophonic communications

An advantage of monophonic communication systems is their simplicity and the existing robust and effective solutions for acoustic echo cancellation. The biggest drawback of monophonic sound capture and reproduction is the inability of the system to convey spatial information. Mainly the problem arises when there are several participants present at the same time during the communication event. Using only one audio channel makes it hard to identify which of the participants is talking. It is a known fact that adding spatial information to the communication makes this easier [13, 14] and increases speech intelligibility. A rather good lateral localization can be achieved by using stereophonic recording and playback methods [47, 1, 9]. The perceived audio sources are usually located somewhere between the two loudspeakers. This already improves the identification of the communication participants. Stereophonic sound reproduction has been widely discussed in [17].

Adding the second loudspeaker and microphone to the system complicates the echo cancellation process and especially degrades the performance of the adaptive filtering methods described earlier in Section 2.2. An example of a stereophonic communication system is shown in Figure 2.2. The far-end microphone signals can be considered as filtered versions of the speech of the user. Effectively this means that both microphone signals have a common origin which is then filtered with the acoustic path (g_1 and g_2) from the user to each microphone. When comparing Figures 2.1 and 2.2, it can be seen that not only is the echo canceller more complex, it also has to cope with two highly correlated loudspeaker signals. For the adaptive cancellers this proves to be a challenging task. Because of the high correlation between the signals, it is not possible for the echo canceller to identify anymore which

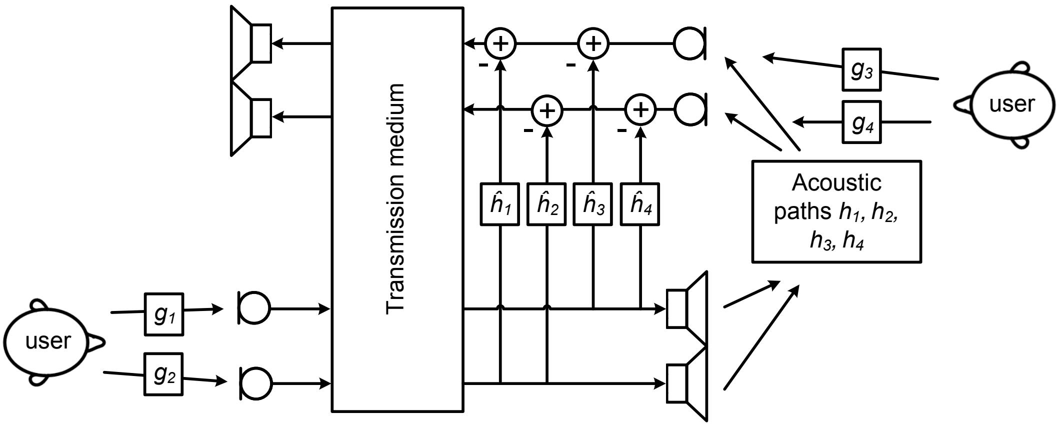


Figure 2.2: *Stereophonic communication system. Loudspeakers and microphones are directly coupled with acoustic paths h_n between them. An acoustic echo canceller estimates the acoustic paths ($\hat{h}_n$) and removes the output of the loudspeakers from the microphone signals. The microphone signals are a combination of residual echo and near-end communication source signal filtered with the acoustic paths (g_n) from the user to the microphones*

signal is coming from the loudspeaker it should be cancelling. The problem is well-known [48], and is usually called the non-uniqueness problem. The derivation leading to (2.1) does not have a unique solution anymore and the convergence of the adaptation slows down or in the worst case does not converge at all.

Multiple solutions have been suggested to solve the problem. One solution is to decorrelate the microphone signals by adding random noise to them. The noise should be low enough in level not to be heard but still high enough to decorrelate the signals. It has been shown that the noise should be 13-15 dB lower in level than the speech signal, and can be hidden with spectral shaping techniques [48]. The most successful methods use non-linear distortion added to both microphone signals. A method adding a nonlinear function of the signal to the signal itself is described in [5]. By modifying both channels according to (2.6)

$$\hat{\boldsymbol{x}}_{\boldsymbol{n}} = \boldsymbol{x}_{\boldsymbol{n}} + \alpha f(\boldsymbol{x}_{\boldsymbol{n}}), \tag{2.6}$$

where

$$f(x) = \begin{cases} x & \text{, if } x \geq 0 \\ 0 & \text{, otherwise,} \end{cases} \tag{2.7}$$

the signals are decorrelated and for values up to $\alpha = 0.5$ only slight degradation of the original signal is introduced. In addition the stereophonic spatial localization is not affected [5].

2.4 Multichannel communications

Extending the stereophonic communication system to feature more than two microphones and loudspeakers enables increasing the intelligibility of the speech and the accuracy of the localization of the sound sources [1, 47]. Additional accuracy can also be achieved in the perceived depth of the audio sources. By increasing the number of loudspeaker in the system, we have the possibility to increase the ratio between the direct and the reverberant sound and especially in highly reverberant rooms this increases speech intelligibility [39].

As the case is with moving from the monophonic to the stereophonic communication, stepping up from the stereophonic communication complicates the echo cancellation process. An example of a multichannel communication system is shown in Figure 2.3. The system features N microphone inputs and M loudspeaker outputs and produces therefore $N * M$ acoustic echo paths from the loudspeakers to the microphones. The multichannel echo canceller has to cancel all these paths from all the N microphone signals. The non-uniqueness problem described in the previous Section 2.3 is emphasized with the additional channels in the system. Solutions provided for stereophonic systems can be generalized for multichannel systems.

In communication systems the amount of the loudspeakers could range from 10 to 100 and the amount of the microphones could be even higher. This could produce tens of thousands acoustic paths between the system's components and high computational load for the echo canceller. For this reason, more robust on computationally effective algorithms are needed for high quality acoustic echo cancellation.

It was mentioned in Section 2.3 that all the microphone signals are highly correlated due to the fact that they are derived from the same audio sources. The acoustic echo canceller has to handle these signals, and for adaptive methods this proves to be difficult. One solution to the problem is creating algorithms that take the cross-correlations between the loudspeaker signals into account. The recursive least squares (RLS) algorithm [3] is known to have a good convergence speed. The update equations for the adaptive RLS algorithm are the following [12]:

$$\boldsymbol{\hat{H}_n} = \boldsymbol{\hat{H}_{n-1}} + \boldsymbol{k_n e_n^T} \tag{2.8}$$

$$\boldsymbol{k_n} = \boldsymbol{R_{xx}^{-1} X_n}. \tag{2.9}$$

In the previous equations matrix $\boldsymbol{\hat{H}_n}$ contains all the adapted impulse responses between the loudspeakers and the microphones at a time moment n. The error vector $\boldsymbol{e_n^T}$ contains the residual echo signal of each microphone, $\boldsymbol{R_{xx}^{-1}}$ contains both the auto-correlations and the cross-correlations between the loudspeakers signals, and $\boldsymbol{X_n}$ contains all the loudspeaker signals. The computationally most demanding part of the adaptation is the calculation of

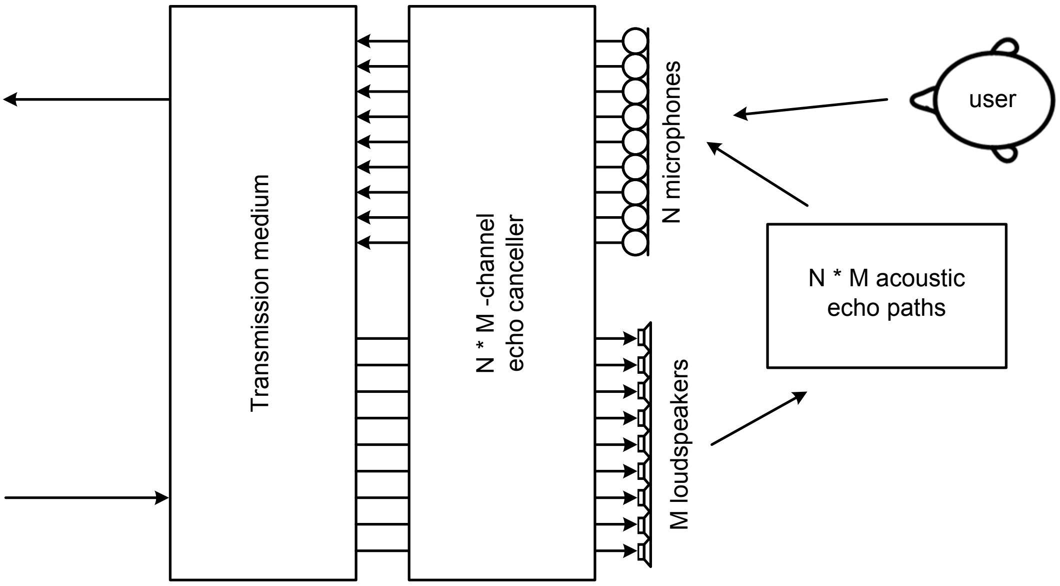


Figure 2.3: *One side of a multichannel communication system with N microphones and M loudspeakers. For echo cancellation an $N * M$ channel canceller is used.*

the *Kalman gain* $\boldsymbol{k_n}$ due to the calculation of all the cross-correlations and inverting the correlation matrices. An efficient solution for the calculation has been introduced in [11]. The method uses the fast Fourier transform (FFT) to calculate the matrix operations in blocks in the frequency domain. As a result of the block processing, the arithmetic complexity of the calculation is significantly reduced. It has been shown that the method clearly outperforms the NLMS method described in Section 2.2 [12].

2.5 Conclusion

Monophonic communication systems have been used for tens of years. For multiuser systems the intelligibility of speech decreases, mainly due to the lack of spatial separation of the sound sources. Monophonic communication cannot provide the spatial information existing on the recording side of the communication. For acoustic echo cancellation the monophonic system is an easy task. Using adaptive filters, the feedback from the single loudspeaker can be easily removed from the microphone signal.

Extending the communication system with an additional microphone and loudspeaker already increases the spatial localization of the audio sources significantly, and therefore increases the intelligibility of the speech. On the other hand, additional channels cause problems with the adaptive echo cancellation filters. The fact that both microphone signals are derived from the same audio sources causes a non-uniqueness problem. The echo canceller tries to cancel the sounds of two highly correlated loudspeaker signals from the microphone signals, but is unable to separate which part of the signal comes from which loudspeaker. This prevents the adaptation from converging. A solution for the problem is decorrelating the two loudspeaker signals with non-linear transformations.

Further extension of the system to feature multichannel recording and reproduction extends the advantages of stereophonic communication systems. In addition, the problems with the acoustic echo canceller increase. Multichannel acoustic echo cancelling algorithms exist and the most effective ones take into account the cross-correlation of the multiple channels and feature frequency domain adaptive filtering for computational efficiency.

In any case, the multichannel acoustic echo cancellers require a lot of computational power. With possibly thousands of echo paths to be cancelled, more efficient algorithms are needed. With the continuously increasing processing power provided by single computers and DSPs, more computationally demanding solutions are becoming feasible. This enables application of multichannel communications even at home environments.

Chapter 3

Spatial hearing

3.1 Introduction

To fully understand the benefits of multichannel communications an understanding of human hearing and its spatial characteristics is essential. Spatial hearing is a complex system that provides us with cues of the direction and the distance of surrounding sound events. This chapter gives an overview of the spatial aspects of human hearing and explains how spatial sound events are perceived.

3.2 Aspects of spatial hearing

The ability to perceive spatial sound events is partly given at birth and partly learnt. Even the smallest child has the ability to orientate towards a loud sound. On the other hand, many aspects of the spatial hearing are learnt through adaptation. This is easily proven, because the human hearing is highly dependant on the shape and size of the head and ears [36]. Human hearing learns to interpret the complex combination of direct and reflected sound waves and forms a spatial auditory image from this information. The human hearing is called *binaural* due to the use of two ears. The direction of sound events relative to the head can be described with a coordinate system described in Figure 3.1, where d is the distance to the sound event, γ is the elevation angle and ξ is the azimuth angle of the sound event.

Median plane divides the head vertically into two identical parts. Frontal plane goes through ears vertically and the horizontal plane, as the name suggests, horizontally. With this mapping, a sound event from front arrives from angles $\gamma = 0°$ and $\xi = 0°$. A sound event from behind arrives from angles $\gamma = 0°$ and $\xi = 180°$.

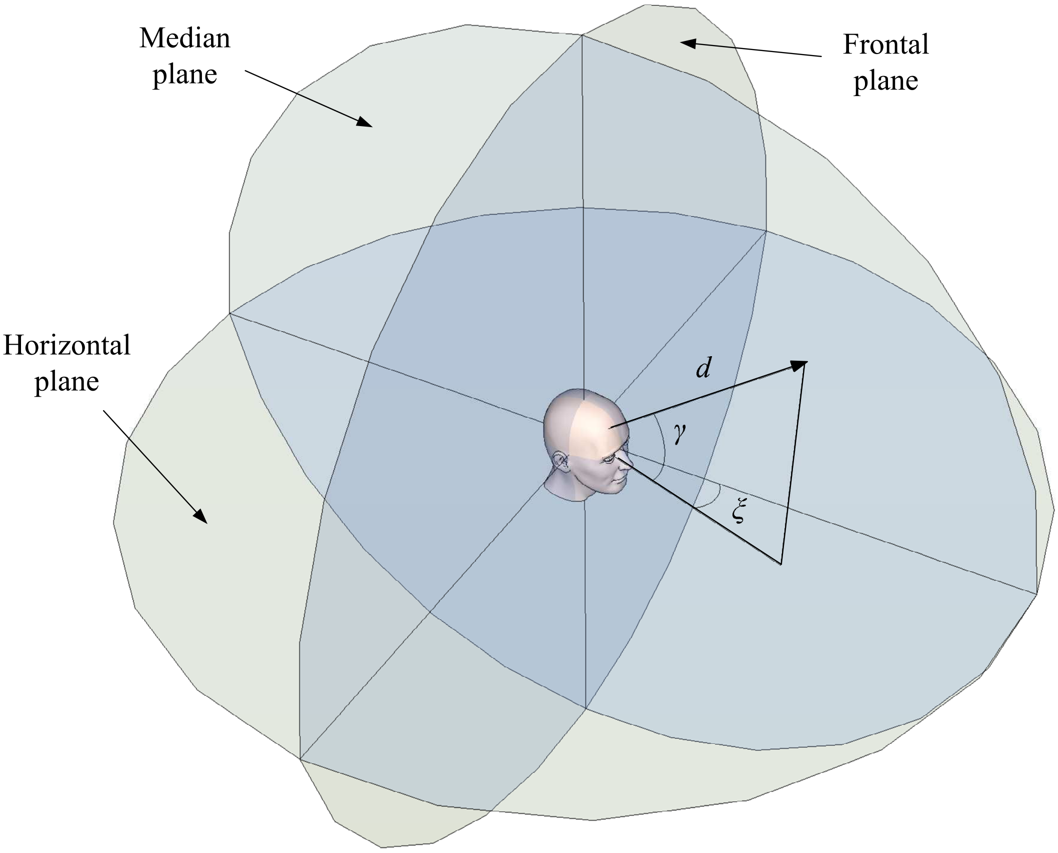


Figure 3.1: *The coordinate system for spatial hearing.*

3.2.1 Interaural time and level difference

Most of the spatial information is received from interaural time difference (ITD) and interaural level difference (ILD) between the signals in the two ears. The ITD can be evaluated using the coordinate system described in Figure 3.1 and approximating the head as a sphere with a radius of D. Now the ITD τ is given by

$$\tau = \frac{D}{2c}(\xi + \sin\xi)\cos\gamma, \tag{3.1}$$

where c is the speed of sound [33]. If the sound event is limited to the level of the ears (i.e. the horizontal plane), the elevation factor $\cos\gamma$ can be omitted from (3.1). The maximum value for the ITD is received when the sound event is arriving directly from the side on the horizontal plane ($\gamma = 0°$, $\xi = \pm 90°$). With an approximation of 18 cm for the diameter of the head the ITD reaches its maximum value of $700\mu s$ [36].

The ITD cues dominate sound localization at low frequencies. For frequencies below 1600 Hz interaural time differences are the major cues for sound source localization (see for example [44]). The role of ITD starts to decrease with frequencies above 2000 Hz [33]. This is mainly due to shorter wave length of the arriving sound creating a weaker detection of phase change between the ears. At higher frequencies ILD starts to dominate. Problems with the sound localization start to occur when ITD and ILD are small or close to zero, mostly at the median plane. In addition, problems arise when the sound event is located at the *cone of confusion*, e.g. when the ITD and ILD cues are ambiguous.

3.2.2 Other spatial cues

Besides the ITD and ILD cues, spatial hearing uses other cues for sound localization. Intersensory cues (i.e. sight) have a large impact on sound localization, especially on distance perception of the sound sources [33]. Furthermore, the cues received from the acoustics of the listening room are known to contribute to the localization of sound sources, especially in a small room and with sources at the median plane. With the acoustic window between two listening rooms we effectively create a combination of two acoustic spaces which both contribute to the localization of the sound sources in the rooms. In addition, moving the head and the ears, spatial perception can be made more accurate, especially at the cone of confusion and median plane.

3.3 Conclusion

Understanding the full advantages of multichannel audio, an understanding of human spatial hearing is needed. Spatial hearing is a complex system which gives us cues of the location of heard sound events. Interaural time and level differences are mainly used in localizing sound events. Problems with spatial hearing arise when the sound is arriving from the median plane or from the cone of confusion. In these situations other cues, i.e. sight, are used for more accurate sound localization.

Chapter 4

Wave field synthesis

4.1 Introduction

Creation of the acoustical window requires a method for sound recording and reproduction as well as hardware to realize the method. This chapter gives a presentation of a method for spatially correct sound reproduction using the concept of wave field synthesis (WFS) first introduced in [8]. Starting from basic principles it will be shown that it is possible to re-create a spatially and temporally correct representation of the original wave field recorded elsewhere. The method of wave field synthesis is then compared to more traditional methods like stereo reproduction of sound. It will be shown, that WFS offers superior performance compared to these traditional methods in re-creating the sound field and is well applicable for creation of the acoustical window. In addition problems of the method are discussed and different solutions are considered. From the results a conclusion is derived for choosing the hardware and selecting signal processing algorithms for building the acoustic opening.

4.2 Theory of WFS

4.2.1 The Huygens' principle

The Huygens' principle - originating from 1690 [34] - is the basis of wave field synthesis. According to the principle, each point of a wave front can be considered as a new center of a new spherical wave front. The envelope of the original wave front can then be considered as a combination of all these elementary wave fronts. The principle is depicted in Figure 4.1. A point source Ω is emitting an impulse at time $t = 0$ in a homogenous medium. $F(t)$ is a spherical wave front caused by Ω at time t, with a radius of ct, c being the speed of sound. Each point Ω' on $F(t)$ can now be considered as a source for a new spherical wave front.

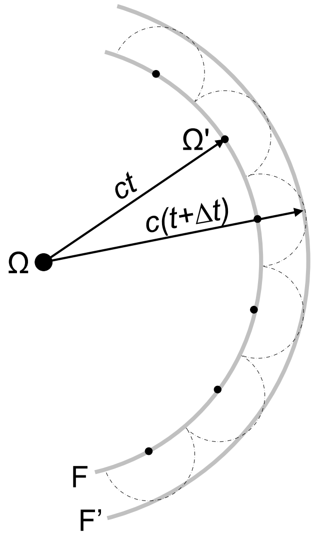


Figure 4.1: *The Huygens' principle.*

As a result the envelope of the wave front $F'(t + \Delta t)$ is being formed by a combination of the new wavefronts with radius of $c\Delta t$. If we can record the sound at points Ω' then it seems intuitive that we can reproduce the sound field outside $F(t)$ by placing sound sources (e.g. loudspeakers) on $F(t)$, while omitting the initial source Ω.

Although intuitive, the Huygens principle does not describe the actual physics of wave field propagation correctly, so we need further mathematical proof for using the principle in practise.

4.2.2 The Kirchhoff-Helmholz integral

The mathematical starting point of WFS is Green's second theorem given by

$$\int_V (f\nabla^2 g - g\nabla^2 f)dV = -\int_S (f\nabla g - g\nabla f)\cdot\boldsymbol{n}dS. \tag{4.1}$$

In the most concrete interpretation, the equation means that a field inside a closed volume can be described if the field is known on the surface enclosing the volume. The layout for the derivation and definition of the variables used are depicted in Figure 4.2, where S is some closed surface containing a source free volume V and $\boldsymbol{n}$ is the normal vector of the surface S pointing inward. The scalar functions f and g are both twice continuously differentiable on S. In case of a sound field we can choose pressure field P of some source distribution Ω outside S satisfying the homogenous wave equation (4.2) as f. Function g is selected so that it satisfies the inhomogeneous wave equation (4.3) inside the surface S. Selecting the pressure field $G(\boldsymbol{r}|\boldsymbol{r_R}, k)$ of a monopole point source at location R as g fills

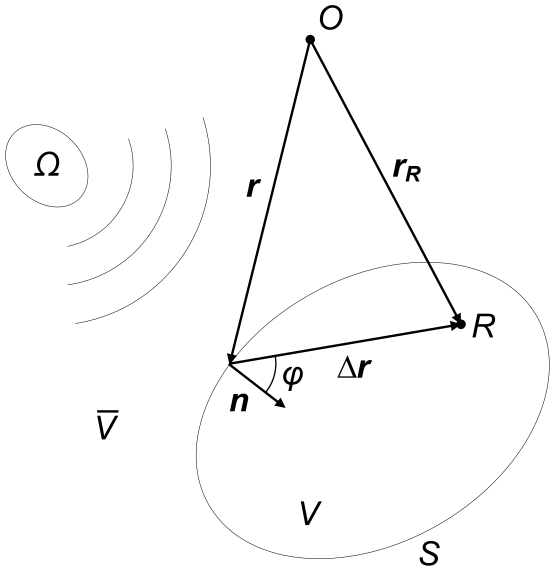


Figure 4.2: *Layout for derivation of the Kirchhoff-Helmholz integral .*

this requirement. The homogenous wave equation is given by

$$\nabla^2 P(\boldsymbol{r}, k) + k^2 P(\boldsymbol{r}, k) = 0, \tag{4.2}$$

where and $k = \omega/c$ is the wavenumber, where ω is the angular frequency . In addition, the inhomogeneous wave equation is qiven by

$$\nabla^2 G(\boldsymbol{r}|\boldsymbol{r_R}, k) + k^2 G(\boldsymbol{r}|\boldsymbol{r_R}, k) = -4\pi\delta(\boldsymbol{r} - \boldsymbol{r_R}), \tag{4.3}$$

where δ is Dirac's delta function [53].

Function $G(\boldsymbol{r}|\boldsymbol{r_R}, k)$ satisfying (4.3) is called *Green's function*, which in signal processing is also known as the impulse response. For a monopole point source we can write [54] an acoustic transfer function

$$G(\boldsymbol{r}|\boldsymbol{r_R}, k) = \frac{e^{-jk|\Delta \boldsymbol{r}|}}{|\Delta \boldsymbol{r}|}, \tag{4.4}$$

where $\Delta\boldsymbol{r} = \boldsymbol{r} - \boldsymbol{r_R}$. The Green's function remains the same while switching places of the source and the receiver producing

$$G(\boldsymbol{r}|\boldsymbol{r_R}, k) = G(\boldsymbol{r_R}|\boldsymbol{r}, k). \tag{4.5}$$

This is known as the reciprocity of the system and for acoustic systems it was introduced in [43] and is also discussed in [57]. Now, using (4.5) and inserting G into (4.1) as g and P as

f we can derive the pressure at point R inside S if the wave field of an external source distribution Ω is known on surface S [54]. Substituting yields *the Kirchhoff-Helmholz integral* :

$$P(\boldsymbol{r_R}, k) = -\frac{1}{4\pi}\int_S [G(\boldsymbol{r_R}|\boldsymbol{r}, k)\nabla P(\boldsymbol{r}, k) - P(\boldsymbol{r}, k)\nabla G(\boldsymbol{r_R}|\boldsymbol{r}, k)] \cdot \boldsymbol{n} dS. \tag{4.6}$$

The relationship between pressure P and the inward pointing normal component V_n of the particle velocity on surface S can be described with the equation of motion (4.7) [54]:

$$\frac{\partial P}{\partial n} = -jck\rho_0 V_n, \tag{4.7}$$

where j is the imaginary unit and ρ_0 is the static density of volume V (Fig. 4.2). Inserting (4.4) and (4.7) into (4.6) produces

$$P(\boldsymbol{r_R}, k) = \frac{1}{4\pi}\int_S [jck\rho_0 V_n(\boldsymbol{r}, k)\frac{e^{-jk\Delta r}}{\Delta r} + P(\boldsymbol{r}, k)\frac{1 + jk\Delta r}{\Delta r}\cos\varphi\frac{e^{-jk\Delta r}}{\Delta r}]dS \tag{4.8}$$

for the pressure P inside S caused by sources on S. Equation (4.8) forms the basis of wave field synthesis. By taking a closer look at the integral we notice that [18] the first term of the integrand represents the pressure at $\boldsymbol{r_R}$ produced by a *monopole* source at position $\boldsymbol{r}$ on surface S (4.2). The source strength is proportional to the normal component V_n of the particle velocity at position $\boldsymbol{r}$. On the other hand the second term of the integrand represents the pressure at the same position $\boldsymbol{r_R}$ that is produced by a *dipole* source at location $\boldsymbol{r}$ on S with strength proportional to the pressure at $\boldsymbol{r}$. With other words, the pressure inside sourceless volume V can be calculated if normal component V_n of the particle velocity and the pressure $P(\boldsymbol{r}, k)$ of some source distribution Ω outside V are known on surface S. It should also be noted that the integral of pressure P everywhere outside surface S is zero [54].

Intuitively equation (4.8) means that the wave field inside some volume V enclosed by S and generated by some source distribution Ω outside V can be synthesized using a continuous distribution of monopole and dipole sources on surface S while omitting Ω. These *secondary* sources are representing the effect of the *primary* source distribution Ω on surface S as (4.8) points out and as depicted in Figure 4.3). In practise this means that we can create virtual sources around a listening space enclosed by some surface S using measured or calculated data to represent source distribution Ω on S.

While the Kirchhoff-Helmholz integral gives us mathematical tools to synthesize a wave field inside some space, it does not provide us with practical solutions for wave field creation. This is due to the requirement of a continuous distribution of secondary sources. Creating such a source distribution is not possible nor practical, as it would call for filling all the boundaries of the listening space with transducers. This creates a need for simplification of the system.

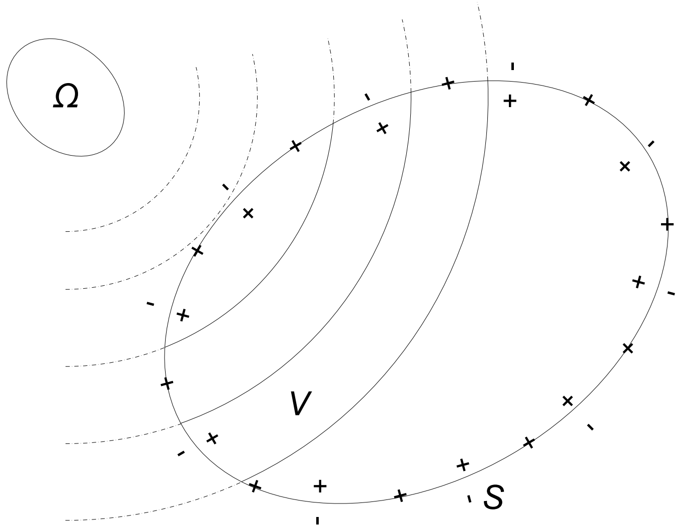


Figure 4.3: *Representation of (4.8). The wave field inside volume V and produced by source distribution Ω can be synthesized with a continuous distribution of monopole and dipole sources on surface S enclosing V. The pressure caused by the source distribution on S is zero outside V.*

4.2.3 The 3D Rayleigh integrals

In simplifying the Kirchhoff-Helmholz integral we keep in mind that the choice of $G(\boldsymbol{r}|\boldsymbol{r_R},k)$ in (4.4) is not unique. Any function satisfying (4.3) inside V (Fig. 4.2) can be used. This does not introduce any restrictions for boundary conditions on surface S and therefore we can use any convenient shape and boundary conditions for surface S. Choosing S so that it consist of a plane surface S_0 and a spherical surface S_1 (Fig. 4.4) produces interesting results. Now, by letting the radius of S_1, $r_1 \longrightarrow \infty$ we are creating an infinite plane S_0 at $z = 0$ between subspaces $z < 0$ and $z > 0$. Now the effect of S_1 vanishes and effectively we have to consider only the plane S_0 [54].

A logical step to simplify the Kirchhoff-Helmholz integral is to choose $G(\boldsymbol{r}|\boldsymbol{r_R},k)$ in (4.6) so that the second term of the integrand vanishes, in other words

$$\nabla G(\boldsymbol{r}|\boldsymbol{r_R},k) \cdot n = 0. \tag{4.9}$$

If this is satisfied, we only need monopole secondary sources on S_0 to synthesize the wave field in $z > 0$ (Section 4.2.2). This is satisfied by choosing $G(\boldsymbol{r}|\boldsymbol{r_R},k)$ as the sum of the

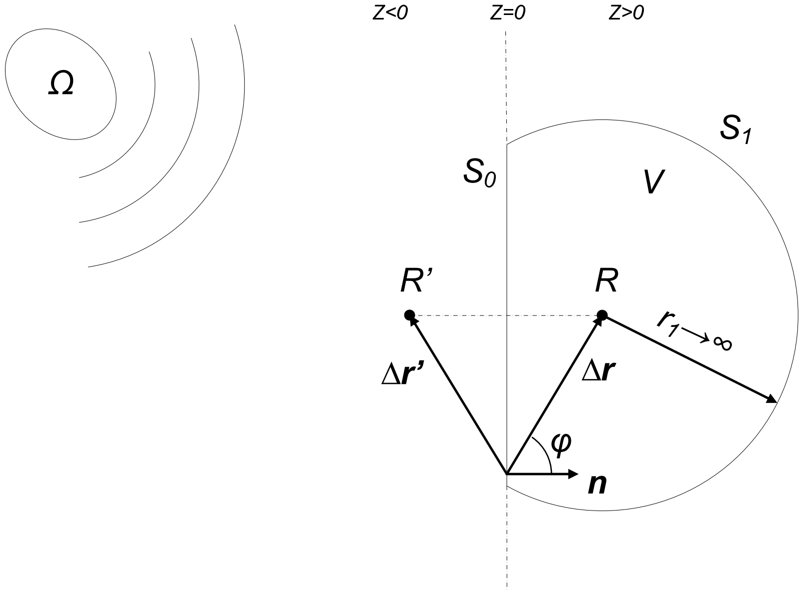


Figure 4.4: *Layout for derivation of the Rayleigh 3D integrals.*

fields of two identical point sources at positions R and R' (Fig. 4.4). This yields

$$G(\boldsymbol{r}|(\boldsymbol{r_R}, \boldsymbol{r_{R'}}), k) = \frac{e^{-jk|\Delta \boldsymbol{r}|}}{|\Delta \boldsymbol{r}|} + \frac{e^{-jk|\Delta \boldsymbol{r'}|}}{|\Delta \boldsymbol{r'}|} \tag{4.10}$$

and creates a fully reflective boundary condition on S_0. R' can be seen as a mirror image of R relative to plane S_0. Therefore the Kirchhoff-Helmholz integral (4.6) instead of producing zero pressure outside the volume V as the case was before is now producing a mirror image of the sound field in subspace $z > 0$ into subspace $z < 0$. This is due to the missing dipole sources that would cancel the mirror image. If we are not interested in subspace $z < 0$ and we can assume that there are no reflections from subspace $z < 0$ to $z > 0$, this does not matter. Using the reciprocity of Green's function (4.5) and substituting (4.10) and (4.7) into (4.6) yields the *3D Rayleigh I integral*:

$$P(\boldsymbol{r_R}, k) = \frac{jck\rho_0}{2\pi} \int_{S_0} V_n(\boldsymbol{r}, k) \frac{e^{-jk\Delta r}}{\Delta r} dS. \tag{4.11}$$

In other words we can synthesize the wave field in subspace $z > 0$ by placing a continuous distribution of monopole secondary sources on plane $z = 0$. The strength of each source is proportional to the normal component of the particle velocity caused by source distribution Ω at the location of each monopole. This is shown in Figure 4.5. The forming of the mirror image R' is now logical. A monopole source radiates evenly into every direction so subspace $z < 0$ is no different from subspace $z > 0$ relative to the sources on plane $z = 0$.

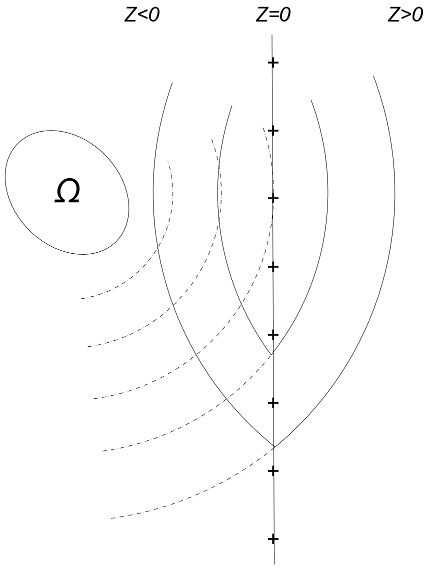


Figure 4.5: *Representation of (4.11). The wave field in subspace $z > 0$ that is produced by source distribution Ω can be synthesized with continuous distribution of monopole sources on infinite plane S_0. The setup also creates a mirrored wave field into $z < 0$.*

Similarly, we can cancel the second term from the integrand in (4.6) so that we would only need dipole sources. The result is known as the *3D Rayleigh II integral* (for derivation, see for example [7]):

$$P(\boldsymbol{r_R}, k) = \frac{1}{2\pi} \int_{S_0} P(\boldsymbol{r}, k) \frac{1 + jk\Delta r}{\Delta r} \cos\varphi \frac{e^{-jk\Delta r}}{\Delta r} dS. \tag{4.12}$$

We have now simplified the Kirchhoff-Helmholz integral into the 3D Rayleigh integrals. For synthesizing a wave field in a half space we only need an infinite plane of either monopole or dipole sources. This is already realizable with loosening the requirement of an infinite plane by approximating it by a finite plane. This seems far more cost efficient than the method proposed in Section 4.2.2. In addition we are able to use just one type of sources, which reduces the complexity of the system. Still even further simplification of the system would be convenient.

4.2.4 The $2^1/_2$D Rayleigh I integral

Instead of using a plane of secondary sources it would be more convenient to use just a linear array of sources. With some restrictions this is indeed possible. We will start by

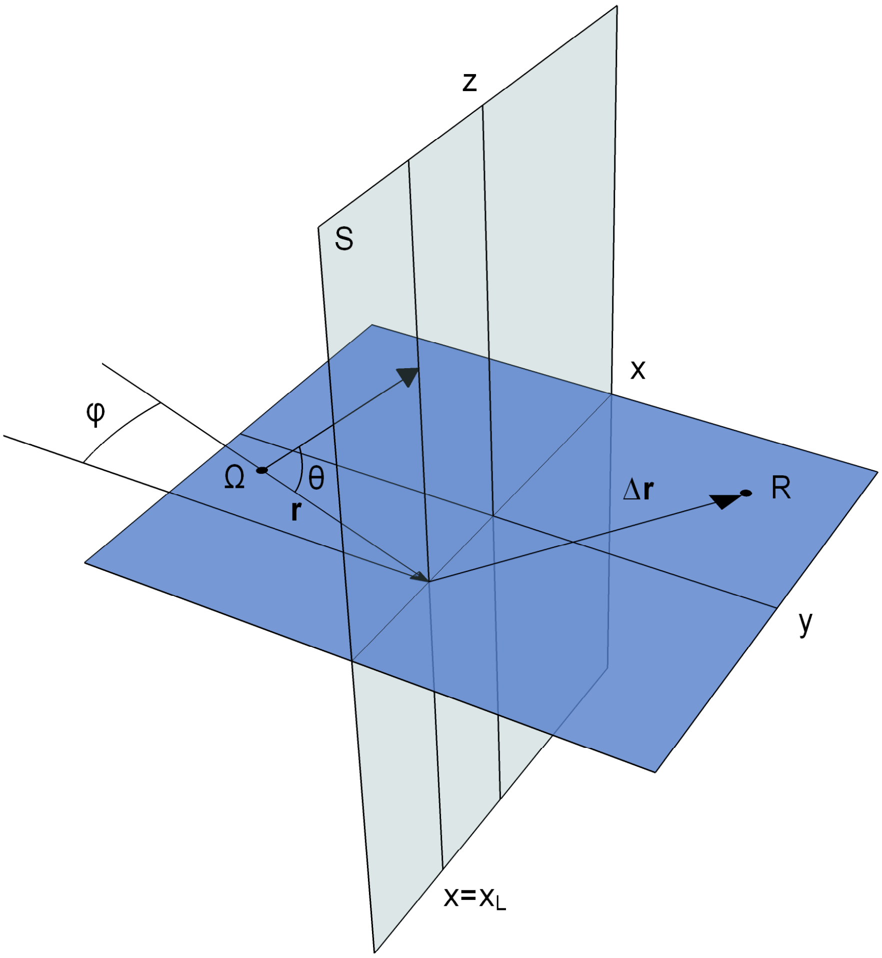


Figure 4.6: *Layout for derivation of the Rayleigh* $2^1/_2D$ *integral and definition of the variables. (Redrawn after [18]).*

transforming the plane integral (4.11) to a line integral [18] for a primary source far away from the array $(kr >> 1)$ (see Fig. 4.6 for geometry):

$$P(\boldsymbol{r_R}, k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} P_l(\boldsymbol{r_R}, x, k) dx. \tag{4.13}$$

Consider a setup where source Ω with directivity function $G(\varphi, \theta, k)$ and receiver R are on the same plane perpendicular to the plane S of secondary monopole sources. The plane S located at $y = 0$ is divided into vertical lines of secondary sources. Each vertical line has a contribution $P_l(\boldsymbol{r_R}, x, k)$ to the total pressure received at position R due to Ω. Now the integration is done over the location of the vertical line (variable x). Using a method called 'stationary phase approximation' it is possible to show (see [54] for details) that the total pressure contribution of a line located at $x = x_L$ received at R can be approximated by a single the secondary source at position $(x_L, 0, 0)$. This leads to the following approximation for $P_l(\boldsymbol{r_R}, x, k)$:

$$P_l(\boldsymbol{r_R}, x_L, k) = S(k)\sqrt{2\pi jk}\sqrt{\frac{\Delta r}{r + \Delta r}}G(\varphi, 0, k)cos\varphi\frac{e^{-jkr}}{\sqrt{r}}\frac{e^{-jk\Delta r}}{\Delta r}. \tag{4.14}$$

Inserting (4.14) into (4.13) and substituting

$$Q(x, k) = S(k)\sqrt{\frac{jk}{2\pi}}\sqrt{\frac{\Delta r}{r + \Delta r}}G(\varphi, 0, k)cos\varphi\frac{e^{-jkr}}{\sqrt{r}} \tag{4.15}$$

as the driving signal of the monopole secondary sources yields

$$P(\boldsymbol{r_R}, k) = \int_{-\infty}^{\infty} Q(x, k)\frac{e^{-jk\Delta r}}{\Delta r}dx. \tag{4.16}$$

Alternatively (4.15) can be expressed in terms of the normal component of particle velocity $V_n(x, k)$ at location $(x, 0, 0)$ due to a source with directivity function $G(\varphi, 0, k)$ at Ω by substituting

$$V_n(x, k) = \frac{S(k)}{\rho_0 c}G(\varphi, 0, k)cos\varphi\frac{e^{-jkr}}{r} \tag{4.17}$$

into (4.15) yielding

$$Q(x, k) = \rho_0 c\sqrt{\frac{jk}{2\pi}}\sqrt{\frac{\Delta r}{r + \Delta r}}\frac{1}{\sqrt{r}}V_n(x, k). \tag{4.18}$$

We have now reduced the plane of monopole secondary sources into a line array of secondary monopole sources with a strength proportional to the normal component of particle velocity at the location of each secondary source due to the source at Ω. Taking a closer look at (4.18) we still note that the value of the second square root - *the amplitude factor* - is proportional to the location of the receiver, through the distance Δr (Fig. 4.6). This is a result of reducing the contribution of the lines to single monopoles on the x-axis. This is unwanted because we would like to see a solution that is independent of the receiver position for deriving the driving functions of the secondary sources. Achieving this is possible with some restrictions using a so called *'reference line'*. Although it is not possible to synthesize the correct amplitude in the whole listening area, it is possible to render the correct amplitude on some reference line. This is done by using the layout depicted in Figure 4.7. Consider a primary source Ω at location $(0, -r_0, 0)$, secondary monopole sources on the x-axis and a receiver at position R on the reference line at $y = \Delta r_0$. It can again be shown

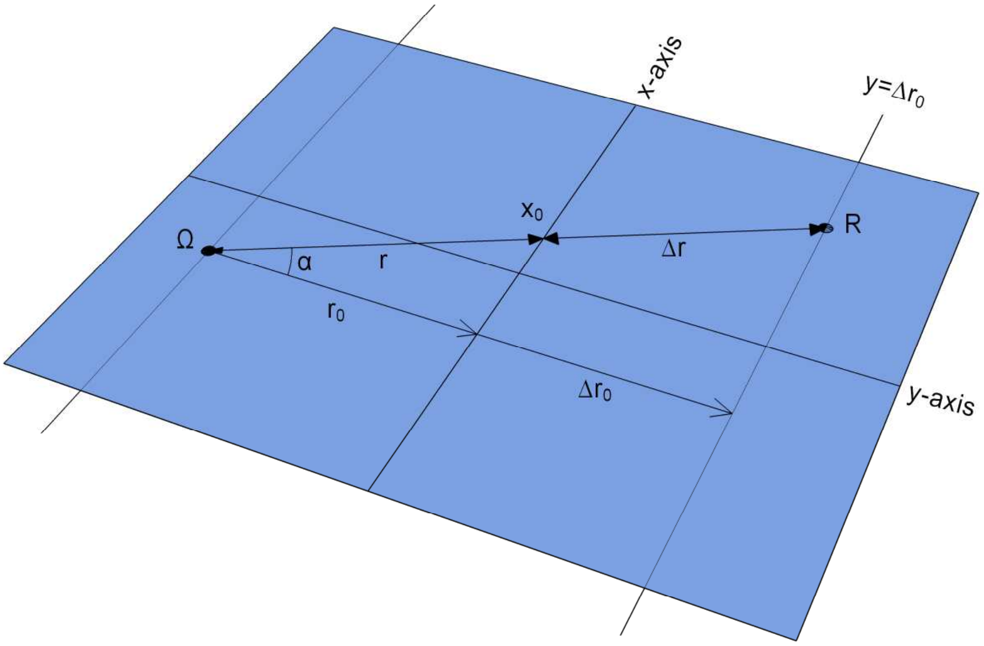


Figure 4.7: *Layout for using the 'reference line' method.*

with a stationary phase approximation (see for example [54] or [32]) that the main contribution to receiver point R comes from the secondary source at location x_0 at the intersection of the x-axis and the line from Ω to R. By fixing r to r_0 and Δr to Δr_0 we get the correct amplitude in a receiver point where the line between Ω and R is perpendicular to the x-axis (Fig. 4.7). By carefully looking at the geometry in Figure 4.7 it can be easily noted that the amplitude factor is correct for the whole line $y = \Delta r_0$ (using positive distances):

$$cos\alpha = \frac{r_0}{r} = \frac{\Delta r_0}{\Delta r} = \frac{r_0 + \Delta r_0}{r + \Delta r} <=> \sqrt{\frac{\Delta r}{r + \Delta r}} = \sqrt{\frac{\Delta r_0}{r_0 + \Delta r_0}}. \tag{4.19}$$

While getting the correct amplitude on the reference line, we are getting too high amplitude between the line of secondary sources and the reference line and too low amplitude when $y > \Delta r_0$. The amplitude errors are quite small, usually not higher than 1.5dB [54]. By choosing the optimal distance (e.g. the most probable listener location) for the reference line we can achieve almost correct amplitude in a large listening area.

Equation (4.18) holds for virtual sources behind the secondary source array. For sources between the array and receiver position the driving signals are different. For these *focusing* sources a driving signal similar to (4.18) can be derived [54] using again the method of the reference line:

$$Q^{foc}(x,k) = \rho_0 c \sqrt{\frac{k}{2\pi j}} \sqrt{\frac{\Delta r_0}{\Delta r_0 - r_0}} \frac{1}{\sqrt{r}} V_n^{foc}(x,k), \tag{4.20}$$

where

$$V_n^{foc}(x,k) = \frac{S(k)}{\rho_0 c} G(\varphi, 0, k) cos\varphi \frac{e^{+jkr}}{r} \tag{4.21}$$

Again Δr_0 is the perpendicular distance of the reference line from the array of secondary sources and r_0 is the perpendicular distance of the virtual source from the array. Again the amplitude of the synthesized wave field is correct at the reference line at $z = r_0$.

We have now simplified the Kirchhoff-Helmholz integral to a situation where we need only a continuous distribution of monopole secondary sources on a infinitely long line to synthesize the wave field produced by some source Ω. This creates only small amplitude deviations in the synthesized wave field in the horizontal plane which can be considered acceptable taking into account the level of simplification. Still, in a real world application the infinitely long line of secondary sources has to be truncated and discretized due to physical and practical limitations. A practical approximation of the line of secondary sources is a linear array of speakers located on some wall of the listening space. The effects of this approximation are considered next.

4.3 Issues of practical implementation

In implementing the method described in Section 4.2.4 in practise, two different issues arise. First the infinite line of the continuous distribution of secondary sources must be discretized. This produces an effect called *spatial aliasing* due to spacing between the discrete sampling points.

Furthermore the infinitely long line has to be truncated to some length allowed by the physical limitations of the listening space. This introduces diffraction artifacts in the synthesized wave field due to both ends of the array. Also the effective listening area is reduced.

4.3.1 Discretization of the WFS array

As mentioned earlier discretizing the linear array of secondary sources introduces spatial aliasing in the synthesized wave field. The effect is comparable to *temporal aliasing* which occurs if a signal is sampled with a sample rate below two times the highest frequency present in the sampled signal [40]. Spatial aliasing occurs if the sampled sound field contains frequencies above

$$f_{max} = \frac{c}{2\Delta x \sin\theta_{max}} \geq \frac{c}{2\Delta x}, \tag{4.22}$$

where c is the speed of sound, Δx is the spacing of secondary sources and θ_{max} is the maximum angle of incidence from which field components reach the array [54, 18]. In worst case scenario the wave field contains components with an angle of incidence of 90 degrees and f_{max} gets its minimum value (4.22). This is the most interesting case because in the general situation no spatial aliasing is introduced below the minimum value of f_{max}. The effect of different speaker spacings are depicted in Figure 4.8 in a room with size of 4 m by 4 m with a 4 m long array on one wall. A virtual source is generated behind (Fig. 4.8a, b, c) and in front (Fig. 4.8d, e, f) of the array.

Note that the spatial aliasing can be removed by low pass filtering the primary signal with a cut-off frequency of f_{max} from (4.22). In addition this reduces the effective frequency range, which in most cases is not wanted.

As described in Chapter 3, the interaural time differences (ITD) are the major cues for sound source localization for frequencies below 1600 Hz. For this reason using speaker spacings below approximately 11 cm (4.22) does not significantly degrade the localization of primary sources ([50]). Even larger speaker spacings produce still rather good localization of virtual sources and if the speaker spacing does not significantly exceed the value of 11 cm, the main perceived effect of spatial aliasing is a place-dependant coloration of the synthesized sound field due to the spectral distortion [18]. It has also been shown that good localization of virtual sources can be achieved with speaker spacing of 17 cm [37].

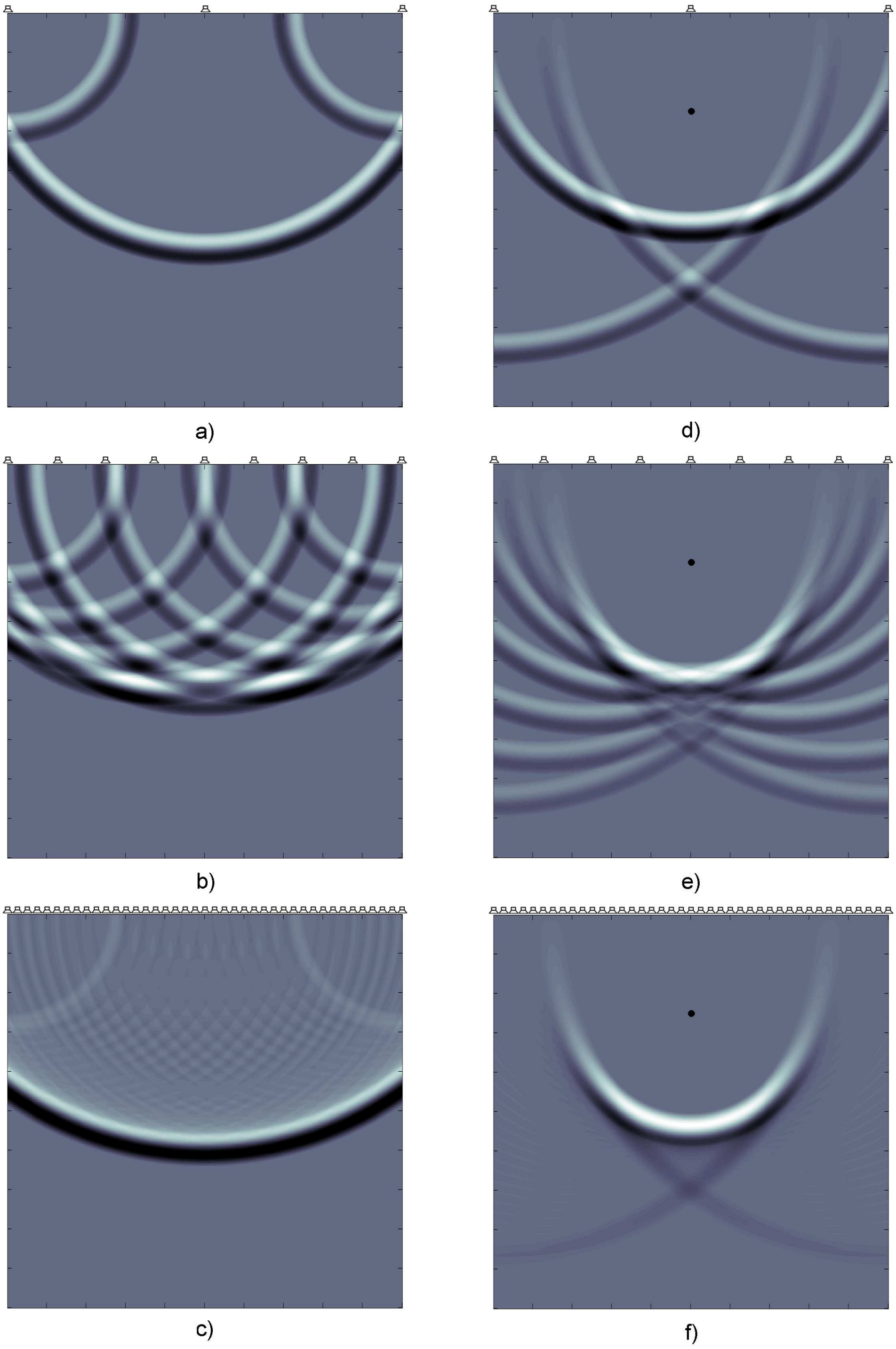


Figure 4.8: *Different speaker spacings and the produced response with spatial aliasing simulated in a 4 m by 4 m room. On the left the virtual source is located 1 m behind the array. a)* $\Delta x = 2m$ *b)* $\Delta x = 0.5m$ *c)* $\Delta x = 0.1m$*. On the right the virtual source is located 1 m in front of the array (marked with a dot). d)* $\Delta x = 2m$ *e)* $\Delta x = 0.5m$ *f)* $\Delta x = 0.1m$*.*

The effect of spatial aliasing can be reduced by applying spatial anti aliasing filtering or using highly directive speakers and/or microphones. The methods can be referred from [54], but remain outside the frame of this thesis, leaving the speaker spacing Δx the main variable in controlling spatial aliasing.

4.3.2 Truncation of the WFS array

Truncating the continuous line distribution of secondary sources reduces the effective listening area. The effect can be compared to the shadowing effect of window borders when light is coming trough it. See Figure 4.9 for further explanation. The only solution for this is to alter the width of the array of secondary sources. This can be done by adding more sources to the array or by increasing the spacing of the sources, though this introduces spatial aliasing artifacts described in Section 4.3.1. The choice of array length is therefore a compromise between array cost and the introduced artifacts – the reduced effective listener area and the spatial aliasing.

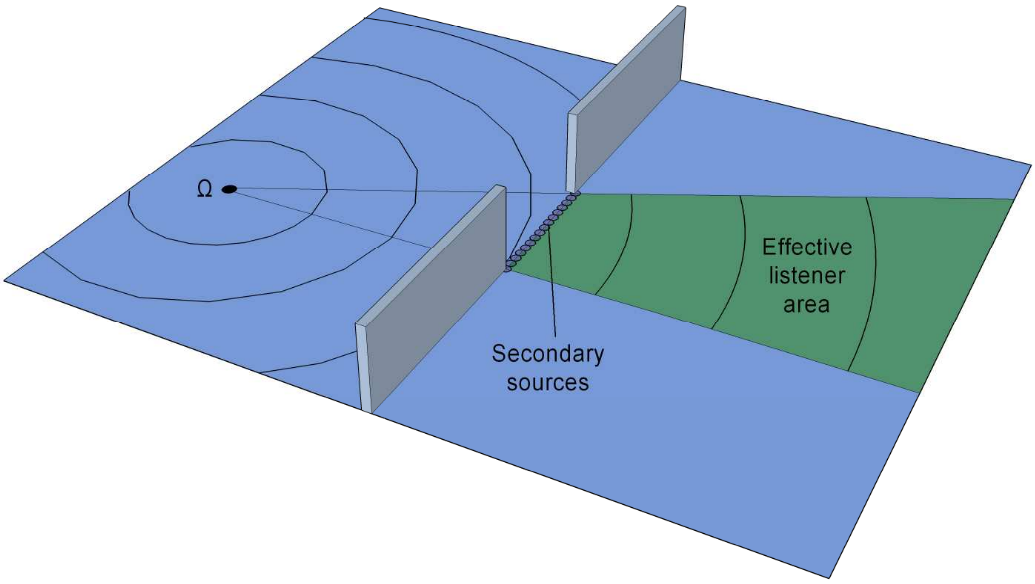


Figure 4.9: *The shadowing effect introduced by truncating the infinitely long line of secondary sources. The effective listening area from which the virtual source Ω can be heard is reduced.*

Truncation of the array also introduces diffraction artifacts from both ends of the array in the synthesized wave field. As shown in Figure 4.9, the effective listener area can be roughly sketched by drawing lines from the virtual source through the end points of the array of secondary sources. Inside this area the diffraction waves interfere with the synthesized wave field and outside the area the diffraction waves bend around the ends of the array. Figure 4.10a shows an array that is truncated from both ends to total length of 2 m. The array is producing a plane wave with a 20 degree angle of incidence into a listening area of 4 m by 4 m. The diffraction artifacts are clearly visible. The centers of the diffraction waves are apparently at the ends of the array.

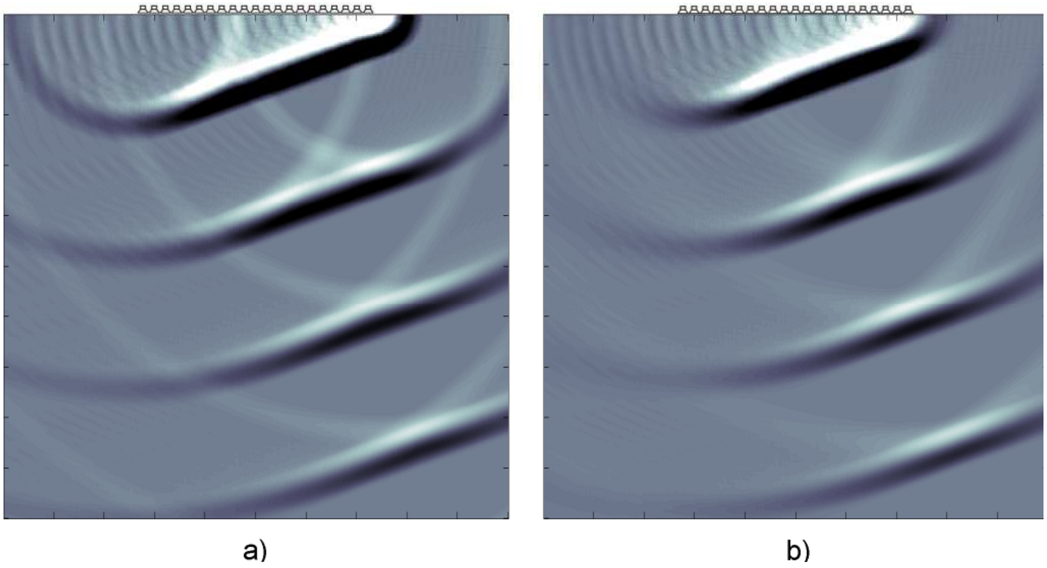


Figure 4.10: *A WFS array with length of 2 m is producing plane wave into a listening area of 4 m by 4 m. a) Diffraction caused by the truncation of the array from both ends. b) a 0.5 m Hanning taper applied to reduce diffraction at the both ends of the array.*

The diffraction artifacts can be effectively reduced by using a method called *tapering*, first introduced in [56]. In tapering the amplitude of the driving signal is gradually decreased towards the truncated end of the array. Best results can be achieved with some cosine taper, for example a half-Hanning window on both ends of the array. The results of tapering are depicted in Figure 4.10b. Attenuation of 6 to 10 dB of the diffraction waves can be achieved using a taper length of 25%. [54]. Applying tapering also introduces too low amplitude on the borders of the effective listening area so choosing the taper length is a compromise between attenuation of the diffraction waves and correct amplitude in the listening area.

In this section it has been show that practical implementation of the infinitely long array of secondary sources introduces artifacts to the synthesized wave field. It is important to take these into account in designing the reproduction and recording system for the acoustic opening. The level of the artifacts can be reduced with careful planning and with quite simple solutions. After reduction, the level of the artifacts is acceptable and does not introduce significant errors in the synthesized wave field.

4.4 Wave field extrapolation

Using the same amount of speakers and microphones in an acoustic window is not always possible and other approaches have to be considered. Furthermore, there is usually some distance between the microphones and the loudspeakers. In addition to WFS, a method for extrapolating the recorded audio signal from N microphones to M loudspeakers, the Wave Field Extrapolation (WFE) was introduced in [8]. In case of using microphones with cardioid directivity pattern, this can be viewed as using each of the microphones signals as an approximation of the particle velocity caused by some source distribution according to (4.18), as described later on in Section 4.6. An example of a WFE system is illustrated in Figure 4.11. For each microphone, the captured signal V_{ni} is extrapolated over the distance $|\Delta r_j|$ to each loudspeaker $(1 \leq i \leq N, 1 \leq j \leq N * M)$.

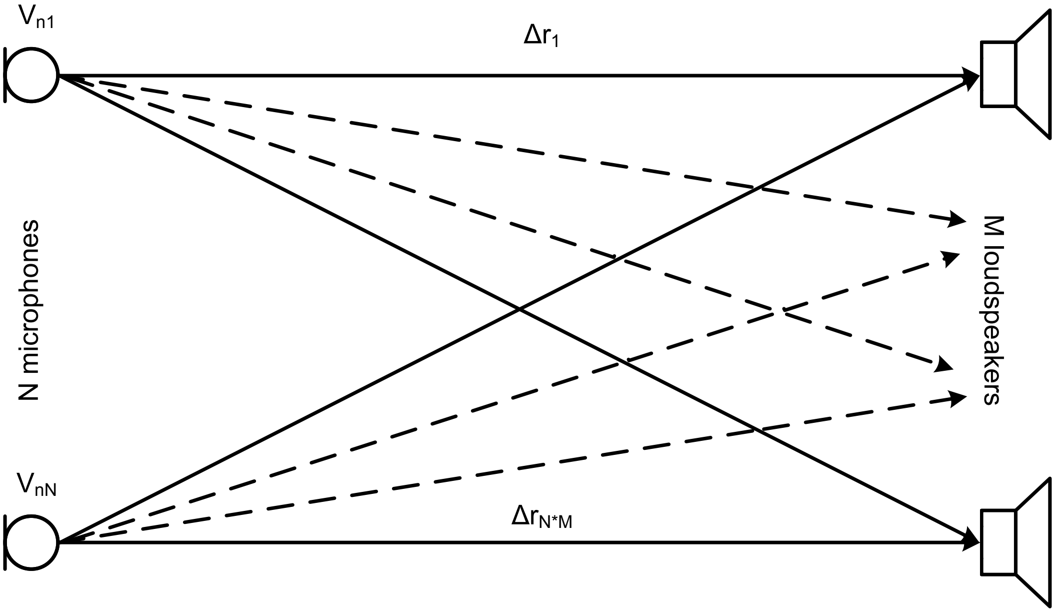


Figure 4.11: *Captured wave field is extrapolated from N microphones to M loudspeakers over the distances $|\Delta r_j|$ using WFE.*

While sampling the wave field with an array of microphones we introduce the same kind of spatial aliasing to the rendered wave field as described in Section 4.3.1. For this reason using the same spacing for microphone and speaker arrays would be optimal, this way we introduce the same spatial aliasing artifacts to the rendered wave field in both cases.

4.5 Comparison to other reproduction methods

In creating the acoustic window, various reproduction methods can be considered. The most obvious ones are stereo reproduction systems and the 5(.1) systems used in home theater applications. These systems have the advantage of being widely used and being quite inexpensive. There are also some serious drawbacks. In creating the acoustic window, the temporal as well as the spatial properties of the wave field travelling through the window should be correct. With stereo and 5 channel setups this can be achieved only in a small area, often only at one *sweet spot* (different speaker configurations have been discussed for example in [20, 10]). With WFS the properties are correct in a large listening area, making WFS superior to the other methods in this aspect. The differences of the methods can be seen in Figure 4.12. Each figure illustrates the wave field in a listening area of 4 m by 4 m and the virtual source is located centered 1 m behind the front wall. The WFS array used is 2 m in width and featuring 0.1 m speaker spacing yielding 21 speakers.

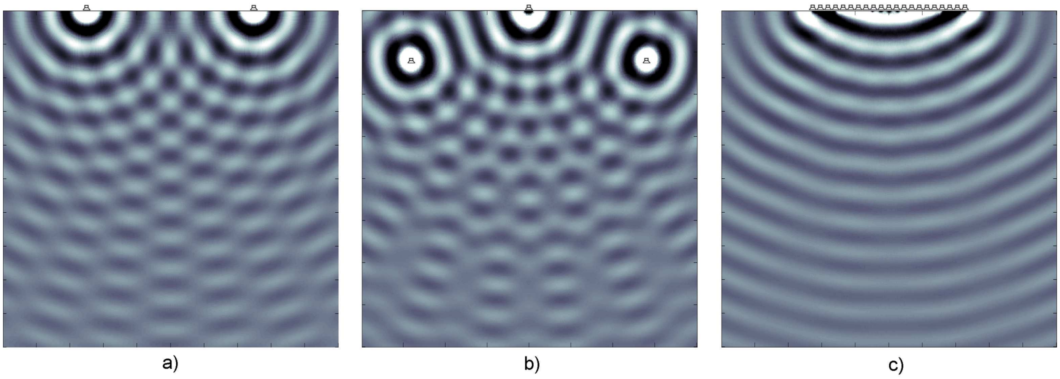


Figure 4.12: *Different speaker configurations synthesizing a virtual source 1 m behind the front wall of a listening space of 4 m by 4 m. a) a stereo setup b) a typical home theater setup with front loudspeakers activated c) a 2 m WFS array with 21 loudspeakers with spacing of 0.1 m.*

The cost of a WFS system exceeds the cost of the other systems by a big margin, but the achieved advantages are obvious. In addition a WFS system requires more computational power to calculate the driving signals of each speaker as described in Section 4.2.4. As the costs of digital signal processors and speaker technology are constantly falling the cost margin of WFS system is constantly decreasing. This makes the solution more interesting considering the advantages.

4.6 Requirements for hardware

In Section 4.2.4 we concluded that using only monopole sources is sufficient in synthesizing the wave field generated by some virtual source. This is easily achieved by using small loudspeakers, which are essentially omnidirectional for not too high frequencies. As concluded in Section 4.3.1 the loudspeakers should not exceed size of 11 cm by much to reduce the effect of spatial aliasing. With this loudspeaker spacing there exists no spatial aliasing below approximately 1600 Hz. To simplify the setup, using active loudspeakers removes the need of separate amplifier. Additionally this approach requires the normal component of the particle velocity at the loudspeaker position as part of the driving signal of the speaker. This could be realized by using velocity microphones (i.e. microphones with *dipole* directivity pattern). In our approach however dipole microphones do not produce the best result.

In creating an acoustic window we place the speaker and microphone arrays on some wall of the listening space, speakers on the reproduction side of the window and microphones on the recording side. Having the microphones against a wall produces reflections from the wall to the back lobe of the dipole microphone. This is unwanted because we just want to sample the wave field travelling through the window. The simplest way to achieve this is to remove the back lobe of the dipole microphone. This however is not possible and microphones with just the front lobe do not exist. [32] has shown that using microphones with *cardioid* directivity pattern produces a very good approximation of the dipole microphones if the direction of the sound propagation is known. Obviously this is the case with the microphones on the wall. Using cardioid microphones also effectively removes the back lobe of the dipole directivity pattern. A comparison of the directivity patterns can is in Figure 4.13.

Additionally, a computer powerful enough to calculate the driving signals (4.18) is needed. In a consumer product a DSP would be used for simplicity, but for versatility a general purpose personal computer was used in the experiments in this report. To realize the requirement of small component spacing the computer should also have a sound card with high number of inputs and outputs. Thus, the hardware requirements that the theory of WFS introduces for the acoustic window are the following:

- For using the method presented in Section 4.2.4 we need speakers with monopole directivity characteristics. The size of the loudspeakers should not exceed 11 cm by much to avoid significant spatial aliasing as concluded in Section 4.3.1. To simplify

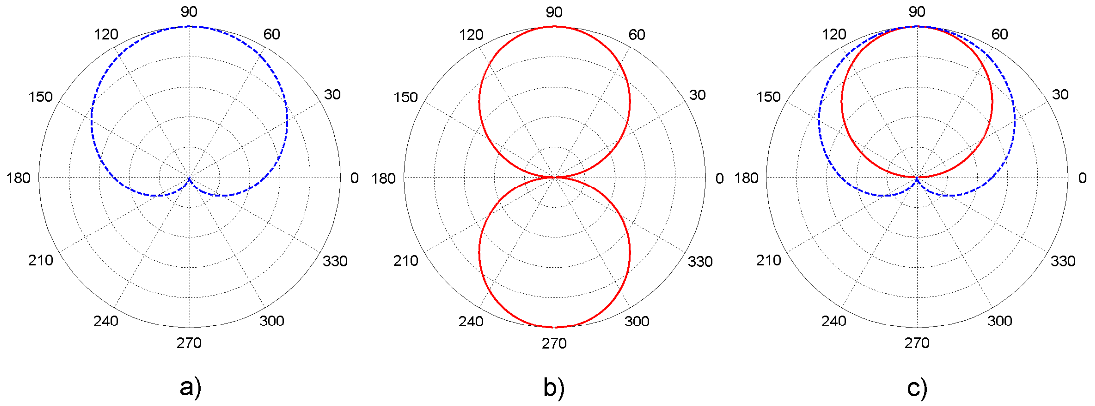


Figure 4.13: *Different microphone directivity patterns compared. The microphones are pointing in the direction of 90 degrees. a) Cardioid, b) Dipole, c) The cardioid compared to the dipole's front lobe.*

the setup, active loudspeakers are preferred.

- A computer to calculate the driving signals for the loudspeakers.
- A sound card for the computer with high number of inputs and outputs to realize the requirement of small loudspeaker and microphone spacing.
- Microphones with cardioid directivity pattern to sample the wave field.

4.7 Conclusion

WFS offers an accurate method for producing a close to natural sound field in a large listening area with an array of loudspeakers and microphones. Realization of the array causes some deviations in the rendered wave field but these can be restricted to a low level with careful consideration of the hardware and signal processing used. Using a small spacing in the arrays and keeping the length of the arrays long enough for the designed listening area, we can achieve high quality spatial audio reproduction. WFS offers versatility that is unmatched by any of the current loudspeaker configurations or communication systems and therefore is selected to realize the acoustic opening.

Chapter 5

Description and verification of the system

5.1 Introduction

When considering building an acoustic opening several factors come up. We have to consider both the hardware requirements of the system and the signal processing involved. Requirements for loudspeaker and microphone setups have been discussed in Section 4.6. In addition, the theory behind the sound reproduction algorithms for the system were introduced in Section 4.2. The system components for building the acoustic opening are selected according to the conclusions derived in these sections. The components are then carefully measured and the achieved performance is measured to verify their usability for the acoustic opening. Furthermore an implementation for a multichannel acoustic echo canceller is needed in the system. The implemented solution is described with performance simulations.

5.2 Hardware

5.2.1 Reproduction system

It was concluded in Section 4.6 that the loudspeaker spacing in the array should not exceed 11 cm by a big margin. It was also concluded that to keep the system simple and to remove the need of a separate amplification system, active loudspeakers should preferably be used. To fulfill these requirements 24 active small size monitor loudspeakers (M-Audio StudioPro 3) were selected, 12 on both sides of the acoustic opening. The width of one loudspeaker is 14 cm yielding a total length of 1.68 m for the full linear array. The setup is shown in Figure 5.1.

This also produces a 14 cm spacing for the loudspeaker elements. According to (4.22)

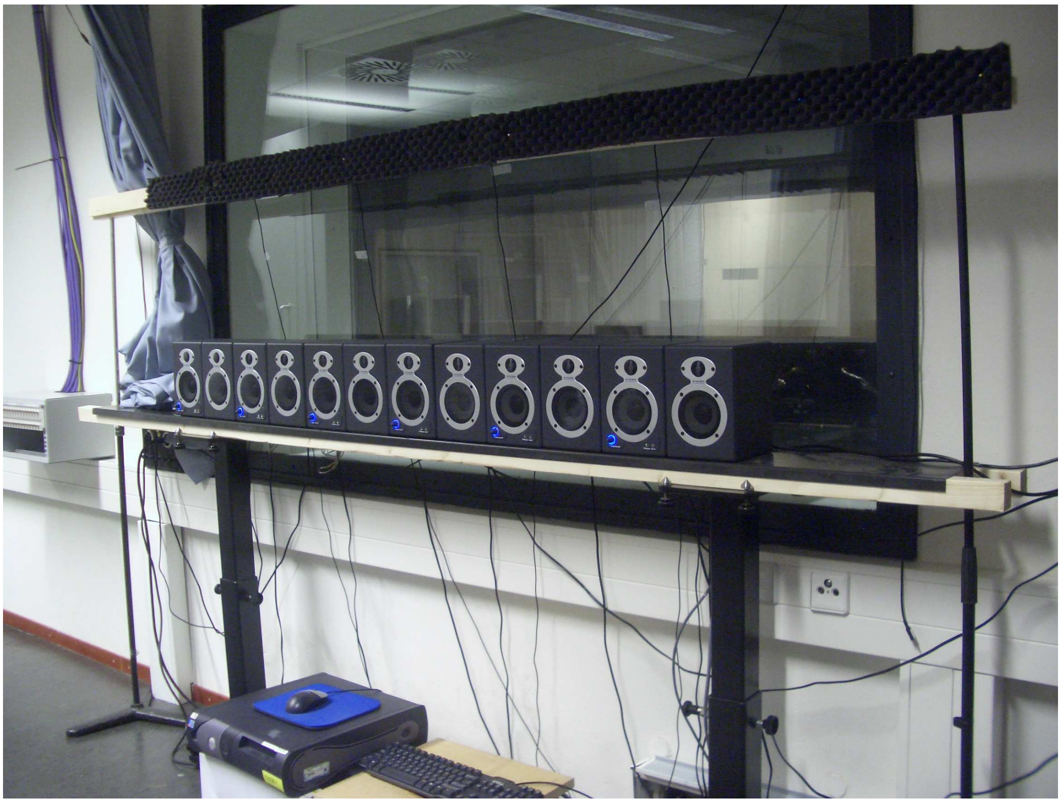


Figure 5.1: *The hardware setup for the acoustic opening. The system features 12 loudspeakers with 14 cm spacing and four microphones with 42 cm spacing installed in a frame above the loudspeakers on both sides.*

this creates spatial aliasing for frequencies approximately above 1214 Hz for sound velocity of 340 m/s. As concluded in Section 4.3.1, this should produce already good results in localization for speech communication.

The loudspeakers used are sold in pairs. The first loudspeaker of the pair contains the amplification circuitry and the second one is without the circuitry. This could cause performance differences between the two loudspeakers. For this reason the loudspeakers were measured to verify their frequency response and directivity characteristics. Using WFS in the sound reproduction system requires that each loudspeaker is identical and produces close to omnidirectional directivity pattern, i.e. radiates evenly into each direction as concluded in section 4.2. The measurements were conducted in an anechoic chamber and the results are displayed in Figures 5.2 and 5.3). Overall imperfections in the frequency response (i.e. the same for all loudspeakers) can be easily compensated for using a filter at the input. It can be concluded that the loudspeakers' frequency responses and directivity patterns are identical enough to use them in a WFS system.

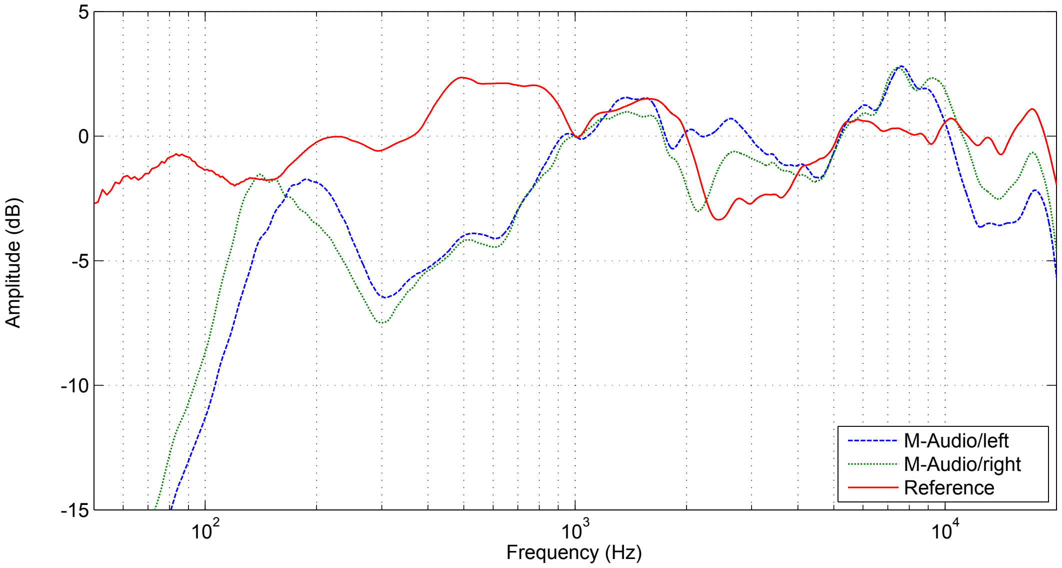


Figure 5.2: *The frequency response of the used active loudspeakers, both the left and right units, were measured in an anechoic chamber and compared against a good quality reference loudspeaker. The responses are normalized to 0 dB at 1000 Hz for easy comparison.*

5.2.2 Recording system

The aim for the recording system is to capture the whole wave field impinging the range of the acoustic window in the recording room and pass the signal to the loudspeaker system on the other side to reproduce the captured wave field as accurately as possible. The sound recording system consists of four microphones (Audio-Technica Pro 45) on both sides of the opening with 42 cm spacing to cover most of the length of the loudspeaker arrays, the total length of the array is 1.26 m. The microphones are located above the loudspeakers to decrease the effect of direct coupling with them. The microphone setup can be seen in Figure 5.1.

As for the reproduction system, the quality and similarity of the microphones needs to be evaluated. Section 4.2 suggests that microphones with a cardioid directivity pattern should be used in conjunction with omnidirectional loudspeakers. Also the sensitivity differences between the microphones should be small to capture the wave field accurately (for details about microphone mismatch, see [21]). The frequency response of each microphone was measured in an anechoic chamber. The results are shown in the upper part of Figure 5.4. After the measurements, the microphone signals were normalized to have the same level at

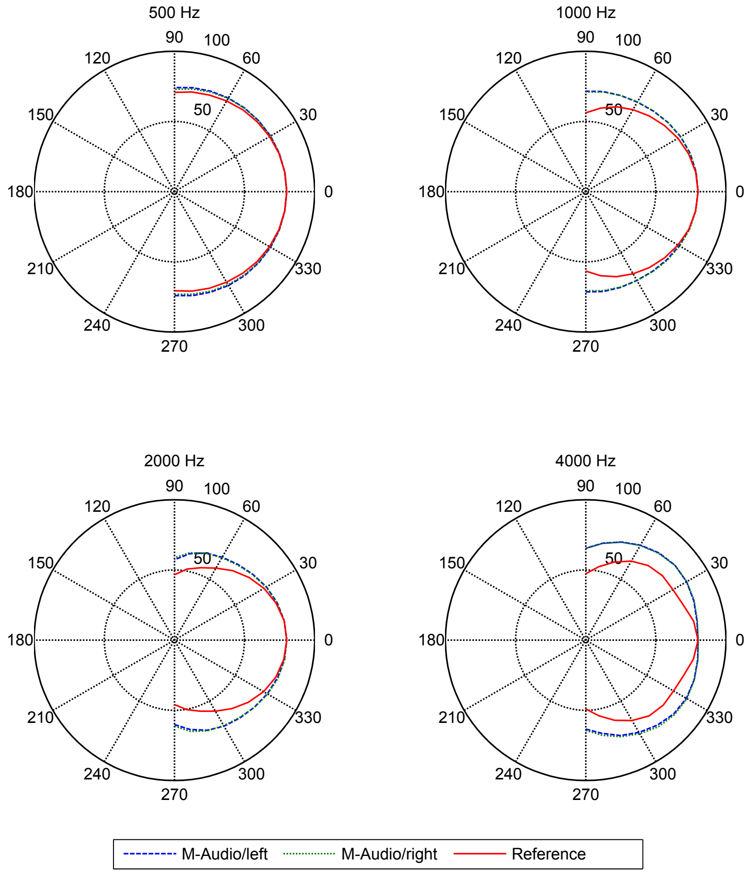


Figure 5.3: *The directivity pattern of the active loudspeakers, both the left and right units, were measured in an anechoic chamber and compared against a good quality reference loudspeaker. The responses are normalized to on-axis response at 80 dB.*

$500Hz$. The results can be seen in the lower part of Figure 5.4.

It can be concluded that the normalization produces good usability for the microphones in acoustic opening due to the similarity of the responses. Also the results from the directivity pattern measurements support this fact. Especially at 1000 and 2000 Hz the pattern is close to a cardioid, and most importantly for all frequencies the sensitivity at the front side is considerably higher than at the back. The results are depicted in 5.5.

The number of microphones was selected smaller than the number of loudspeakers to keep the overall complexity of the system low. To render the captured wave field on the reproduction side, the microphone signals need to be extrapolated to 12 loudspeakers using WFE. The technique is described in Section 4.4. By this selection we introduce strong spatial aliasing to the captured wave field due to the large spacing of the microphones. Ac-

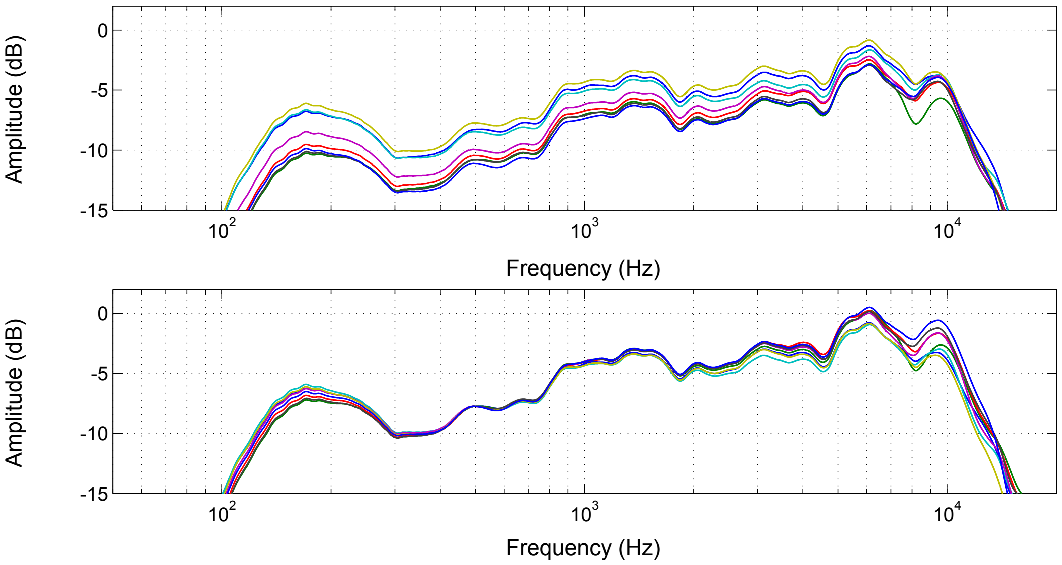


Figure 5.4: *Upper figure: The measured frequency response of all of the eight microphones. Lower figure: Normalized frequency response of the microphones to produce similar results for speech frequencies. The responses are normalized to have the same level at 500 Hz.*

cording to (4.22) frequencies above approximately 404 Hz show some aliasing effects. Due to this fact the quality of the recording and reproduction system is investigated and listening tests are performed to compare WFE against other recording methods in conjunction with WFS. The listening tests are described in Section 6.

5.3 Multichannel acoustic echo canceller

While building a two-way communication system, we always run into one fundamental problem. By placing microphones and loudspeakers on both sides of the communication path we create a closed acoustical loop between the the two sides of the system. A signal recorded with microphones on one side is reproduced with loudspeakers on the other side. In addition to propagating into the communication room, the reproduced sound is also picked up by the microphones in the room. The same also happens on the other side of the communication and this produces an audible echo in the communication if the delay between the two sides is high. If the level of the echo increases on each loop and the delay between sides is small, the system becomes unstable causing a loud, howling sound - known as *the Larsen effect* [52] - that prevents the systems use for communication.

Considering a real opening between two separate rooms, the room reflections from each room would propagate to the other room and be a part of the resulting sound field there, and

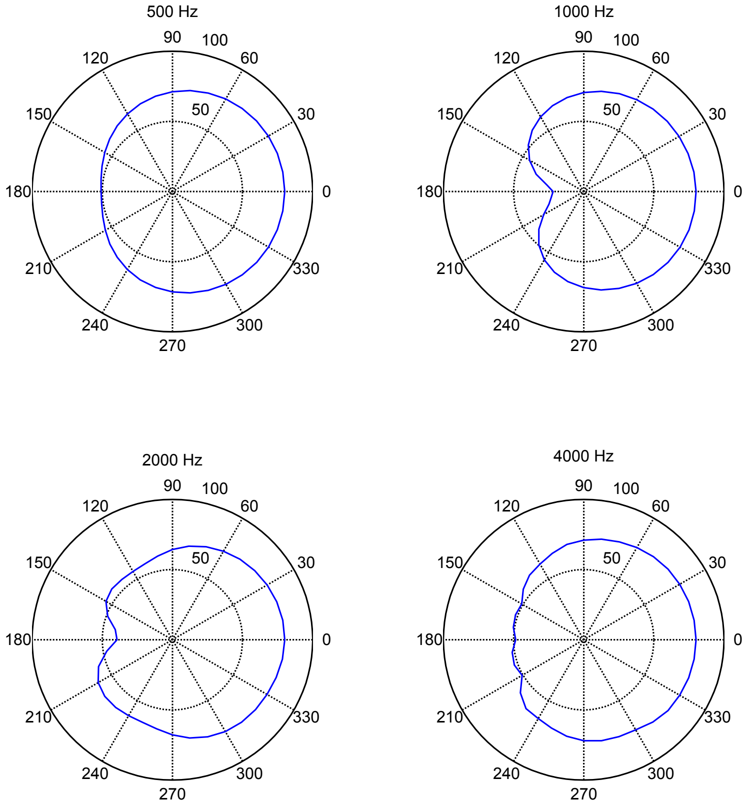


Figure 5.5: *The directivity pattern of the used microphones. The response is normalized to on-axis response at 80 dB.*

therefore they should also be reproduced by the virtual opening. For this reason only the direct sound from the loudspeakers to the microphones should be cancelled.

Building a two-way communication system as described in this section creates a need for a multichannel acoustic echo canceller. The scale of the problem increases quickly when additional components are introduced to the system. An ordinary communication system with just one microphone and two loudspeakers has only two possible direct paths for the sound to reach the microphone. The described solution using four microphones and 12 loudspeakers on each side has a total of 48 acoustic paths that cause direct coupling between the components. The effect of these paths should be removed from the recorded and reproduced sound as they are not characteristics of a real physical opening.

A real time two-way communication system with multichannel acoustic echo canceller was implemented and is described in Figure 5.6. The solution is a combination of techniques described in Section 2. The input-output system was running at 44.1 kHz sampling rate, but to decrease the computational load, all processing was done with downsampled signals and at a sampling rate of 22050 Hz. Also antialiasing filtering was applied before downsampling and after upsampling according to Nyquist-Shannon sampling theorem [40]. The anti-aliasing filter was a 5th order IIR filter with a cut-off frequency at 7 kHz.

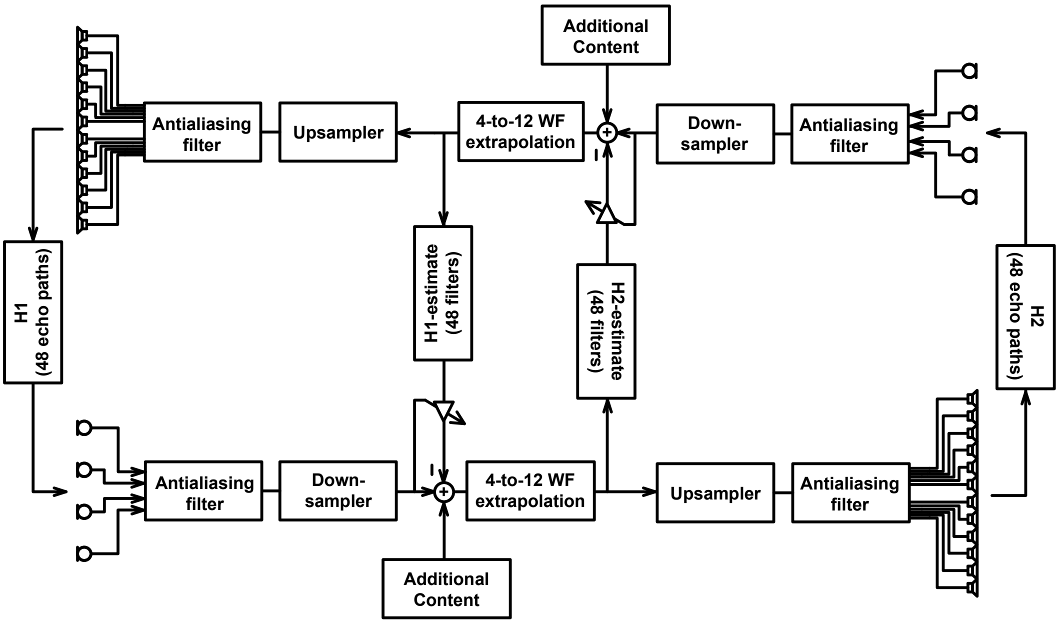


Figure 5.6: *An implementation of a multichannel acoustic echo canceller.*

All the 96 acoustic echo paths were measured off-line and the acquired impulse responses were then truncated to m coefficients to have control over the total computational load of the system. Each loudspeaker signal is continuously filtered with the corresponding impulse responses and the result is multiplied with an adaptive normalized least mean squares (NLMS) gain [24] (see Section 2.2 for details). The end results are then subtracted from the corresponding microphone signals to remove the directly coupled sound. The adaptation tries to minimize the energy of the microphone signals. This can be done because the direct sound from the loudspeakers and the other sounds from the room excluding the room reflections are uncorrelated. Therefore, in the optimal case the adaptation maximizes the amount of echo removed while preserving all other sound events. An example of a measured impulse response without the initial delay between one loudspeaker and one microphone with the truncated version is shown in Figure 5.7.

The truncated impulse responses represent estimates of the real acoustical paths and therefore different performance can be expected with different lengths. The results of the echo canceller were simulated with the recorded data and with a wide-band white noise signal played from each loudspeaker. The amount of attenuation in the direct sound arriving at one microphone from 12 loudspeakers is represented in Figure 5.8 as a function of the filter

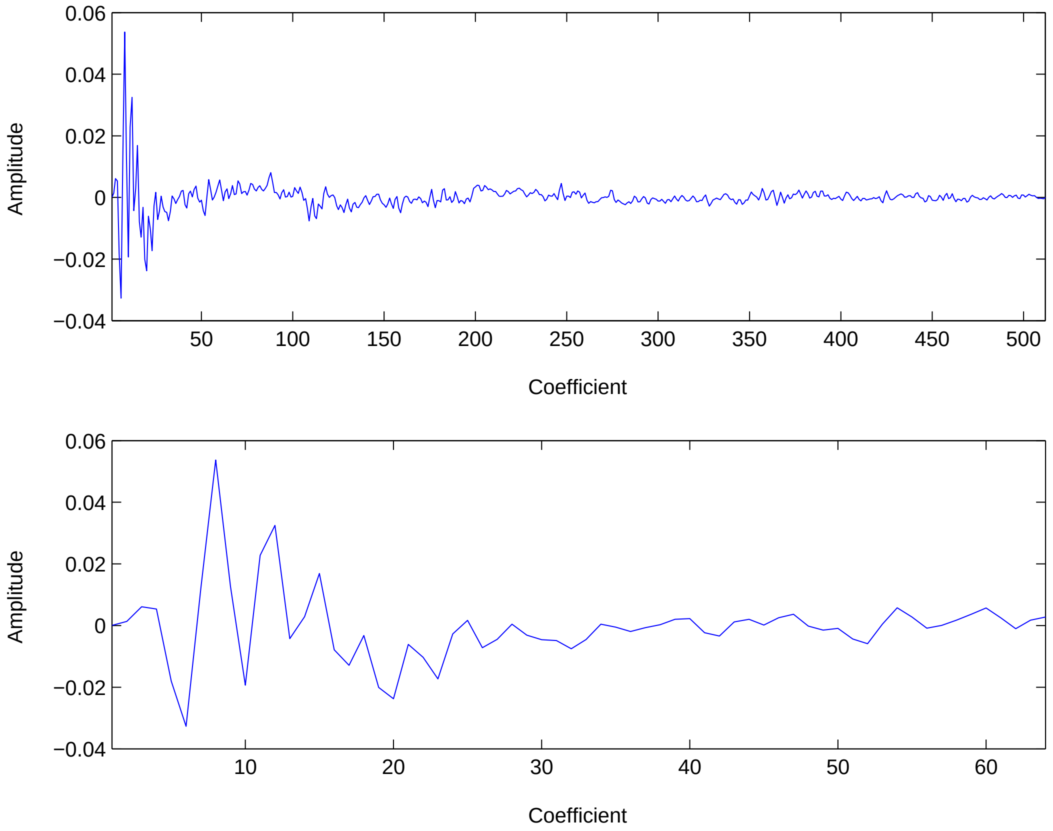


Figure 5.7: *Upper part: One example of an impulse response measured between system loudspeaker and microphone without the delay between the components. Lower part: The same response truncated to 64 coefficients.*

length.

The realtime system was running on a single Pentium4 class computer and the filter length was set to 16 coefficients. This already provides almost 5 dB attenuation in the direct sound received by the microphones. In practical use it makes normal conversation through the acoustic opening possible with almost no audible echo in the conversation. It can be seen from Figure 5.8 that doubling the filter length from 16 to 32 coefficients would also double the attenuation to approximately 10 dB. Additional doubling to 64 would not provide similar improvement, but moving from 64 to 128 coefficients provides 5 dB more attenuation.

It should be kept in mind that the results provided here are acquired from simulations and the achievable real world performance is lower. This is mainly due to the fact that the impulse responses are measured offline. Even small room temperature changes affect the

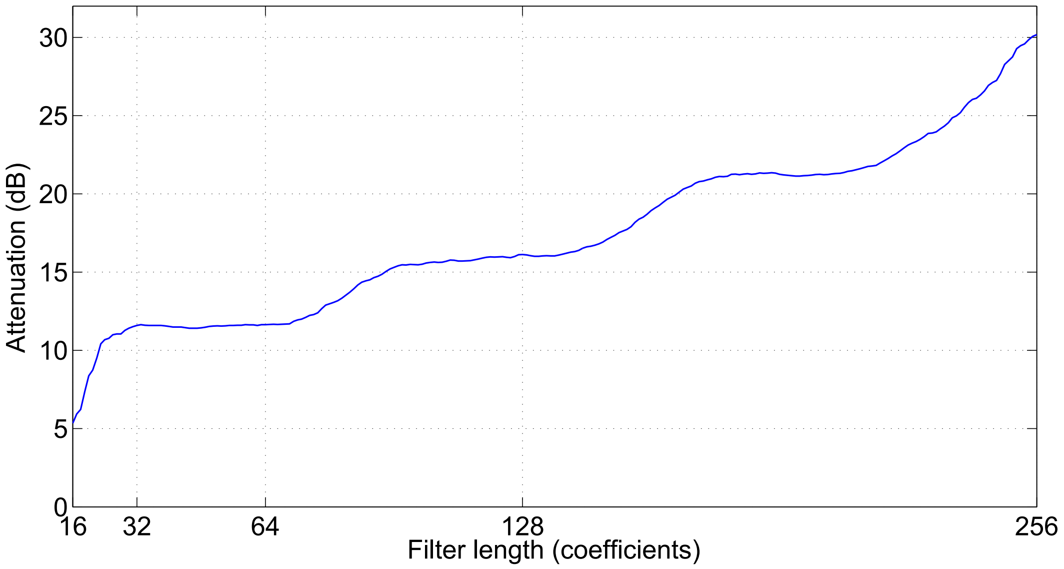


Figure 5.8: *Results of echo cancellation. Attenuation of the direct sound arriving to one microphone from 12 loudspeakers as a function of filter length.*

velocity of sound (see [22] for details) and this also affects the acoustical paths between the loudspeakers and the microphones. In addition the configuration in the room changes, i.e. people and objects move, causing again changes in the acoustical paths. A system that adapts the impulse responses on line would for this reason be preferred. For computational reasons this was not possible to perform in realtime in the system introduced in the current thesis.

The implemented acoustic echo canceller provided a stable multichannel communication system. Some minor artifacts existed in the reproduced sound field, but the system provided good usability in communication tasks. Due to the fact that relatively short filters were used, some frequencies in the system had an audible, ringing echo that attenuated slowly. The level of the ringing was low enough to not to disturb the communication through the system.

5.4 Software

A two part realtime software was developed to run the measurements needed for the multichannel acoustic echo canceller and to control the acoustic window system. In addition, a realtime software was written for the listening tests. Screen captures of the graphical user interfaces (GUI) of the software can be seen in Section B.

5.4.1 Measurement of the impulse responses

The first part of the software measures all the 2 x 48 = 96 acoustic echo paths involved in the system described above. The measurement was performed playing a logarithmic sweep signal through each of the system's loudspeakers one by one. The sweep is recorded by the 4 microphones on the same side of the acoustic window. This is repeated for both sides. From the recorded data, the transfer functions between all the loudspeakers and the microphones are calculated using frequency domain convolution. The impulse responses required by the multichannel echo canceller are then acquired with inverse Fourier transform (IFT). The results are saved into an audio file that is read later by the acoustic echo canceller.

5.4.2 The acoustic window

On startup, the software for the acoustic window reads the filter coefficients from the previously recorded file in addition with initial values for the adaptive NLMS filter gains. The software provides control over the adaptation process, both the filtering and adaptation can be turned on or off. This provides the ability to compare the performance of the system with different processing enabled. Furthermore the adaptation coefficients can be saved or loaded from a file during realtime processing. The user interface provides also functionality for other possible echo cancellation methods for future work described in Section 8. The basic functionality of the acoustic echo canceller and the WFE part are described in Figure 5.6.

Due to the flexible nature of the processing system, it provides ability to render additional sound events with the system. An example of this is playing different background music on each side of the window while having a conversation. The acoustic echo canceller makes this possible without crosstalk of the different music samples on each side.

5.5 Conclusion

The theory of WFS introduces requirements for the hardware used in building the acoustic opening. According to these requirements components for the system were selected. The components were carefully measured to verify their usability in the system. The loudspeakers were measured in an anechoic chamber to verify the similarity between the units. In addition, all the microphones used were measured also in the anechoic chamber to avoid level differences between them. According to this data, the microphone signals were equalized to produce similar results. It can be concluded that the selected hardware has the characteristics needed to be used in building the acoustic opening.

Direct coupling between the system's loudspeakers and the microphones in communication systems is a known problem and causes audible echo into the conversation. An implementation of a multichannel acoustic echo canceller was introduced and its effectiveness was tested by simulations. The echo canceller uses pre-measured data in conjunction with simple adaptation to cancel the acoustic echo from the system. The presented results show that the solution provides good usability and makes normal speech communication possible. Despite the scale of the problem with multiple channels, it is feasible with a Pentium4 class PC.

Chapter 6

Listening test

6.1 Introduction

In this section the sound field capture aspects of the acoustic window system described in Section 5 are evaluated with listening tests. To limit the attributes affecting the captured and reproduced wave field, only a one-way system is used for the tests, omitting the multichannel echo canceller described in Section 5.3. In the listening tests the WFE is compared against three other known methods for capturing and reproducing the wave field in the acoustic window. After the tests, the results are investigated and a conclusion of WFE's usability in a virtual acoustic window is derived.

6.2 Description of the test

The listening test described here is essentially the same that the author has presented earlier in [26].

6.2.1 Sound capture methods

For recording the sound and reproducing it at the other side of the acoustic window four different methods are compared. The first capture method investigated is based on a linear array of microphones and Wave Field Extrapolation (WFE), in which the captured wave field is extrapolated from N microphones to M loudspeakers. The method aims at capturing the wave field impinging on the wall segment and reproducing it on the other side of the window (see Section 4.4 for details). In the listening test, a 4-to-12 extrapolation is used with a linear microphone array of four microphones with 0.42 m spacing, so the total length matches the dimensions of the loudspeaker array. Using such a large spacing for the microphones introduces spatial aliasing in the recorded wave field above approximately

400 Hz as described in Section 5.2.2. For this reason, the localization accuracy of the audio sources could be expected to be decreased with the system.

With the second and the third methods the aim is to capture the sound sources as cleanly as possible assuming that no reflections exist in the room. The captured source signal is then used together with location parameters to synthesize the wave field as a combination of individual point sources on the other side using WFS (Eq. 4.18). In addition, we take into account the relation described in Equation (4.17) and assume monopole directivity for the recorded sources.

While WFE is a 'blind' method, i.e. it does not need any information about the source locations, the second and the third methods require either a priori knowledge of the source location or an implementation of source tracking to render the sound sources at the correct position.

The second method uses a generalized side lobe canceler beamformer [25] which is used to pick up the audio sources as dry as possible. An array with four microphones is used again, now with a 0.06 m spacing. The third method uses close-talk microphones to record the sources from close distance. The fourth method uses the signals from the close-talk microphones and renders them through single speakers in the array without processing. This method is used as a reference.

6.2.2 The test setup

The four capturing methods described above in Section 6.2.1 were compared in a listening test using the one-way system. The layout for the experiment is shown in Figure 6.1. The listener was seated in the acoustically treated room facing the loudspeaker array placed next to the window and on the level of the listeners ears. The listening distance was 2.5 m from the loudspeaker array. On the recording side, the two microphone arrays were placed directly under the window, matching the position and height of the loudspeaker array on the other side. Two audio sources were placed in the room with the microphones, symmetrically placed at angles of -30°and 30°(Fig. 6.1). The audio sources were loudspeakers and they were visible to the listener through the window. This should produce an accurate spatial localization of the sound sources as described in Chapter 3 if the acoustic localization is correct. The perpendicular distance from the microphone arrays was 2.5 m. Close-talk microphones were simulated using the original source signals that were fed to the loudspeakers.

In the beginning of the listening test the subjects were able to experiment with all the methods and sound samples used to familiarize themselves with the content. The test was performed in two parts. In the first part the subject listened to four sound samples captured with each of the four methods described above. The samples included a man-to-man conver-

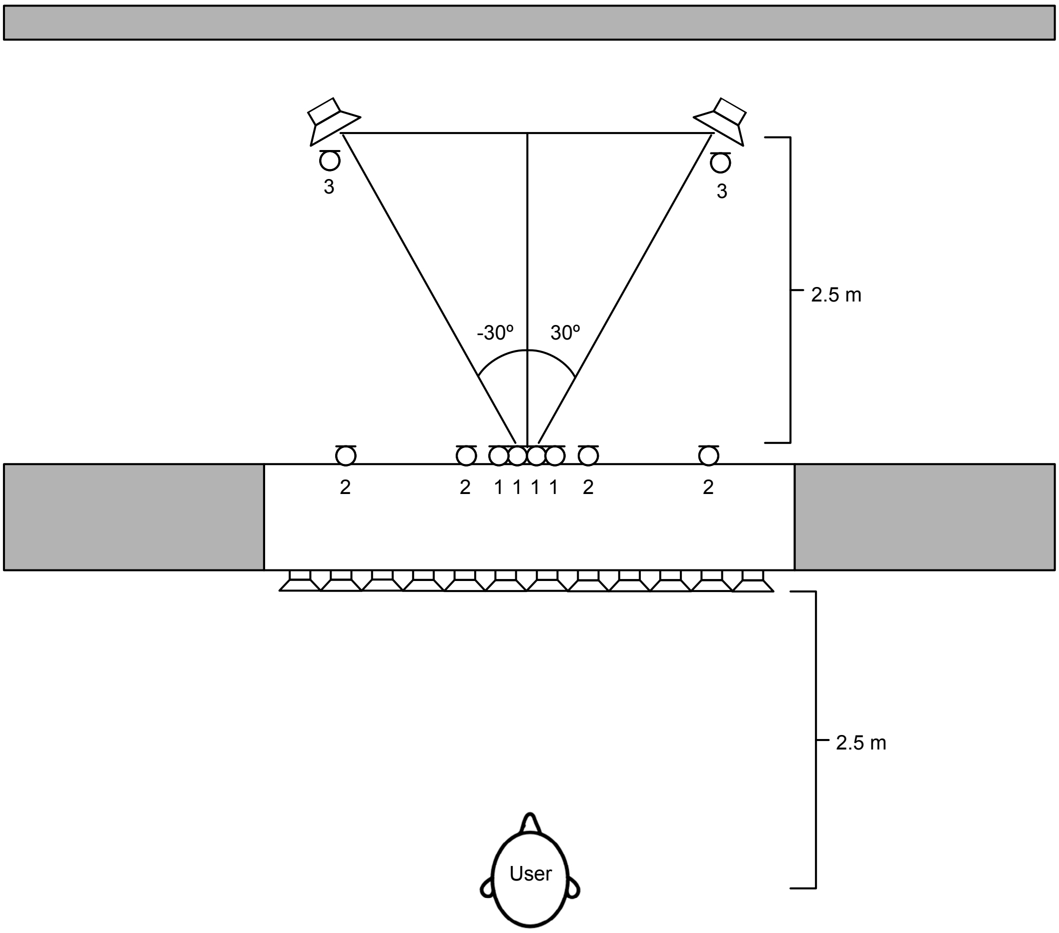


Figure 6.1: *Layout for the listening test. Two different microphone arrays are used, the first for the generalized side lobe canceller beamformer (1) and the second for WFE (2). Also close-talk microphones (3) are used. The wave field is synthesized with a loudspeaker array.*

sation, female-to-man conversation with overlapping speech, a duet of a string instrument and a brass instrument, and a male and a female singer in duet.

The spatial quality of audio multichannel reproduction systems and related attributes have been investigated in [6]. By combining their suggestions and the nature of the acoustic window system, four separate attributes were selected to be investigated. The subject evaluated the spatial naturalness (the quality of the spatial image) of the produced sound field. The subjects were instructed to imagine listening to two separate loudspeakers through an opening in the wall. In addition the coloration of the produced sound field was evaluated. In the evaluation a scale from 0 to 100 was used, 100 meaning ’good’ for spatial naturalness and 'bad' for coloration.

In the second part the listener was presented with the exact same stimuli, but now the distance of the sound sources from the listener was evaluated. The scale ranged between the back wall of the recording room and the subject's own position. In addition, the separation between the two sources was evaluated between 0 and 100, 100 equaling high separation.

The subjects were accustomed to participating in listening tests. The number of subjects participating in the test was 12.

6.3 Analysis of the results

The results from the listening tests are shown in Figures A.1 - A.20. The four methods are labelled as 'WFE', 'Beamformer', 'Close-talk' and 'Direct' in the graphs. For each combination of a sample and a method the results are depicted with a boxplot, which presents the median of the results in the middle of the box and the upper and lower quartiles. Additionally, the range of rest of the results is shown with a whisker plot (maximum whisker length is 1.5 times the interquartile range). Possible outliers are shown with crosses outside the whiskers.

6.3.1 Results with dialog samples

The listening test results with the sample featuring man-to-man dialog are shown in Figures A.1 - A.4 and results with the female-to-man overlapping dialog are illustrated in Figures A.5 - A.8. For the dialog, a clear influence of the effect of the recording room acoustics can be seen in the spatial naturalness scores (Fig. A.1). The highest values are given to the WFE, but the beamformer also provides relatively high scores for the spatial naturalness. A low but significant level of room reverberation remains in the signals captured with the beamformer and this probably gives an impression of higher spaciousness. Capture with close-talk microphones and the direct playback yield lower values for the naturalness, but also the results are more spread. This is probably due to the differences in the interpretation of the spatial naturalness scale among the listeners.

With the coloration aspect the methods are divided into two groups (Fig. A.2). The WFE and the beamformer cause a lot of coloration to the reproduced sound. The close-talk microphones and the direct reproduction receive low coloration results, mainly due to the use of original source signals.

The WFE's ability to capture a part of the room acoustics also enables creation of virtual audio sources far behind the loudspeaker array. This can be seen in all the distance graphs (Fig. A.3, A.7, A.11, A.15 and A.19). The best performance is achieved with the dialog sample. Also the beamformer is able to create a perceived position for the sound source clearly behind the loudspeaker array. In the two other methods, the close-talk microphones and the direct reproduction, the sources are typically localized close the loudspeaker array. This is a somewhat unexpected result because the WFS system should be able to produce an illusion of having the sources behind the loudspeaker array. Probably the results were influenced by the fact that the WFE method used in the same listening experiment was able to produce such a strong illusion of the distance of the source and the fact that the sources and the listener were in static locations.

There is a clearly visible trend in the separation of the audio sources. WFE produces the worst separation, but still the score is surprisingly high and comparable to the other methods, given the large spacing of the microphones. The best separation score is achieved with direct reproduction (the reference method) followed by the close-talk microphones and then the beamformer.

6.3.2 Results with the sample with instruments

The next sample contained a duet of a string instrument and a brass instrument. The results show similarities with the dialog samples, but the beamformer performs differently. The achieved spatial naturalness (Fig. A.9) is decreased and the amount of coloration is clearly higher than with other methods (Fig. A.10). The most probable reason for this is the fact that the beamformer was optimized for speech signals. The sound of the instruments has a broader frequency content than the dialog samples and therefore also the coloration artifacts are emphasized.

Again the WFE is able to re-create audio sources far behind the loudspeaker array, while the rest of the methods produce audio sources close to the array (Fig. A.11. The separation of the sound sources produces similar results than before with the dialog samples.

6.3.3 Results with the sample with singers

With the two singers (Fig. A.13 - A.16), results are again similar to the earlier results with the instruments. The coloration with the beamformer is again the highest, while close-talk microphones and direct reproduction show almost no coloration at all. WFE is able to render an impression of distant sources, but also surprisingly the beamformer reaches a high score, but with a high variance. Separation of the audio sources follows the same trend than with previous samples.

6.3.4 All results combined

When all the results are combined in the last four graphs (Fig. A.17 - A.20) , the same trends described above can be observed. However, the variances are now larger and one can see that in most cases there are no statistically significant differences between the beamformer, close-talk microphones, and two-channel reproduction.

6.4 Conclusion

Four different methods for capturing and reproducing the sound in the transmitting room were compared in listening tests. The methods used were the wave field extrapolation (WFE), an adaptive beamformer combined with wave field synthesis (WFS), close-talk microphones combined with WFS, and direct reproduction of the sound using only two loudspeakers. The results of the listening test show the clear potential of the WFE method in creating an illusion of an acoustic window between the two rooms. It enables the creation of audio sources far beyond the used loudspeaker array while preserving satisfactory separation of the different sources.

Chapter 7

Future research and development

The topic of multichannel communications has been under investigation for several years. However, there are few examples of multichannel audio communications systems that have actually been built and tested in full scale. Therefore, the system described in this thesis represents almost pioneer work in the field and the system's performance could be improved in the future with additional research. In the following sections, a couple of interesting topics are described.

7.1 Hardware setup

It was clearly stated in section 4.3 that increasing the number of loudspeakers and microphones decreases the amount of spatial aliasing introduced in the reproduced wave field. Therefore, it also increases the perceived audio quality of the system and enlarges the sweet spot as verified in [28].

The current system uses small size loudspeakers and the loudspeaker spacing is already quite small and should produce good localization of virtual audio sources created with the array. On the other hand, the system uses a rather large spacing for the microphones and this is clearly causing coloration and inaccuracy in the reproduced wave field, as can be seen in the listening test results (Sect. 6.3). The performance could be improved by using a smaller microphone spacing and more microphones. This however increases the computational load of the system due to the WFE. In the ideal case the amount of microphones and loudspeakers would be the same, eliminating the need of the extrapolation assuming that the the microphones are at the same positions as the loudspeakers.

For home use of the acoustic opening, it would be preferred that the system components could be hidden out of sight. In the case of the microphones this is rather easy due to the small size of the elements. Ordinary electromagnetic loudspeakers can be hidden by flush mounting them into structures, but this is inconvenient, especially if the setup is not built together with the structure itself. The solution for this is to use special panel loudspeakers that can be surface mounted and painted to completely hide them from the user. An example of these elements is a distributed mode loudspeaker (DML) [2]. The acoustic characteristics of DML panels [42] and the performance of DML arrays [15] have been investigated in literature. In addition, the panels have been tested with WFS in [19] with promising results.

7.2 Signal processing

7.2.1 Improvements on WFS

The WFS processing described in section 4.2 is a 'blind' method, i.e. some wave field is reproduced with the array assuming that there is no reverberation in the space where the wave field is synthesized. In real applications this is not the case and room reverberation introduces a significant effect to the reproduced wave field. Some methods for adaptive WFS have been proposed in [23]. In general the adaptation to the room can be performed so that the wave field is measured constantly in several points in the reproduction room and the WFS process is adapted according to the measurements to produce the correct wave field. Furthermore, adaptive WFS solutions have been investigated in [49, 16]

7.2.2 Multichannel acoustic echo canceller

The implemented multichannel acoustic echo canceller has room for improvement. The current method does not take into account the cross-correlations of the different loudspeaker signals and provides only a simple adaptation process. Therefore it cannot fully take into account changes in the acoustical paths between the system's components, for example due to users movement or opened or closed doors. For future research an implementation of a frequency-domain echo canceller described in [12], taking also the cross-correlations into account, could be implemented for both increased echo attenuation and decreased computational complexity. Furthermore, other advanced time domain [3] and frequency domain [4] methods exist.

The fact that we are using WFS to create a wave field in the reproduction room gives us also possibilities for acoustic echo cancellation. By controlling the reproduced wave field so that the sound from the loudspeakers gets cancelled at the position of the microphones, we effectively implement acoustic echo cancellation. The method is described in detail in [38] and provides highly promising results.

Furthermore, additional methods for acoustic echo cancellation exist. By shifting the signals of the microphones of one side of the system in frequency, we can attenuate the

Larsen effect [45]. Frequency shifting is used usually in conjunction with some of the other methods to increase the attenuation. In addition to cancelling part of the echo by itself, it can also be used for decorrelating the signals in the system in conjunction with adaptive filters [35]. If used alone, it might produce unwanted artifacts to the reproduced sound.

7.2.3 Speech enhancement and user tracking

There was a clearly visible trend in the listening test results. The methods that were able to convey a part of the acoustics of the recording room to the other side also provided higher values for spatial naturalness. In addition, the methods were able to produce sound sources further away from the loudspeaker array than the methods that used only the original dry signals. The results suggest that the recording room acoustics and reverberation have a large influence on the perceived naturalness and the distance of the sound sources. Adding artificial reverb to the used signals could provide better results for especially the method with WFS and the dry signals. A method for blind reverberation estimation was proposed in [55]. The method could be used in conjunction with the microphones in the system to enhance the loudspeaker signals.

The methods that were able to produce high spatial naturalness in the listening tests also introduced a lot of coloration to the reproduced wave field. Different speech enhancement techniques are investigated in [41] and [31]. It could be possible to decrease the amount of noise and coloration in the reproduced wave field, but this should be done without affecting the spatial properties.

The WFE does not need any information on the location of the sounds sources, but the methods using plain WFS to render virtual sound sources need this information. In the listening test, a priori knowledge of the source positions was used. In optimal case the system would be able to track the audio sources and use the information for the WFS rendering. Several techniques used in audio source tracking have been proposed in literature (see for example [27] for angle of arrival estimation or [51] for estimation of the time differences between the microphone signals).

Chapter 8

Conclusion

In this thesis a two-way multichannel audio communication system was introduced. The aim was to create a virtual acoustic window between two rooms, providing correct spatial localization of multiple audio sources on both sides.

To motivate the use of multiple microphones and loudspeakers, current standard communication systems were reviewed starting from monophonic systems and expanding the setup to feature stereophonic recording and reproduction of sound. The monophonic systems lack the ability to convey the most part of the spatial properties of the acoustic scene in the recording space. In addition, with just a single audio channel, it is impossible to render correct spatial localization of the sources from the recording side. This also degrades the speech intelligibility during the conversation and the problem is emphasized when multiple people are participating in the conversation.

Acoustic feedback in communication systems is a known problem, but with single channel systems acoustic echo cancellation is an easy task to handle and modern adaptive filters achieve high performance.

Stereophonic recording and reproduction offers a clear improvement over monophonic system. Sound sources can easily be rendered between the two loudspeakers, but the acoustic echo cancellation problem gets more difficult. Due to the known non-uniqueness problem the acoustic echo cancellers that work for single channel systems do not work anymore with stereophonic systems. The convergence of the adaptation process slows down drastically or seizes totally. Solutions for the problem exist and the most effective ones use techniques to de-correlate the signals.

Extending communication systems to feature multichannel sound capture and reproduction increases the achievable sound quality even further. Adding multiple channels to the system also increases the complexity of the acoustic echo cancellation. Methods known from stereophonic systems extend to multichannel systems. In addition, more sophisticated methods can be used, taking into account the cross-correlations of all the loudspeaker signals and featuring frequency-domain processing for decreased computational load.

Understanding the full advantages of multichannel audio, an understanding of human spatial hearing is needed. Spatial hearing is a complex system which gives us cues of the location of heard sound events. Interaural time and level differences are mainly used in localizing sound events. Problems with spatial hearing arise when the sound is arriving from the median plane or from the cone of confusion. In these situations other cues, i.e. sight, are used for more accurate sound localization.

By using arrays of microphones and loudspeakers it becomes possible to try to capture the entire wave field and reproduce it at a different location. A method for achieving this is wave field synthesis (WFS). It enables us to create virtual audio sources with loudspeaker arrays. Extending the technique further it becomes possible to record a wave field by N microphones and extrapolate the signal to M loudspeakers. The method is called wave field extrapolation (WFE).

The theory of WFS sets strict requirements for the hardware used to create the acoustic opening. Using standard loudspeakers with monopole directivity requires the use of microphones with cardioid directivity pattern. In addition, the system components have to be similar, i.e. each loudspeaker used has to feature identical directivity and frequency response characteristics. The same holds for the microphones.

According to these requirements, components for the system build-up were selected. The components were then carefully measured to verify the required characteristics. When needed, the components were equalized to compensate for the differences. After verification of the components a symmetrical two-way system was built featuring 12 loudspeakers and 4 microphones on both sides of the system. WFE was used to extrapolate the wave field from the microphones to the loudspeakers.

To solve the acoustic feedback problem, a 48 channel acoustic echo canceller was implemented. It featured 2 x 48 = 96 static filters using pre-measured data of the acoustical paths between each loudspeaker and microphone in the system. Furthermore, to maximize the achieved echo attenuation, adaptive gains were used for each filter. The implementation provided a stable solution that made normal conversation through the window possible. The entire two-way system was running in realtime in one Pentium4 class computer.

To verify the quality of the system, a listening test was performed. In the test, WFE was compared against three other recording and reproduction methods in four different aspects. The participants were asked to evaluate the perceived spatial naturalness, the amount of coloration, the achieved distance of sound sources and the systems ability to create separated virtual sound sources. The results show that WFE offers clear potential to be used in multichannel communication systems and in creation of the acoustic opening. Especially its ability to convey a high quality spatial image of the acoustical scene from the recording side makes it a potential candidate for future research and development.

There is space for improvement with the system, especially with the multichannel echo canceller. The implemented solution uses static filters with simple adaptive gains. In future research, various, more sophisticated solutions can be experimented with, but the system described in this report represents a well performing starting point for multichannel communication systems.

Appendix A

Listening test results

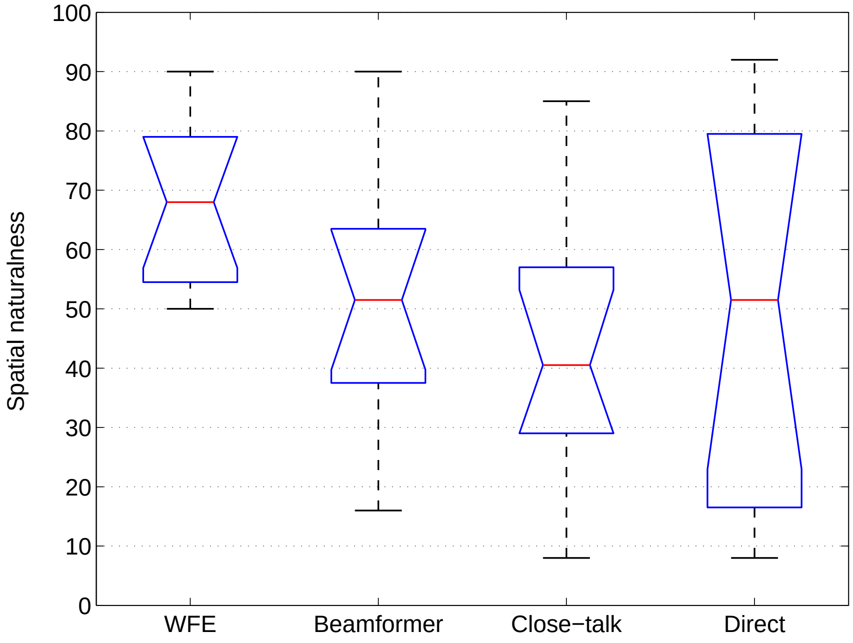


Figure A.1: *The results: dialog & spatial naturalness*

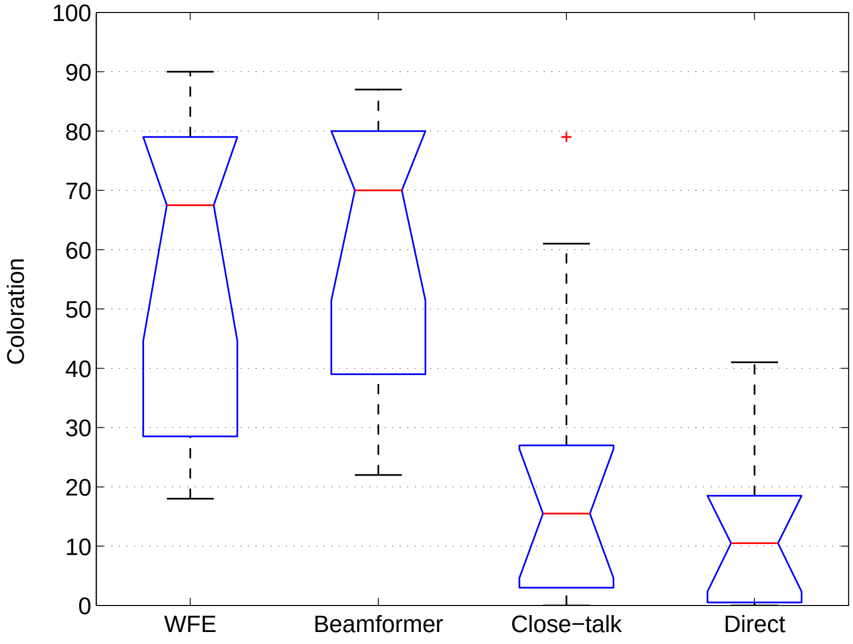


Figure A.2: *The results: dialog & coloration*

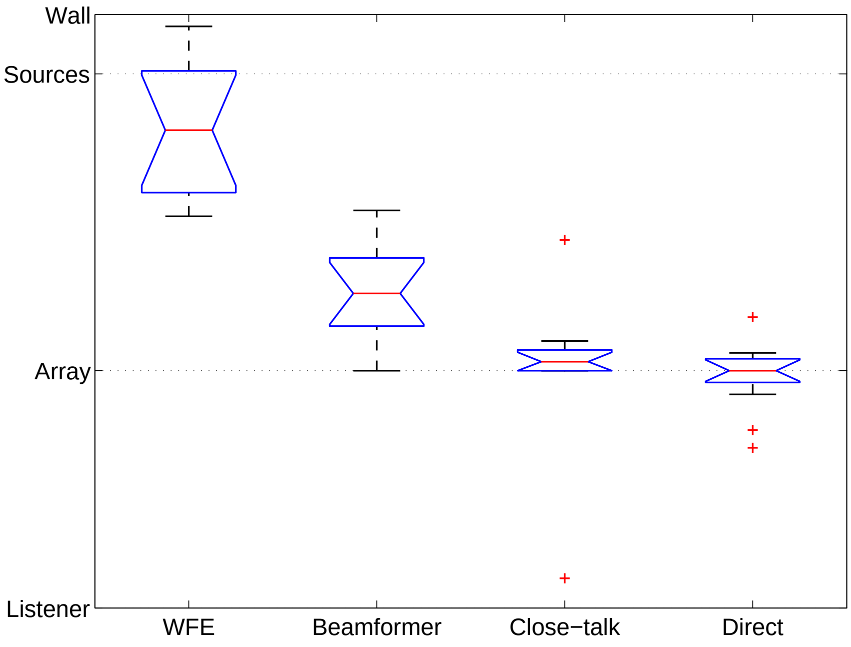


Figure A.3: *The results: dialog & distance*

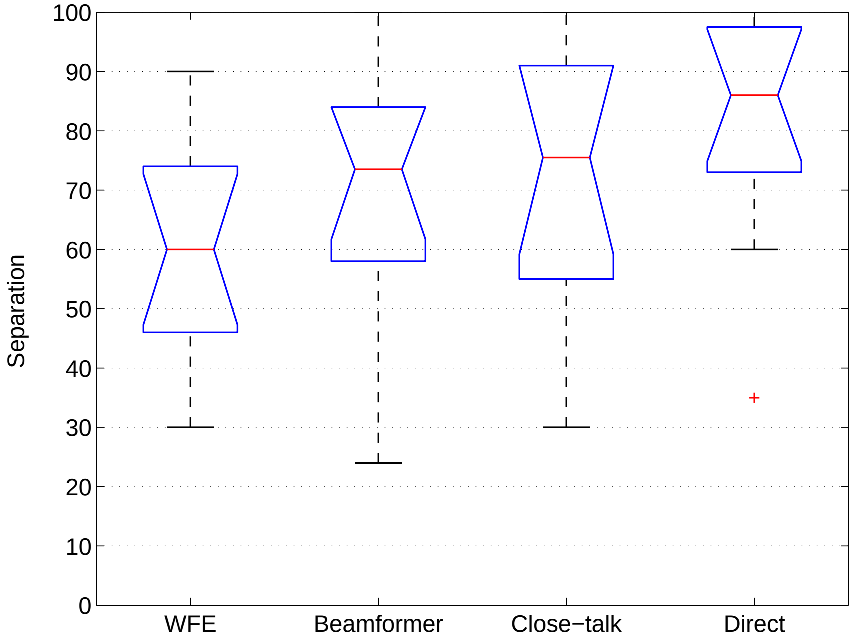


Figure A.4: *The results: dialog & separation*

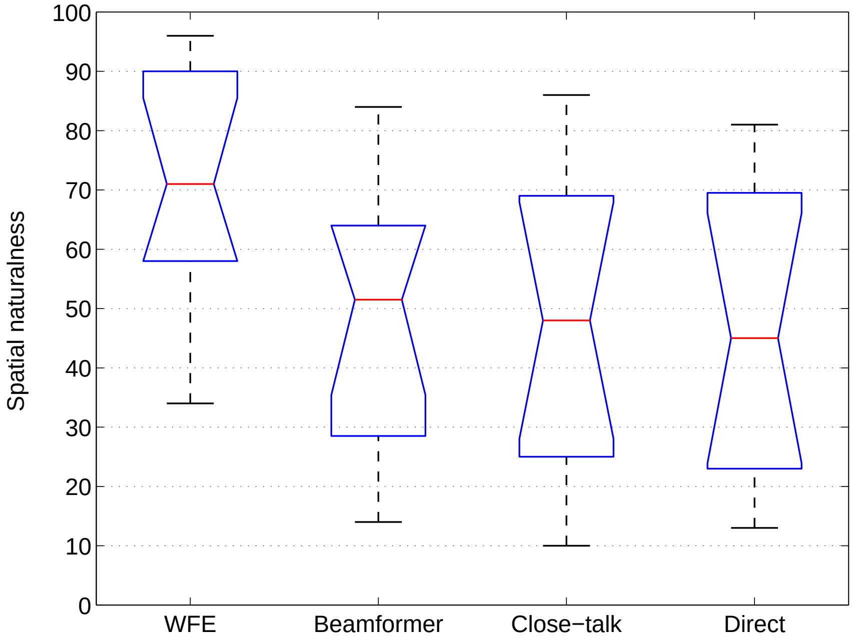


Figure A.5: *The results: overlapping dialog & spatial naturalness*

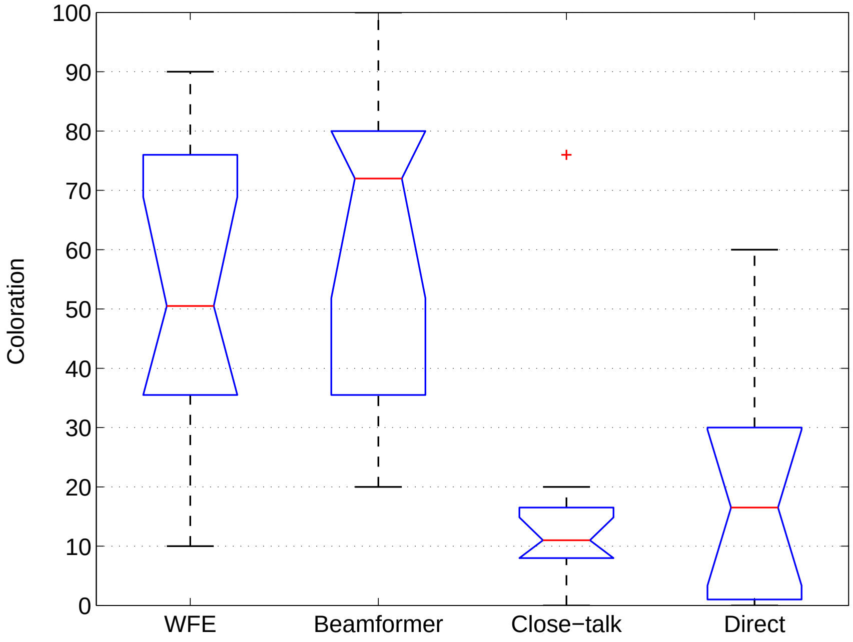


Figure A.6: *The results: overlapping dialog & coloration*

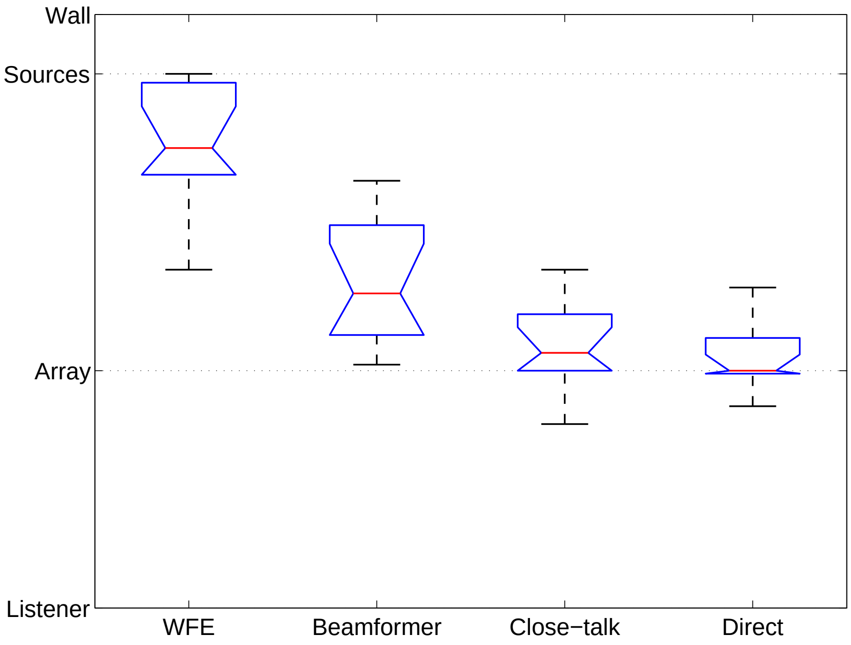


Figure A.7: *The results: overlapping dialog & distance*

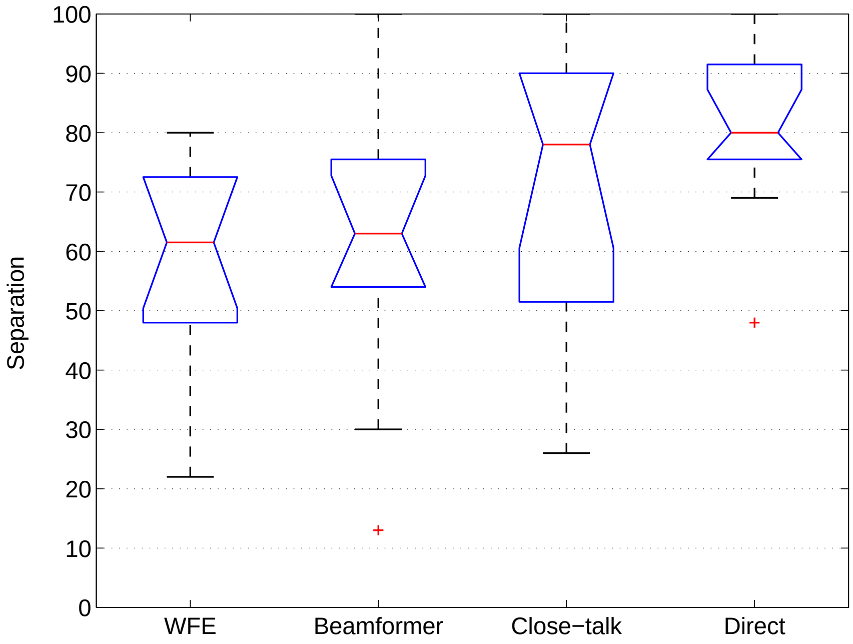


Figure A.8: *The results: overlapping dialog & separation*

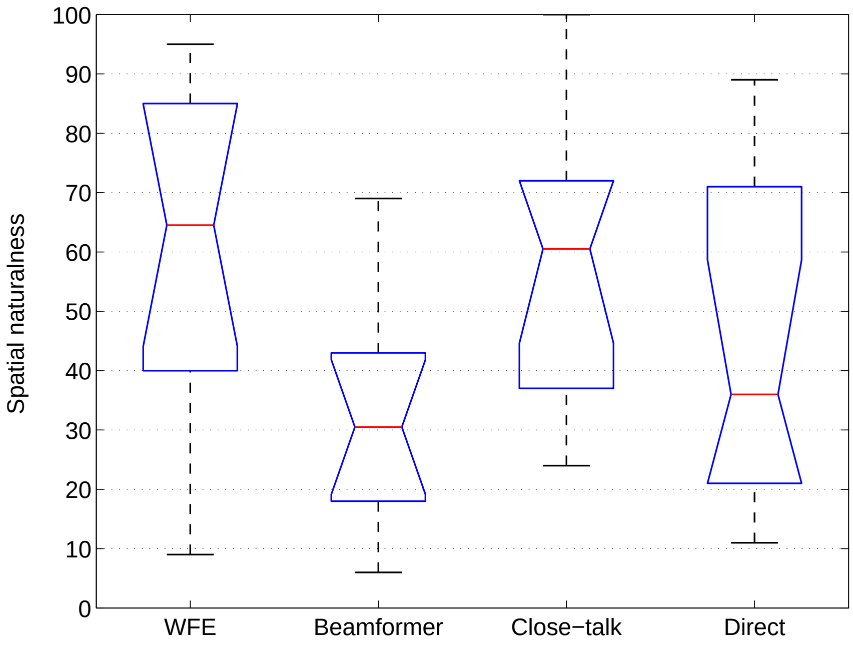


Figure A.9: *The results: instruments & spatial naturalness*

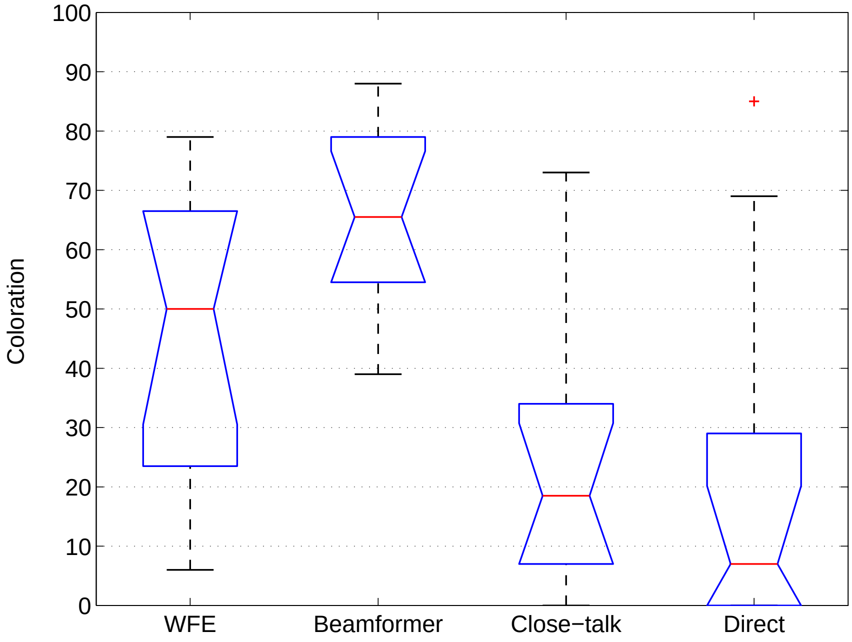


Figure A.10: *The results: instruments & coloration*

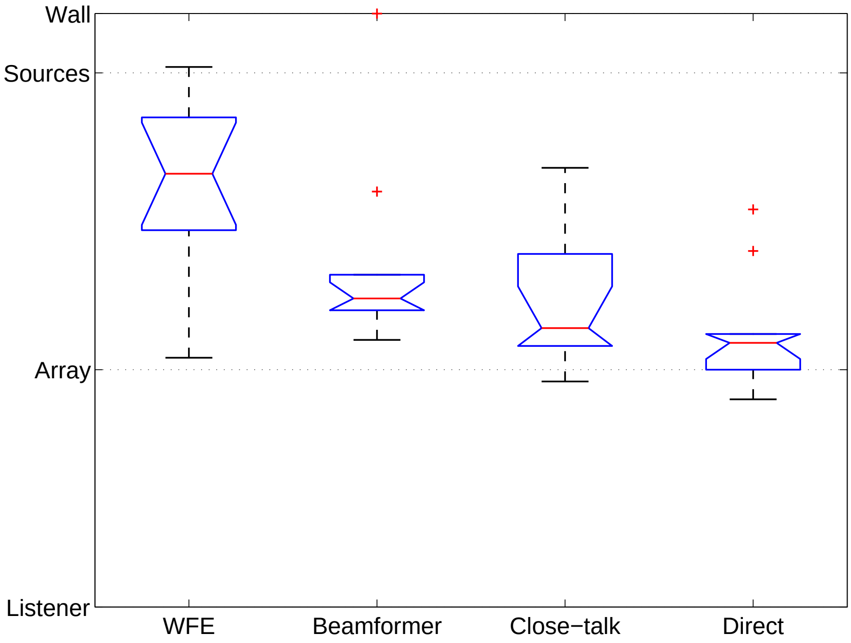


Figure A.11: *The results: instruments & distance*

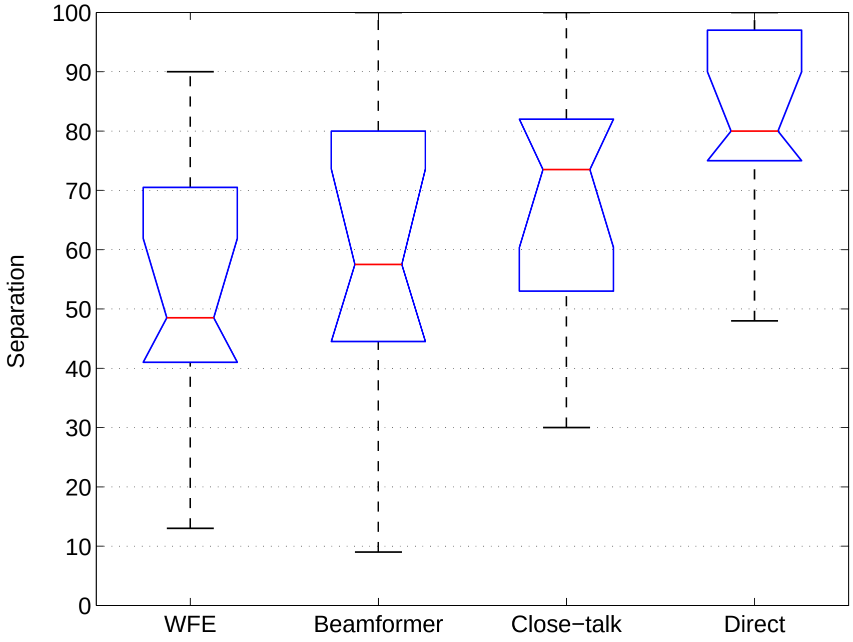


Figure A.12: *The results: instruments & separation*

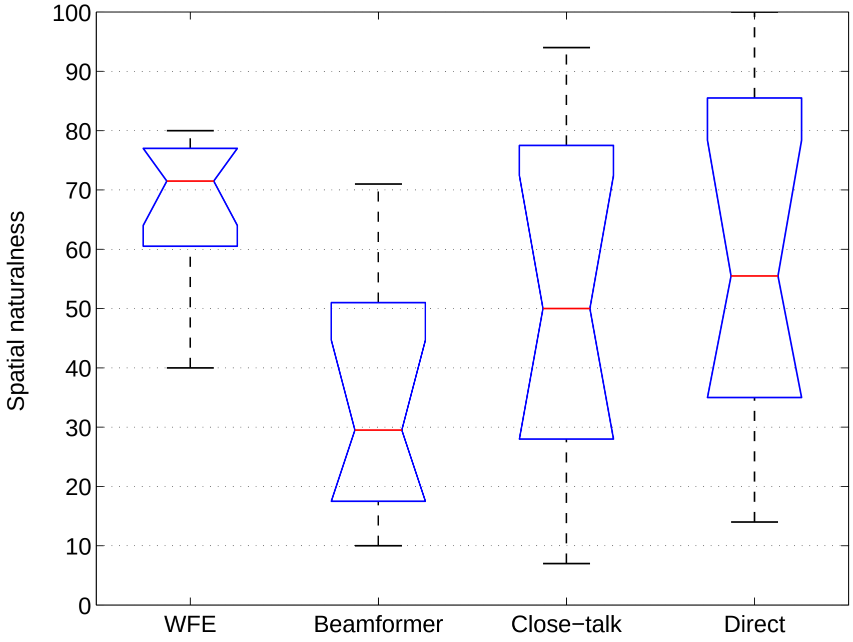


Figure A.13: *The results: singers & spatial naturalness*

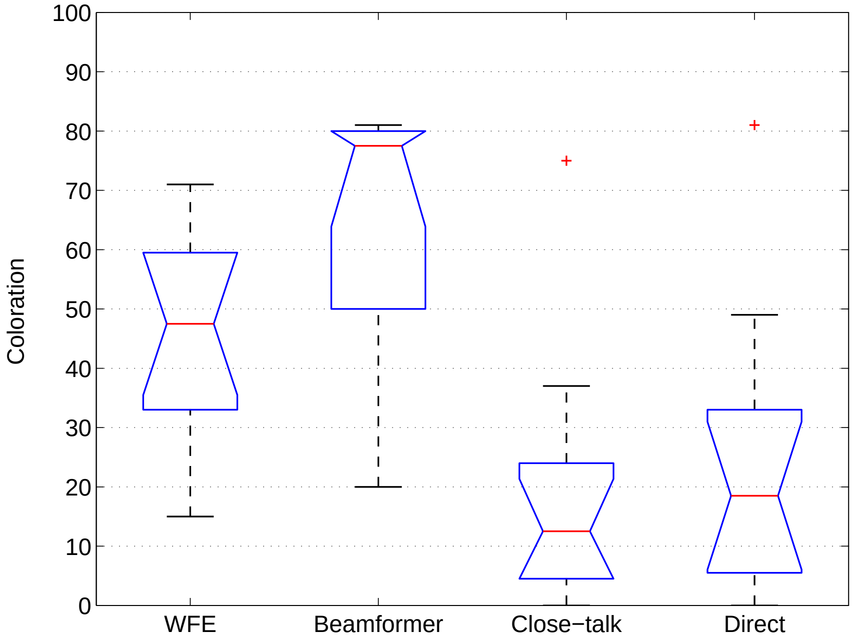


Figure A.14: *The results: singers & coloration*

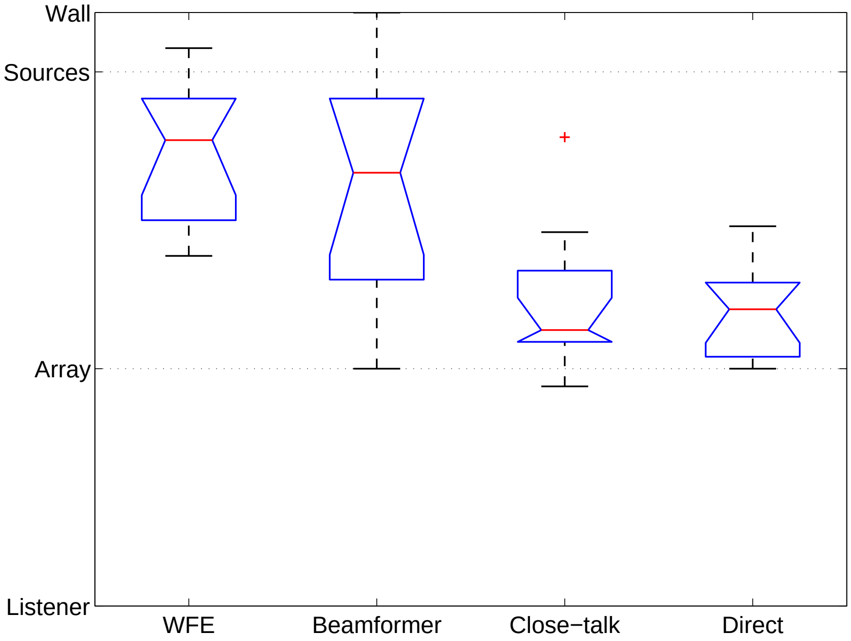


Figure A.15: *The results: singers & distance*

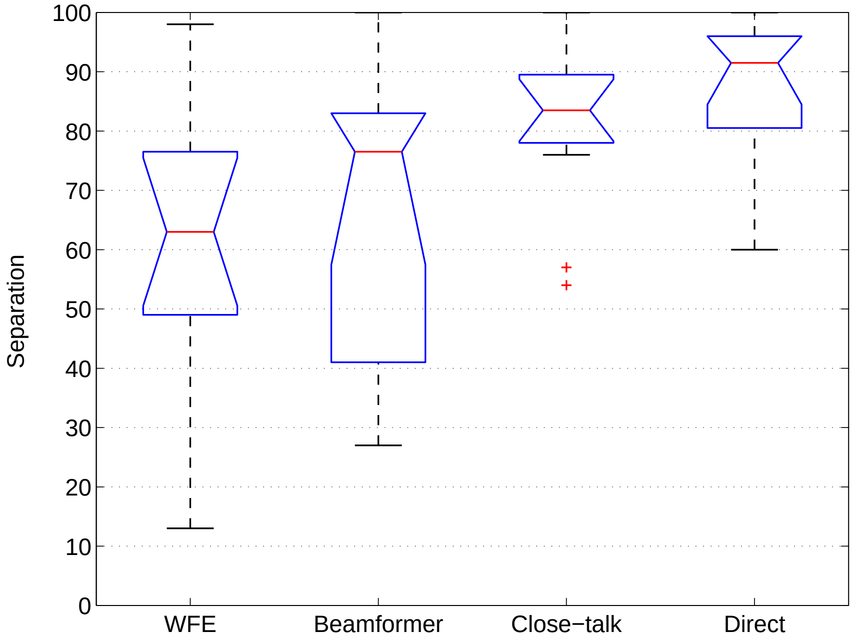


Figure A.16: *The results: singers & separation*

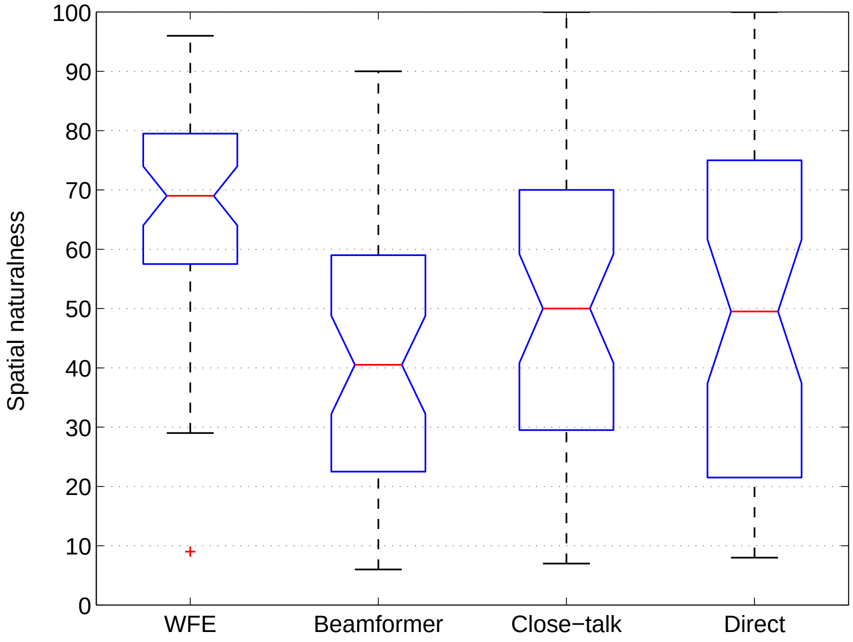


Figure A.17: *The results: all samples & spatial naturalness*

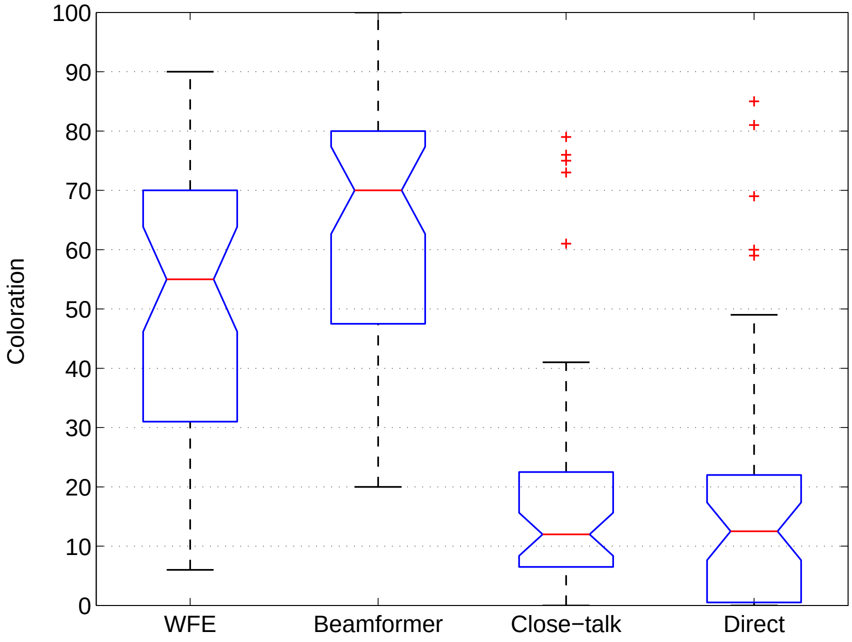


Figure A.18: *The results: all samples & coloration*

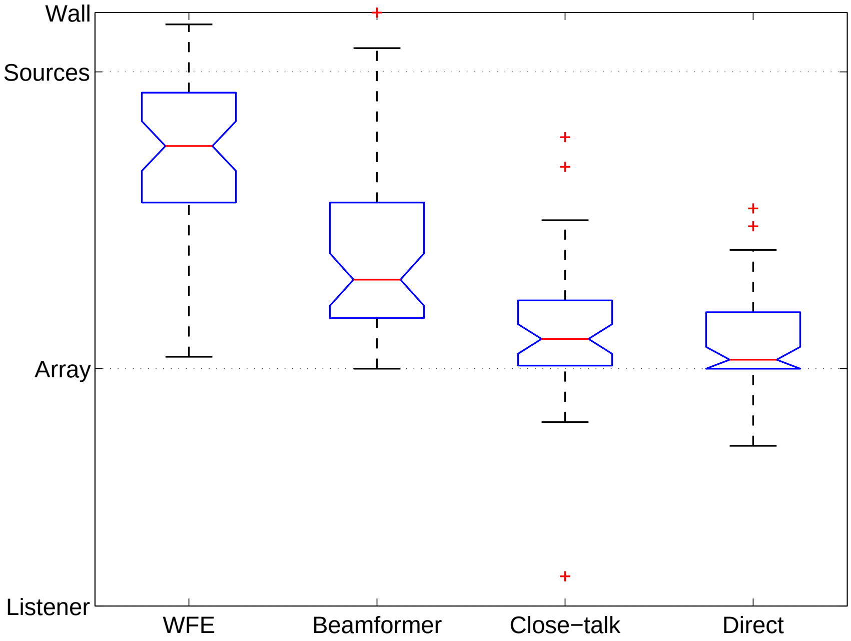


Figure A.19: *The results: all samples & distance*

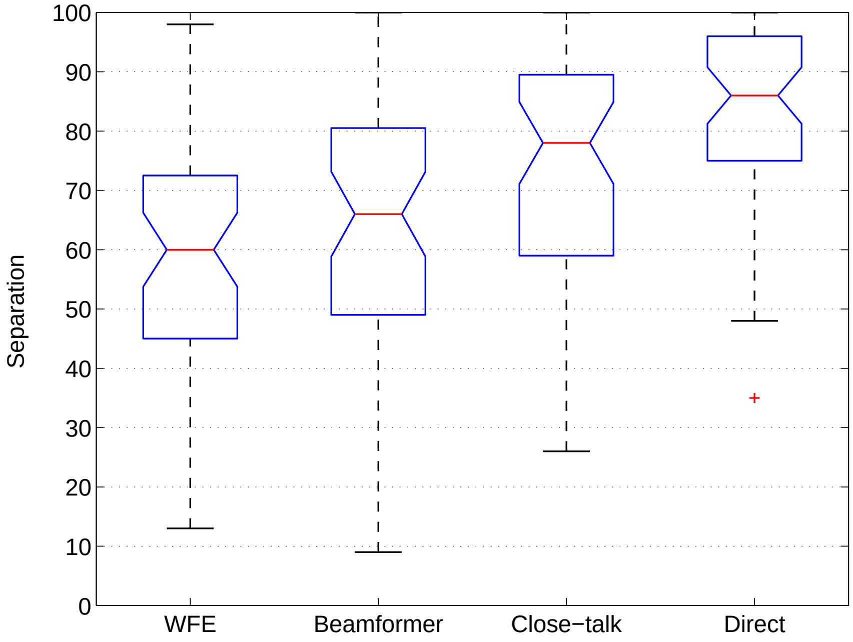
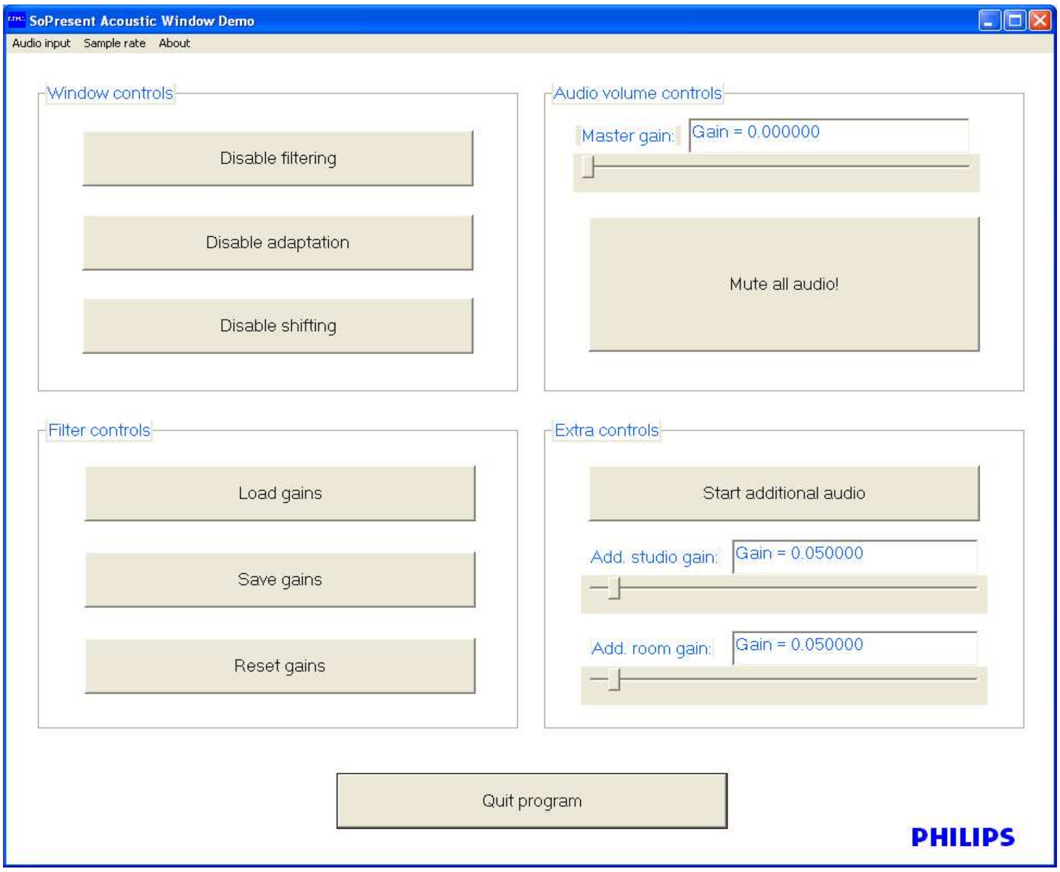


Figure A.20: *The results: all samples & separation*

Appendix B

Graphical user interfaces of the system



Master gain: Gain = 0.000000

Add. studio gain: Gain = 0.050000

Add. room gain: Gain = 0.050000

Figure B.1: *Graphical user interface of the acoustic window*

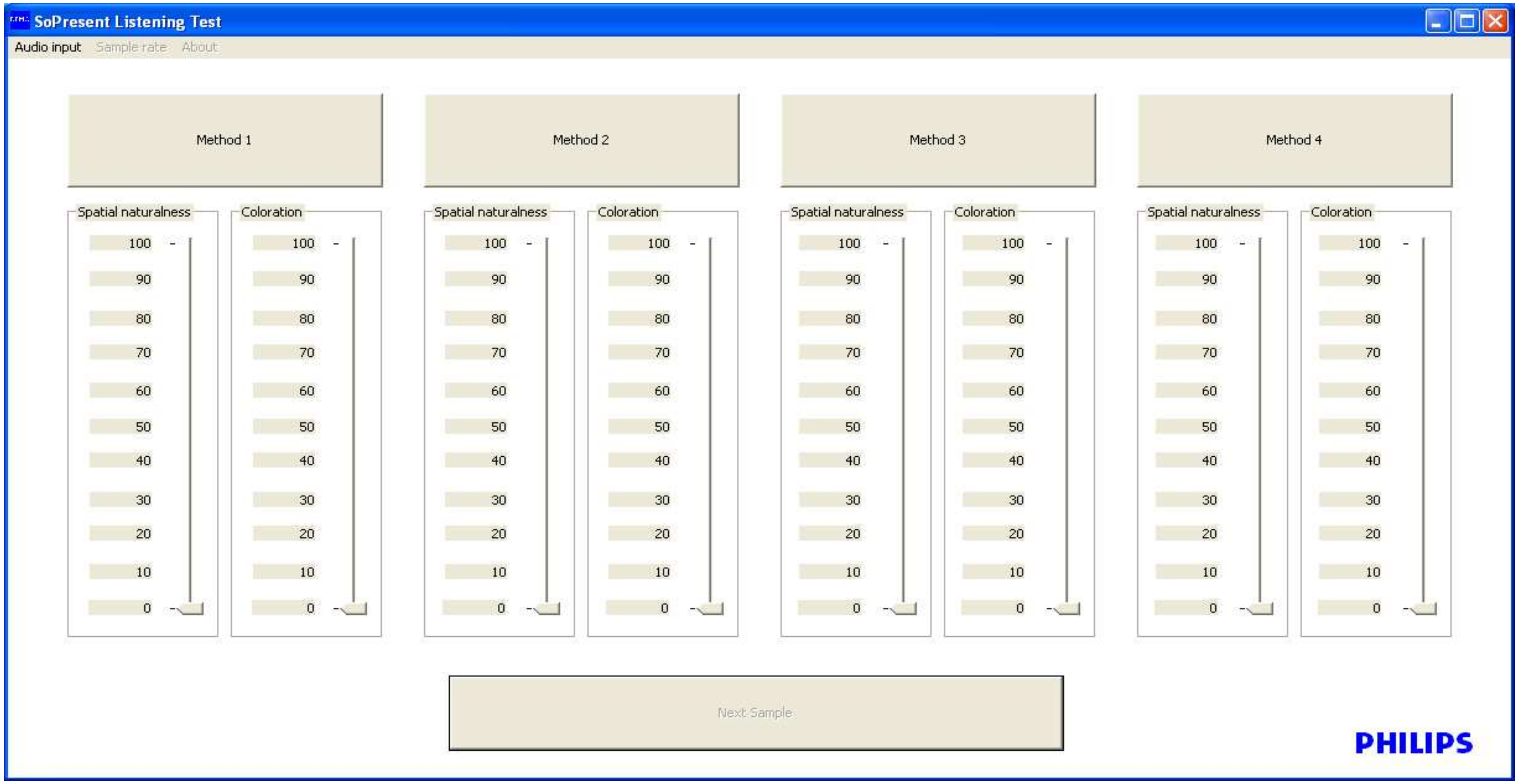


Figure B.2: *Graphical user interface of the first part of the listening test*

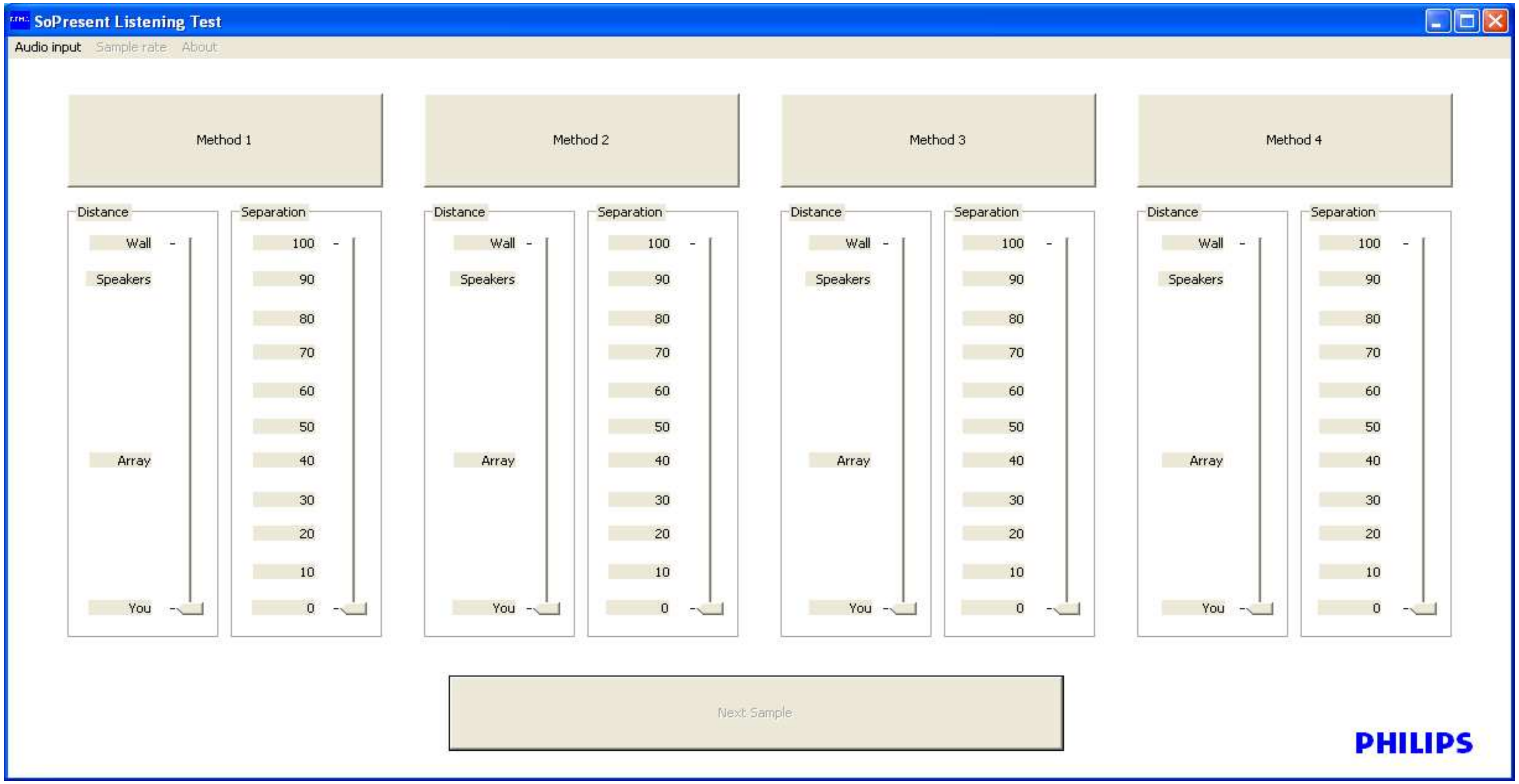


Figure B.3: *Graphical user interface of the second part of the listening test*

Bibliography

[1] S. Aoki and N. Koizumi. Expansion of listening area with good localization in audio conferencing. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 12:149–152, April 1987.

[2] G. Bank and N. Harris. The Distributed Mode Loudspeaker - Theory and Practice. In *AES UK Conference: Microphones & Loudspeakers*, The United Kingdom, March 1998.

[3] J. Benesty. Multi-Channel Sound, Acoustic Echo Cancellation, and Multi-Channel Time-Domain Adaptive Filtering. In *Acoustic Signal Processing for Telecommunications*, chapter 6, pages 101 – 120. Kluwer Academic Publ., Boston, USA, 2000.

[4] J. Benesty and D. R. Morgan. Multi-Channel Frequency Domain Adaptive Filtering. In *Acoustic Signal Processing for Telecommunications*, chapter 7, pages 121–134. Kluwer Academic Publ., Boston, USA, 2000.

[5] J. Benesty, D.R. Morgan, and M.M. Sondhi. A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation. *IEEE Transactions on Speech and Audio Processing*, 6(2):156–165, March 1998.

[6] J. Berg and F. Rumsey. Systematic evaluation of perceived spatial quality. In *AES 24th International Conference on Multichannel Audio*, Banff, Canada, June 2003.

[7] A.J. Berkhout. *Applied seismic wave theory*. Elsevier Science Publishers, 1987.

[8] A.J. Berkhout. A holographic approach to acoustic control. *Journal of the Audio Engineering Society*, 36(12):977–995, 1988.

[9] R. Botros, O. Abdel-Alim, and P. Damaske. Stereophonic speech teleconferencing. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 11:1321–1324, April 1986.

[10] S. Brix, T. Sporer, and J. Plogsties. CARROUSO - An European Approach to 3D-Audio. In *AES 110th Convention Paper 5314*, Amsterdam, The Netherlands, May 2001.

[11] H. Buchner and W. Kellerman. Improved Kalman gain computation for multichannel frequency-domain adaptive filtering and application to acoustic echo cancellation. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2:1909–1912, 2002.

[12] H. Buchner, S. Spors, W. Kellermann, and R. Rabenstein. Full-duplex communication systems using loudspeaker arrays and microphone arrays. In *IEEE International Conference on Multimedia and Expo*, pages 509 – 512, 2002.

[13] E. C. Cherry. Some Experiments upon the Recognition of Speech, with One and with Two Ears. 25(5):975–979, 1953.

[14] E. C. Cherry and W. K. Taylor. Some Further Experiments upon the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, 26(4):554 – 559, 1954.

[15] L. D. Copley, T. J. Cox, and M. R. Avis. Distributed mode loudspeaker arrays. In *AES 112th Convention Paper 5610*, Munich, Germany, May 2002.

[16] E. Corteel and R. Nicol. Listening room compensation for Wave Field Synthesis. What can be done? In *23rd AES International Conference*, Copenhagen, Denmark, 2003.

[17] K. de Boer. Stereophonic sound reproduction. In *Philips Technical Review, volume V*, pages 107 – 115, Eindhoven, The Netherlands, 1940.

[18] W. P. J. de Bruijn. *Application of Wave Field Synthesis in Videoconferencing*. PhD thesis, Delft University of Technology, The Netherlands, 2004.

[19] W. P. J. de Bruijn and M. M. Boone. Application of Wave Field Synthesis in life-size videoconferencing. In *AES 114th Convention Paper 5801*, Amsterdam, The Netherlands, March 2003.

[20] D. de Vries and M. Boone. Wave Field Synthesis and Analysis Using Array Technology. In *1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, October 1999.

[21] G. W. Elko. Superdirectional Microphone Arrays. In *Acoustic Signal Processing for Telecommunications*, chapter 10, pages 181–237. Kluwer Academic Publ., Boston, USA, 2000.

[22] G. W. Elko, E. Diethorn, and T. Gänsler. Room impulse response variation due to thermal fluctuation and its impact on acoustic echo cancellation. In *International Workshop on Acoustic Echo and Noise Control*, Kyoto, Japan, Septempber 2003.

[23] P. Gauthier and A. Berry. Adaptive wave field synthesis with independent radiation mode control for active sound field reproduction: Theory. *The Journal of the Acoustical Society of America*, 119(5):2721 – 2737, 2006.

[24] S. L. Gay and J. Benesty. An Introduction to Acoustic Echo and Noise Control. In *Acoustic Signal Processing for Telecommunications*, chapter 1, pages 1–19. Kluwer Academic Publ., Boston, USA, 2000.

[25] L. J. Griffiths and C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennas and Propagation*, AP-30(1):27–34, January 1982.

[26] T. Haapsaari, W. de Bruijn, and A. Härmä. Comparison of different sound capture and reproduction techniques in a virtual acoustic window. In *AES 122nd Convention*, Vienna, Austria, May 2007.

[27] S. Haykin. Radar array processing for angle of arrival estimation. In *Array Signal Processing*, chapter 4, pages 194–292. Prentice-Hall, Inc, New Jersey, USA, 1985.

[28] A. Härmä. Coding Principles for Virtual Acoustic Openings. In *AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, June 2002.

[29] A. Härmä. Social Presence Technology. In *Philips Technical Note TN-2006/00519*, Eindhoven, Netherlands, 2006.

[30] A. Härmä, T. Lokki, and V. Pulkki. Drawing quality maps of the sweet spot and its surroundings in multichannel reproduction and coding. In *AES 21st International Conference*, St. Petersburg, Russia, June 2002.

[31] Y. Huang, J. Benesty, and J. Chen. Separation and dereverberation of speech signals with multiple microphones. In *Speech Enhancement*, chapter 12, pages 271–298. Springer-Verlag, 1985.

[32] E. Hulsebos. *Auralization using Wave Field Synthesis*. PhD thesis, Delft University of Technology, The Netherlands, 2004.

[33] J. Huopaniemi. *Virtual Acoustics and 3-D Sound in Multimedia Signal Processing*. PhD thesis, Helsinki University of Technology, Finland, 1999.

[34] C. Huygens. *Traité de la lumière - ou sont expliquées les causes de ce qui arrive dans la reflexion, & dans la refraction. Et particulièrement dans l'étrange refraction du cristal d'Islande*. Pierre van der Aa, Leiden, 1690.

[35] S. Kamerling, K. Janse, and F. van der Meulen. A new way of acoustic feedback suppression. In *AES 104th Convention*, Amsterdam, The Netherlands, May 1998.

[36] M. Karjalainen. *Kommunikaatioakustiikka*. Espoo, Finland, 2000.

[37] B. Klehs and T. Sporer. Wave field synthesis in the real world: Part 1 - in the living room. In *AES 114th Convention*, Amsterdam, The Netherlands, March 2003.

[38] S. Miyabe, Y. Hinamoto, H. Saruwatari, K. Shikano, and Y. Tatekura. Interface for barge-in free spoken dialogue system based on sound field reproduction and microphone array. *EURASIP Journal on Advances in Signal Processing*, 2007.

[39] H. Nomura and M. Tohyama. Loudspeaker Arrays for Improving Speech Intelligibility in a Reverberant Space. *Journal of the Audio Engineering Society*, 39(5):338 – 343, 1991.

[40] H. Nyquist. Certain topics in telegraph transmission theory. *Proceedings of the IEEE*, 90(2):280–305, 2002.

[41] D. O'Shaughnessy. *Speech Communication: Human and Machine*. Addison Wesley, 1987.

[42] E. Prokofieva, K. V. Horoshenkov, and N. Harris. Intensity measurements of the acoustic emission from a dml panel. In *AES 112th Convention Paper 5609*, Munich, Germany, May 2002.

[43] J. W. S. Rayleigh. *The Theory of Sound II*. Dover Publications, Inc. (Reprint 1945), 1878.

[44] T.D. Rossing, F.R. Moore, and P.A. Wheeler. *The Science of Sound*. Addison Wesley, 2001.

[45] J. Scheuing and B. Yang. Frequency shifting for acoustic feedback elimination. In *European DSP Education and Research Symposium (EDERS)*, München, Germany, April 2006.

[46] M. R. Schroeder. Improvement of Acoustic-Feedback Stability by Frequency Shifting. *The Journal of the Acoustical Society of America*, 36(9):1718 – 1724, 1964.

[47] W. B. Snow. Auditory Perspective. *Bell Laboratories Record*, 12(7), March 1934.

[48] M. M. Sondhi, D. R. Morgan, and J. L. Hall. Stereophonic acoustic echo cancellation-an overview of the fundamental problem. *Signal Processing Letters, IEEE*, 2(8):148–151, August 1995.

[49] S. Spors, H. Buchner, and R. Rabenstein. Efficient active listening room compensation for Wave Field Synthesis. In *AES 116th Convention Paper 6119*, Berlin, Germany, 2004.

[50] E. Start. *Direct sound enhancement by Wave Field Synthesis*. PhD thesis, Delft University of Technology, The Netherlands, 1997.

[51] S. Tervo. Aikaeron estimointimenetelmien suorituskyky reaalitilanteessa. Master's thesis, Tampere University of Technology, Finland, 2006.

[52] Wikipedia the free encyclopedia. *Audio feedback*. February 2007.
http://en.wikipedia.org/wiki/Audio_feedback.

[53] Wikipedia the free encyclopedia. *Dirac delta function*. March 2007.
http://en.wikipedia.org/wiki/Dirac_delta_function.

[54] E. Verheijen. *Sound reproduction by Wave Field Synthesis*. PhD thesis, Delft University of Technology, The Netherlands, 1997.

[55] S. Vesa. Estimation of reverberation time from binaural signals without using controlled excitation. Master's thesis, Helsinki University of Technology, Finland, 2004.

[56] P. Vogel. *Application of Wave Field Synthesis in room acoustics*. PhD thesis, Delft University of Technology, The Netherlands, 1993.

[57] C. P. A Wapenaar. Reciprocity theorems for two-way and one-way wave vectors: a comparison. *The Journal of the Acoustical Society of America*, 100(6):3508 – 3518, 1998.