Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

A systematic study of binaural reproduction systems through loudspeakers

A multiple stereo-dipole approach

Lacouture Parodi, Yesenia

Publication date:
2010

Document Version
Accepted author manuscript, peer reviewed version

Link to publication from Aalborg University

Citation for published version (APA):
Lacouture Parodi, Y. (2010). *A systematic study of binaural reproduction systems through loudspeakers: A multiple stereo-dipole approach*. Section of Acoustics, Aalborg University.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

? Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
? You may not further distribute the material or use it for any profit-making activity or commercial gain
? You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: November 29, 2020

A Systematic Study of Binaural Reproduction Systems Through Loudspeakers:

A Multiple Stereo-Dipole Approach

PhD Thesis

Yesenia Lacouture Parodi

November 2010

Section of Acoustics
Department of Electronic Systems
Aalborg University, Denmark

A Systematic Study of Binaural Reproduction Systems Through Loudspeakers:
A Multiple Stereo-Dipole Approach

ISBN: 978-87-92328-47-2

Copyright ©2010 by Yesenia Lacouture Parodi
All rights reserved. Address correspondence to:
Section of Acoustics, Department of Electronic Systems
Aalborg University, Frederik Bajers vej 7-B5, DK-9220 Aalborg
Phone: +45 9940 8710, E-mail: acoustics@es.aau.dk

Ph.D. thesis defended on November 12, 2010
Ph.D. degree awarded on December 15, 2010
Aalborg University, Denmark

Assessment Committee:
Professor Dorte Hammershøi, Aalborg University, Denmark (chairman)
Professor Philip Arthur Nelson, University of Southampton, UK
Professor John N. Mourjopoulos, University of Patras, Greece

Supervisors:
Professor Henrik Møller, Aalborg University, Denmark
Associate Professor Per Rubak, Aalborg University, Denmark

'We shall not cease from exploration
and the end of all our exploring
will be to arrive where we started
and know the place for the first time'

T.S. Eliot, 1942

PREFACE

This PhD thesis is submitted to the Faculty of Engineering and Science at Aalborg University, Denmark, in partial fulfillment of the Ph.D. study program. The work of this thesis has been carried out at at the Department of Electronic Systems, Section of Acoustics at Aalborg University during the period September 2007 to July 2010. The project has been supervised initially by Henrik Møller and later by Per Rubak.

I would like to thank Henrik Møller for his ideas that gave the foundation and starting point to the research done. I would also like to extend my gratitude to Per Rubak for his support and guidance throughout the research process and his valuable inputs to the work done.

The outcome of this Thesis would not have been possible without the help of our technician Claus Vestergaard Skipper for whom no experimental setup was an impossible task. Further thanks go to my former office mates Daniela Toledo and Pablo Hoffmann who were a great support during the first years of the research and made the working atmosphere more amenable. I would like to thank all the people at the section of Acoustics for supporting my work in one way or another and the subjects that participated in the listening experiments.

Special thanks go to Professor Mingsian Bai and his working group at the National Chiao Tung University in HsinChu, Taiwan, for their contributions to my work and for helping to make my stay in Taiwan an invaluable experience both from the professional and personal point of view. Finally, I would like to thank my beloved husband Martin for his unconditional support and love, his valuable help proofreading my manuscripts and participating as subject in my pilot experiments. Last but not least, I would like to thank my family who contribute to an integral life and that in spite of the long distance has never failed to give me all the support and love needed to keep the good work. This thesis is dedicated to them.

Yesenia Lacouture Parodi
Aalborg, July 2010

SUMMARY

With binaural technology it is possible to simulate a virtual environment where the listener perceives an acoustic source located at a position where no physical source exists. Based on the assumption that our sound perception is controlled by the sound pressures at the ears, one can simulate any virtual source if complete control over the sound pressures at the left and right ear is achieved. Normally, headphones are used to reproduce the binaural signals due to the perfect channel separation that they provide. However, reproduction through headphones might not be practical or comfortable for some applications and therefore loudspeakers are preferred.

When reproducing binaural signals through loudspeakers, the signals that are to be heard only by the right ear are also heard by the left ear and vice versa. This is called crosstalk and it is possible to reduce it - even eliminate it - by adding proper filters into the reproduction chain. These techniques are known as crosstalk cancellation. In general, crosstalk cancellation lacks robustness with respect to several parameters in the physical setup, such as listener's position and head movements.

This PhD study investigated the acoustical and psycho-acoustical characteristics of binaural reproduction systems through loudspeakers. The study was mainly focused on the characteristics of the stereo-dipoles when placed in different positions, e.g. in front of, above or in the back of the head. The main objective of the study was to set the foundations for the design of a binaural reproduction system that is robust to head rotations and reduces front-back confusions.

In the first part of the study, different crosstalk cancellation techniques were investigated and their performance with respect to loudspeaker configurations was evaluated. Different parameters such as regularization, filter length and bandwidth were varied and the performance of the methods was evaluated for each iteration. It was found that the techniques based on least square approximations yield the best performance. Additionally, the use of more than two channels did not exhibit an advantage over the conventional two-channel configurations. The choice of the regularization value showed not to be critical, but the optimum value proved to be a compromise between performance and filter gain.

The second part of the study investigated the robustness of different loudspeaker configurations placed at different elevation angles with respect to head misalignments. The previously studied crosstalk cancellation techniques were also evaluated. The analysis was done looking not only at changes in the magnitude ratios of the crosstalk to the direct signals but also at the changes in the interaural time delays of the binaural signals. Results show that closely spaced loudspeakers are more robust to lateral displacements than wider span angles. Additionally, the sweet spot as a function of head rotation increases systematically when the loudspeakers are placed at elevated positions.

In another experiment, the audibility of crosstalk in binaural signals was investigated. It was shown that audible crosstalk results in lateralization of the virtual image and that the maximum acceptable channel separation is much below the assumed values encountered in the literature. It is also argued that at asymmetric listener positions, the delay differences between channels can also result in lateralization effects when wide span angles are used.

Finally, based on the findings of the investigations carried out, a binaural reproduction system that uses three-stereo dipoles is proposed: one pair in front, one pair behind and one pair above the listeners. The

results of preliminary evaluations indicate that a reduction of front-back confusions is obtained with the proposed system. Yet, there are still some pending issues regarding coloration differences between loudspeaker pairs and lack of dynamic cues rendered by the upper loudspeakers.

RESUMÉ

Med binaural teknologi er det muligt at simulere et virtuelt miljø, hvor tilhøreren opfatter en akustisk kilde lokaliseret et sted, hvor der ingen fysisk kilde er. Baseret på den antagelse, at vores lydopfattelse bliver styret af lydtrykkene ved ørerne, kan man simulere enhver virtuel kilde, hvis man har fuldstændig kontrol med lydtrykkene ved venstre og højre øre. Normalt bruges hovedtelefoner til at reproducere de binaurale signaler, fordi de har perfekt kanaladskillelse. Men for nogle anvendelser er reproduktion gennem hovedtelefoner måske ikke praktisk eller komfortabelt, hvorfor højtalere kan foretrækkes.

Når man reproducerer binaurale signaler gennem højtalere, høres de signaler, som er beregnet for det højre øre også af det venstre øre og vice versa. Dette kaldes krydstale. Det er muligt at benytte højtalere, hvis man tilføjer nogle filtre til reproduktionskæden. Disse teknikker kaldes krydstaleundertrykkelse. Generelt er krydstaleundertrykkelse følsomt overfor en lang række faktorer så som lytterens position og hovedbevægelser.

Dette ph.d.-studie undersøgte de akustiske og psyko-akustiske egenskaber af binaurale reproduktionssystemer gennem højtalere. Undersøgelsen blev primært fokuseret på de særlige egenskaber ved stereo- dipoler (to tætliggende højtalere), når de placeres i forskellige positioner, fx foran, over eller på bagved hovedet. Hovedformålet med undersøgelsen var at skabe et grundlag for udformningen af et binauralt reproduktionssystem, der er robust over for hoved rotationer og reducerer front-back forvekslinger.

I den første del af undersøgelsen blev forskellige krydstale undertrykkelsesteknikker undersøgt og deres effektivitet med hensyn til højttaler konfigurationer blev evalueret. Forskellige parametre så som regularisering, filter længde og båndbredde blev varieret og performance af de forskellige signalbehandligsmetoder blev evalueret for en lang række forskellige højtalerkonfigurationer. Det blev konstateret, at teknikker baseret på least-squares tilnærmelser giver de bedste resultater. Desuden har anvendelsen af mere end to kanaler ikke udvist fordele i forhold til den konventionelle to-kanals konfiguration. Valget af regulariseringsværdi har vist sig at være ukritisk, men den optimale værdi viste sig at være et kompromis mellem ydelse og filterforstærkning.

Den anden del af projektet undersøgte robusthed (i forhold til hoved forskydninger) af forskellige højttaler konfigurationer placeret i forskellige elevationsvinkler. De tidligere studerede krydstale undertrykkelsesteknikker blev også evalueret. Analysen blev gennemført ikke kun med hensyn til interaurale niveauforskelle, men også med hensyn til ændringer i interaurale tidsforskydninger (ITD) i de binaurale signaler. Resultaterne viser, at tætliggende højttalere (stereo-dipoler) er mere robuste over for lateral forskydninger end når brede spændvidde vinkler anvendes (større afstang mellem højtalerne). Derudover bliver sweet-spot som funktion af hovedrotation øget systematisk, når højttalerne er placeret ved forhøjede positioner.

I et andet eksperiment blev hørbarheden af krydstale i binaurale signaler undersøgt. Det blev påvist, at hørlig krydstale resulterer i lateralization af den virtuelle lydkilde og at den nødvendige kanalseparation (ca. 15-20 dB) er væsentlig større end den værdi (12 dB) der hidtil er anført i litteraturen. De udførte lytteforsøg viste også , at ved asymmetriske lyttepositioner, kan forsinkelsen mellem kanalerne også resultere i lateralizations virkninger, når brede spændvidde vinkler anvendes.

Baseret på de opnåede forskningsresulter blev et binauralt reproduktionssystem med tre stereo-dipoler foreslået: et par i front, et par bagved og et par over lytteren. Resultaterne af de foreløbige evalueringer viser, at en reduktion af front-back forvekslinger opnås med det foreslåede system. Alligevel er der stadig nogle uafklarede spørgsmål vedrørende farvning (tonebalance) af lydreproduktionen fra de forskelle højttalerpar og mangel på dynamiske features leveret af det øverste højtalerpar.

CONTENTS

1 OVERVIEW OF THE THESIS

1.1 Introduction

The idea behind binaural technology is to be able to simulate a virtual environment where the listener perceives an acoustic source located at a position where no physical source exists. Based on the assumption that our sound perception is controlled by the sound pressures at the ears, one can simulate any virtual source if complete control over the sound pressures at the left and right ears is achieved. Normally, headphones are used to reproduce binaural signals due to the perfect channel separation that they provide. However, reproduction through headphones might not be practical or comfortable and in some cases can even reduce the virtual reality perception. Additionally, if head rotations are to be allowed, a head-tracker should be included so that the location of the source relative to the listener is updated properly. This can be an expensive solution and it requires extra computational power. In contrast, when reproducing binaural signals through loudspeakers the virtual images are easily externalized and an automatic update could be achieved when the head is turned within some angle constraints.

One of the biggest challenges of reproducing binaural signals through loudspeakers is to be able to achieve complete control over the sound pressures at the ears. This is due to the fact that the signal that is to be heard on the left ear is also heard on the right ear and vice-versa. This problem is known as crosstalk and can be alleviated by introducing proper filters into the reproduction chain. Figure 1.1 shows a simplified diagram of a crosstalk cancellation system with n loudspeakers.

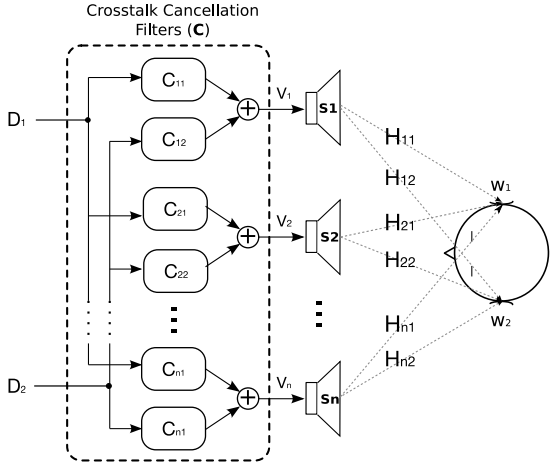


Figure 1.1: *Acoustic paths from loudspeakers to the ears. The functions H_{ij} represents the transfer functions between loudspeakers and the ears. The signals D_j are the desired binaural signal to be reproduced and the signals W_j are the signals which are reproduced at the ears.*

In theory, perfect crosstalk cancellation is achieved when the system is capable of reproducing sound in one ear while nothing is heard in the other (i.e. $W_j = D_j$). From the figure we have that

$$\mathbf{w} = \mathbf{H} \cdot \mathbf{C} \cdot \mathbf{d} \tag{1.1}$$

where **w** is a vector that contains the signals that reach the ears, **H** is the matrix containing the transfer functions from the loudspeakers to the ears - known as the plant matrix -, and **C** is a matrix containing the crosstalk cancelation filters.[1]

Even though a number of techniques to implement crosstalk cancelers have been developed in the past decades, there are still some limitations such as inversion constraints, robustness to head misalignments, front-back confusions, among others.

One of the first attempts to recreate a virtual acoustic environment by using loudspeakers was done by Atal and Schroeder in the late 60's [1, 28]. They proposed a system for reproducing virtual acoustic images by using a simple stereo setup. The filters used were derived from mannequin measurements and the system was built so it could simulate concert halls, but it was constricted to anechoic environment and head rotations no greater than $\pm 10^\circ$.

Damaske tried to simulate the diffractions caused by the head and in that way compensate for the crosstalk [10]. By placing two loudspeakers in a stereo arrangement and including a delay line, an attenuator and an inverter on the right channel, the magnitude and phase of what he called the "90° filters" was measured. Results show a rather accurate localization performance of the subjects when the 90° filters were used. Especially front-back confusions did not seem to be a problem. Despite of the cumbersome approximation of the head diffractions, it is claimed that "under optimal listening conditions, the system allows a true reproduction of all directional information".

Møller presented in 1989 mathematical descriptions of the acoustical paths and the ideal crosstalk canceler [21]. Results of subjective experiments indicated that when using a pair of loudspeakers, a better localization is achieved in the frontal region in comparison with headphones.

Cooper et. al reported improvements of the Atal-Schroeder scheme [8, 9]. They limited the filters at high frequencies and introduced additional equalization filters based on a spherical head model. They claimed that by limiting the crosstalk canceler to frequencies below 10 kHz a more robust system concerning head movements and inter-individual differences is achieved. Additionally, it is stated that equalization of the diffractions of the typical stereo setup ($\pm 30^\circ$) can be approximated with the diffractions obtained at 0° and similar results should be obtained.

Some studies have proposed different span angles and loudspeaker arrangements to improve the performance and widen the sweet-spot [2, 5, 7, 13, 15, 29]. The sweet-spot is usually defined as the area in which the amount of head movements is within the maximum allowed such that the introduced errors are negligible.

Kirkeby et al. proposed a new loudspeaker arrangement in order to overcome most of the drawbacks of the previous art systems [15, 17]. The arrangement consists of two closely spaced loudspeakers, better known as the "stereo-dipole". Comparisons with different spans showed that the controlled area obtained with the stereo dipole is much larger. However, it was also found that for small span angles, low frequencies are to be boosted. Nevertheless, due to its practical applications in home entertainment applications e.g. virtual imaging through portable computers or mobile phones, the stereo-dipoles have kept attracting the attention of many researchers and their development is still of high relevance [4, 14, 25, 31].

[1] In principle the transfer functions contain not only the acoustical path from the loudspeakers to the ears, but also the acoustical characteristics of the loudspeakers and the reproduction chain.

The sweet spot size of the stereo-dipole at off-axis asymmetric listener's positions was investigated by Rose et al. [27]. By means of computer simulations, the sound field generated by the loudspeakers was analyzed under free field conditions with the rigid sphere model and HRTFs from a dummy head. Additionally, subjective experiments with band-passed noise between 300 Hz and 3 kHz in an anechoic chamber were carried out. They concluded that the stereo-dipole is robust to lateral head translations at asymmetric listener locations that are offset up to 25 cm from the inter-source axis. However, it is pointed out that the head shadowing effect decreases the sweet spot size at low frequencies.

Bai et al. also analyzed the listening angle with respect to performance and robustness to lateral head movements [2]. In their study, crosstalk cancelers for different spans were simulated with the HRTF and the free field model. Channel separation, filter gains and condition numbers versus frequency, angle and head displacement were evaluated objectively and subjectively. Results suggest that small spans lead to a relatively large sweet spot due to the small changes on the time-of-arrival with head displacements. Nevertheless, they suffer from poor conditioning, high gain and poor performance at low frequencies. A span angle of $\pm 60^{\circ}$ is recommended as optimum.

Ward and Elko analyzed the effect of loudspeaker position on the robustness of crosstalk cancellation [32]. They suggested that the loudspeaker span should be inversely proportional to the operating frequency in order to obtain a numerically robust crosstalk canceler. Based on these results Takeuchi and Nelson proposed the so called "optimal source distribution" system which consists on multiple loudspeaker pairs covering different frequency bands each [29].

In order to increase robustness, Yang et al. proposed to introduce a third loudspeaker right in front of the listener [20]. They simulated, assuming free field conditions, a span of $\pm 15^{\circ}$ for the side loudspeakers and a crosstalk canceler band limited between 300 Hz and 6 kHz. Their results showed that the three speaker configuration is more robust with respect to head movements than the two speaker configuration. It is also noted that the robust region increases when the span angle is decreased, but at low frequencies the robustness is degraded. However, they concluded that a span of $\pm 15^{\circ}$ is a good compromise between robustness and working frequency and gives the optimum performance.

Adopting a more mathematical approach, Nelson and Rose made an analysis of the connection between the condition number of the inversion matrix and the performance of the crosstalk canceler [23]. The significance of the condition number was evaluated regarding geometry and they concluded that the best conditioned spans are the stereo-dipole and the "optimal source distribution" system proposed by Takeuchi et. al [29].

Other studies focused on developing different methods to design crosstalk cancellation filters. For example, in [11] a crosstalk canceler based on minimum-phase approximation is presented. This system is modeled in terms of the Interaural Transfer Functions (ITF) which are calculated as the ratio between the minimum-phase components of the ipsilateral and contralateral transfer functions. The excess-phase components are approximated with a frequency independent Interaural Time Delay (ITD). A more elaborated algorithm is described by Kirkeby et al. in [18], where the crosstalk cancellation filters are obtained through a least square method based on the fast fourier transform (FFT). Another least square approach is proposed in [16], but in the time domain. In general, methods based on least square approximation are the most commonly used and a number of approaches have been introduced in the last decades [3, 24, 31].

Bai et al. proposed the implementation of a crosstalk cancelation system based on subband filtering [3]. Instead of using a conventional stereo setup, they proposed a crosstalk canceler implemented in a 5.1 speaker array placed over the screen of a PC. The crosstalk canceler proved to be effective within the desired bandwidth. However, it demanded a high computational power when all the features such as HRTF and reverberation were included. They also presented a comparison between different inverse

filtering techniques in [4]. This study focused on the stereo-dipoles and in particular their application in mobile phones. They concluded that the performance of the crosstalk canceler improves with the increase of filter length. Additionally, results indicated that the direct filtering technique has a better performance than the filter bank method. From the subjective experiments it was concluded that the best spatializing performance and sound quality it is achieved when HRTFs and a conventional crosstalk cancellation system are used.

Front-back confusion is also an issue that needs to be solved in binaural reproduction systems through loudspeakers. Most researchers agree on that a significant reduction of the front-back confusions is obtained when the sound pressures at the ears change due to head rotations [12, 26, 33]. However, when the listener rotates his/her head to a certain degree, a sufficient amount of rotation will make the loudspeakers be located in the same side of his/her head. As a result, the virtual images will be wrongly placed at the loudspeakers location as a consequence of precedence effect or the law of the first wavefront [6].

In 1998, Kahana et al. made an attempt to tackle this problem by implementing a multichannel system based on a rigid sphere model, which simulates the ears as two microphones placed on each side of the sphere [13]. With this arrangement, the respective inverse filters were calculated using a set of four loudspeakers: two in front at $\pm 30^\circ$ and two in the back at $\pm 110^\circ$ with 0° being right in front. Localization experiments were performed and their results showed a decrease in front-back confusions with the suggested system. They also claimed that the system is robust with respect to head rotations and that they in fact obtained an improved performance with respect to other approaches. However, the amount of rotation was not measured and the cross-talk cancellation showed not to be very efficient at frequencies larger than 4 kHz.

A study carried out by Hill et al. [12], showed through simulations that head rotations indeed help resolving front-back confusions. By means of a spherical head model, two binaural reproduction systems were evaluated: a two- and a four-channel configuration. Both systems were simulated and compared with a real source. The interaural time delays (ITD) were obtained by means of the interaural cross-correlation function and calculated for a static head and a rotated head with angles of $\pm 5^\circ$. Their results show that when virtual images are rendered through a four-channel configuration, the ambiguities are resolved in a similar manner as for the real source.

Richard Clemow also proposed a system including two pairs of loudspeakers, one in the front and one in the back [7]. The way the system approaches the problem is by letting the frontal and rear loudspeakers reproduce only the images situated in their respective hemisphere and by adding cross-fading to account for the transitions from the front to the back. He also proposed to place additional loudspeakers at positions outside the spanning angles where an accurate localization performance is desired e.g. at the sides of the listener for images on the sides. The results obtained by Hill et. al. suggest that the system proposed by Clemow might be a good solution for virtual sources located in the front and back. However, when the virtual source is placed in locations where the ITD update gives little information, e.g. sources placed above, below and at the sides, it is expected that the virtual images will be misplaced to the region near the loudspeakers.

If one places a pair of loudspeakers above the listener, in theory head rotations will not change the interaural differences of the signals that reach the ear significantly [22]. This can be clearly seen in figure 1.2, where HRTFs at the left ear from 0° to 180° at 84° elevation are displayed.

Thus, it is hypothesized that virtual images placed above, below and at the side of the listener will be more accurately reproduced with such a system. Furthermore, by placing an additional pair of dipoles, one in the front and one in the back, it is expected that a more accurate binaural reproduction can be achieved, as well as a wider sweet spot compare to the state of the art approaches.

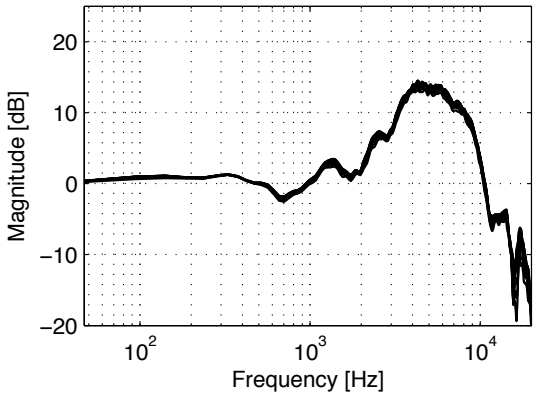


Figure 1.2: *HRTFs at the left ear for sources placed at azimuth angles from* 0° *to* 180° *at* 84° *elevation.*

1.2 Scope of the Thesis

This PhD thesis presents a systematic study of the acoustical and psycho-acoustical characteristics of binaural reproduction systems through loudspeakers. The study is mainly emphasized on the characteristics of the stereo-dipoles when placed at different locations, e.g. in the front of, above or behind the head. The investigations described in this thesis are mainly directed towards the design of a binaural reproduction system that is robust to head rotations and tackles the problem of front-back confusions discussed before. For this purpose, the investigations were organized in three main stages as follows:

1. **Analysis of crosstalk cancellation techniques**

   Most crosstalk canceler filters are designed by using head related transfer functions (HRTF). Since the HRTFs are functions of direction, it can be deduced that the optimum crosstalk cancellation filters have a high dependency on direction as well. Thus, in the initial stage, different crosstalk cancellation techniques were thoroughly studied for several loudspeaker configurations. This gave a better understanding of their characteristics with regards to design parameters and loudspeaker configurations.

2. **Physical and psycho-physical characteristics of the sound field reproduced by loudspeakers placed at different locations**

   After obtaining a better understanding of the different crosstalk cancellation techniques, a set of objective and subjective experiments was carried out in order to investigate the influence of loudspeaker placement on the robustness to head misalignments. This stage covers an extensive study of the sweet spot size of several loudspeaker configurations and some of their perceptual characteristics. Different crosstalk cancellation techniques were also applied.

3. **Estimation of an optimal loudspeaker configuration**

   Gathering the results obtained in the first and second stages of the research, a binaural reproduction system that uses three stereo-dipoles is proposed. The proposed system is compared with more conventional implementations. In this stage, the influence of head rotations in the localization performance is also evaluated for different loudspeaker configurations. Results from this stage set a basis for further research and development of an optimal binaural reproduction system through loudspeakers.

1.3 Synopsis of the Thesis

This thesis is organized in five manuscripts which are the result of the studies carried out. They are related in the following way: Manuscript A focuses on the analysis of different crosstalk cancellation techniques and aims to obtain a general understanding of the methods and their performance with regard to loudspeaker position. Manuscripts B and C present an objective study of the robustness to head misalignments of the different crosstalk cancellation techniques analyzed in Manuscript A, also applied to different loudspeaker configurations. Motivated by some of the assumptions made to evaluate the sweet-spot size in Manuscript B and C, Manuscript D explores some perceptual attributes of the crosstalk. Based on the results observed in Manuscript B and C, a binaural reproduction system using three stereo-dipoles is proposed in Manuscript E.

[Manuscript A] [2] Lacouture Parodi, Y. and Rubak, P. (2010), Analysis of Design Parameters for Crosstalk Cancellation Filters Applied to Different Loudspeaker Configurations, *Journal of the Audio Engineering Society*, re-submitted after first revision from reviewers.

[Manuscript B] [3] Lacouture Parodi, Y. and Rubak, P. (2010), Objective Evaluation of the Sweet Spot Size in Spatial Sound Reproduction Using Elevated Loudspeakers, *Journal of the Acoustic Society of America*, 128(3), 1045 -1055, September 2010.

[Manuscript C] [4] Lacouture Parodi, Y. and Rubak, P. (2010), Sweet Spot Size in Virtual Sound Reproduction: A Temporal Analysis, chapter in *Principles and Applications of Spatial Hearing*, World Scientific, in press.

[Manuscript D] [5] Lacouture Parodi, Y. and Rubak, P. (2010), A Subjective Evaluation of the Minimum Audible Channel Separation in Binaural Reproduction Systems Through Loudspeakers, *Journal of the Audio Engineering Society*, submitted.

[Manuscript E] [6] Lacouture Parodi, Y. and Rubak, P. (2010), Binaural reproduction system using three stereo-dipoles, *Journal of the Audio Engineering Society*, submitted.

This thesis is divided into three main parts which follow the order of the stages of research described in section 1.2: the first part covers the theory of crosstalk cancellation techniques and presents an analysis of their characteristics when applied to several different loudspeaker configurations (Manuscript A). The second part investigates the issues regarding the sweet-spot size of different loudspeaker configurations. This analysis is done from the spectral (Manuscript B) and temporal (Manuscript C) point of view. The perceptual attributes of the crosstalk and the sweet-spot are also investigated (Manuscript D). The third part of the thesis presents an evaluation of a proposed system which makes use of three stereo-dipoles (Manuscript E). A tentative approach to widen the sweet-spot size by means of loudspeaker arrays was also explored during the research process. A short discussion on its advantages and possible limitations are discussed in Appendix A.1.

[2] Manuscript A was presented at the 125^{th} Convention of the Audio Engineering Society, San Francisco, CA, USA, 2008 October 2-5

[3] Portions of Manuscript B were presented in "Preliminary Evaluation of the Sweet Spot Size in Virtual Sound Reproduction Using Dipoles", Proceedings of the 126^{th} Convention of the Audio Engineering Society 2009, 7-10 May, Munich, Germany

[4] Portions of Manuscript C were in "Sweet Spot Size in Virtual Sound Reproduction: A Temporal Analysis", Proceedings of the International Workshop on the Principles and Applications of Spatial Hearing, Miyagi, Zao, Japan, November 11 - 13, 2009

[5] Manuscript D was presented at the 128^{th} Convention of the Audio Engineering Society 2010, May 22-25, London, UK and received **the AES 128^{th} Convention Student Technical Paper Award**

[6] Portions of Manuscript E were presented in "Evaluation of a Binaural Reproduction System Using Multiple Stereo-Dipoles", Proceedings of the 128^{th} Convention of the Audio Engineering Society 2010, May 22-25, London, UK

1.3.1 Crosstalk Cancellation Techniques

Manuscript A: Analysis of design parameters for crosstalk cancellation filters applied to different loudspeaker configurations

This paper presents a systematic study of three different crosstalk cancellation techniques applied to several different loudspeaker configurations. The theory behind crosstalk cancellation is reviewed in the introduction of the paper, followed by a detailed description of the three evaluated techniques. The first crosstalk cancellation technique is based on a minimum-phase approximation. The inverse filters are modeled in terms of the Interaural Transfer Functions (ITF) which are calculated as the ratio between the minimum-phase components of the ipsilateral and contralateral transfer functions. The excess-phase components are approximated with a frequency independent Interaural Time Delay (ITD). The two other methods are based on a least square approximation, one in frequency domain and the other in time domain. These crosstalk cancellation techniques are applied to several different loudspeaker configurations and their performance is evaluated through Matlab simulations. The different loudspeaker configurations analyzed were divided into two- and four-channel configurations. The configurations were simulated with different span angles and at different elevation angles with respect to the head of the listener.

To evaluate the performance of the techniques two indexes were defined: the channel separation index and performance error index. The former is an average over frequency of the channel separation, which is the magnitude ratio of the cross-path to the direct signal. In other words, is a measure of the level of crosstalk that leaks into the direct signal.

The performance error index is also an average over frequency of the performance error, which is defined as the magnitude ratio of the reproduced signal to the desired signal:

$$PE_i(z) = \frac{D_i\, z^{-m}}{R_{ii}(z) + R_{ji}(z)} \tag{1.2}$$

where $R_{ii} + R_{ji}$ corresponds to the signal that reaches the ears and $z = e^{j\frac{2\pi}{N_c}k}$. It is thus a measure of the equalization performance of the crosstalk cancellation filter. Being an average over frequency, these indexes can be considered as indicators of the performance of the filters. This is, some peaks and notches present in the signals can be reduced by the averaging process and thus possible coloration changes or artifacts in the signals are not taken into account in the analysis.

The regularization constant, the filter length and the bandwidth of the filters were varied for each loudspeaker configuration and the performance indexes were calculated for each case. It is shown that in general the least square methods yield better channel separation and less performance errors than the method based on the minimum-phase approximation. Additionally, the choice of regularization shows not to be critical for most configurations and small regularization constants are sufficient to avoid singularities. However, the gain of the filters were not taken into account in the analysis.

The indexes evaluated in this manuscript are calculated under ideal conditions, which means that the same impulse responses used to calculate the filters $\mathbf{C}$ were used to evaluate them. In a real application, the acoustical paths between the loudspeakers and the ears can differ from the original acoustical paths used to design the filters, e.g. head misalignments or different head related transfer functions. Thus, a further evaluation of the crosstalk cancellation techniques described in this manuscript is presented in Manuscripts B and C, where the loudspeakers impulse response and head misalignments are taken into account.

1.3.2 The Sweet Spot Problem

Manuscript B: Objective evaluation of the sweet spot size in spatial sound reproduction using elevated loudspeakers

As mentioned before, this paper presents a continuation of the evaluation of the three different crosstalk cancellation techniques presented in Manuscript A. In this study, a set of measurements of the techniques applied to different loudspeaker configurations is analyzed. The robustness analysis is based on the magnitude ratios of the crosstalk to the direct signal when the head of the listener is not at the center position. The crosstalk cancellation techniques evaluated in Manuscript A were applied to several loudspeaker configurations and the channel separation was measured as a function of lateral displacement, head rotations and frontal displacement. The measured loudspeaker configurations are comprised by three different span angles placed at four different elevations with respect to the listener. Measurements were done with a dummy head and the sweet spot was calculated using the channel separation index defined in Manuscript A.

Two sweet spot definitions were used in this analysis: the absolute sweet spot and the relative sweet spot. The first one is defined as the maximum displacement from the nominal center position[7] that results in a channel separation index not larger than -12 dB. The second one is defined as the maximum displacement from the nominal center position that results in a channel separation index degradation of 12 dB with respect to the nominal center position. This is,

$$\text{Absolute sweet spot} := max\left\{d:\ \overline{CHSP}_d \leq -12dB\right\} \tag{1.3}$$

$$\text{Relative sweet spot} := max\left\{d:\ \overline{CHSP}_d \leq \overline{CHSP}_o + 12dB\right\}, \tag{1.4}$$

where d is the amount of displacement or rotation from the nominal center position, $\overline{CHSP}_d$ is the channel separation index at the displaced position and $\overline{CHSP}_o$ is the channel separation index at the nominal center position. The limit of -12 dB was taken from [2], though it is based on personal experiences of the author.

Closely space loudspeakers showed to be more robust to head misalignments than wider span angles which is in good agreement with previous studies found in the literature [2, 27, 30]. When the loudspeakers are placed above the listener head, the sweet spot size increases significantly. Additionally, a smoother channel separation is also obtained at these particular positions. It is hypothesized that this smoothness might result in less coloration changes of the virtual image.

When comparing the performance of the two- and four-channel configurations, it is observed that whereas with the two-channel configurations the sweet spot increases with narrower span angles, it increases with wider span angles with the four-channel configurations. Yet, the four-channel configurations are in general less robust to head misalignments than the two-channel configurations.

This analysis was based on the magnitude ratios of the direct signal and the crosstalk. Thus, in order to complete this study an analysis of the temporal changes with head misalignments was also carried out (Manuscript C).

Manuscript C: Sweet Spot Size in Virtual Sound Reproduction: A Temporal Analysis

When the listener moves his/her head, not only the magnitude differences between direct signal and crosstalk changes, but also the time information of the binaural signals. In this paper the sweet

[7] The nominal center position is defined at the position where the transfer functions used to calculate the inverse filters were measured.

spot problem is analyzed from a temporal point of view. Based on the measurements described in Manuscript B, temporal variations with respect to head displacements and rotations are analyzed for the different loudspeaker configurations. Delay differences are calculated between the interaural time delays (ITD) of the binaural signals measured at the optimal position (i.e. center position) and the ITD of the same signals measured at a laterally displaced or rotated position. The sweet spot is thus defined as the maximum displacement from the nominal center position, such that the ITD difference between the nominal center position and the new position does not exceed $10\mu s$:

$$\text{ITD sweet spot} := max\{d : \ |ITD_o - ITD_d| \leq 10\mu s\}, \tag{1.5}$$

The $10\mu s$ limit is taken from subjective experiments of the minimum audible ITD [19]. The three crosstalk cancellation techniques that were evaluated in Manuscripts A and B are also compared in this paper.

When evaluating the temporal changes of the binaural signals, a narrower sweet spot is observed in comparison to the results obtained with the magnitude ratios (Manuscript B). Supporting the observations made in Manuscript B, the closely spaced loudspeaker showed to be more robust to head misalignments than wider span angles. Additionally, when the loudspeaker are placed above the head, the delay differences with respect to head rotation showed to be negligible. As oppose to the observations made in Manuscript B, the four-channel configurations show a slightly better performance than the two-channel configurations.

Contrary to what was expected, differences in performance between the crosstalk cancellation techniques were observed. In general, the least square method in the time domain results in narrower sweet spots than the frequency domain approach and the minimum-phase approximations.

Manuscript D: A subjective evaluation of the minimum audible channel separation in binaural reproduction systems through loudspeakers

As described before, the sweet spot size analysis presented in Manuscript B was done under the assumption that a channel separation of -12 dB is sufficient to render accurate virtual images. That value was found in the literature and its choice was based on personal experiences of the referred authors. Motivated by the lack of a systematic evaluation of the maximum acceptable level of crosstalk, a subjective study of the minimum audible channel separation was carried out. The crosstalk was simulated through headphones and the channel separation was modeled by a gain factor and delays between channels.

The minimum audible channel separation was thus defined as the minimum amount of channel separation needed such that the binaural signals with crosstalk are perceived equal to the binaural signals without crosstalk. Delays equivalent to two different span angles were introduced (12° and 60°). The listener was simulated to be positioned at symmetric and asymmetric locations with respect to the loudspeakers. Eight different stimuli were evaluated including broad-band and narrow-band signals.

At high frequencies, the crosstalk shows to be more audible than at lower frequencies. The position of the source has also an effect on the audibility of the crosstalk when simulating a span angle of 60°, where the crosstalk showed to be more audible for sources at the side of the listener. In general, the minimum audible channel separation was found to be around -15 dB for most narrow-band stimuli and about -20 dB for the broad-band signals when the listener is placed at a symmetric position. When the listener is placed at a asymmetric position the minimum audible channel separation goes down to -25 dB for the broad-band signals. According to these results, the sweet spot sizes presented

in Manuscript B are thus expected to be smaller, given that the constraint of maximum acceptable channel separation was underestimated.

1.3.3 The Multiple Stereo-Dipole Approach

Manuscript E: Binaural reproduction system using three stereo-dipoles

In manuscript B and C was observed that closely spaced loudspeakers result in wider sweet-spot. What's more, when placed above the head, those configurations showed to be rather robust to the head rotations. In addition to that, it is assumed that when the binaural signals are rendered through loudspeakers placed on the horizontal plane, head rotations will cause them to be located near the loudspeaker pair that is reproducing them. Thus, based on these assumptions and in an attempt to exploit the observed robustness to head rotations of the loudspeakers placed above the head, a binaural reproduction system comprised by three stereo-dipoles is proposed: one pair in the front of, one pair behind and one pair above the head of the listener. The reproduction of virtual images is thus divided into regions and each loudspeaker pair reproduces only virtual images in their proximity as shown in figure 1.3.

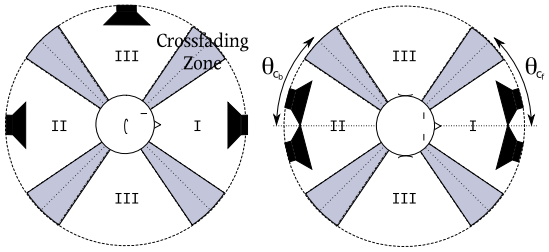


Figure 1.3: *Scheme of the different zones in the three stereo-dipoles (TSD) setup proposed in Manuscript E.* $2\theta_{c_i}$ *is the aperture of the cones that define the limits for the different zones.*

A crossfading is applied to smooth the transitions between regions and is defined by the angles θ_{c_i}. It is hypothesized that with such system, front-back confusions will be reduced and a better localization performance will be obtained.

The implementation and design criteria of the proposed system are described in the paper. Two subjective experiments are carried out in order to evaluate the aforementioned hypothesis and the overall performance of the system. The first subjective experiment consists on a localization evaluation with static sources. The second experiment evaluated the localization performance with dynamic sources. The sources moved between the defined zones around the listener on the horizontal plane. Real and virtual sources were evaluated in both experiments. The virtual sources consisted on binaural recordings made with a dummy head and reproduced by each of the three loudspeaker pair. Virtual sources reproduced by the three loudspeaker pairs working together were evaluated in the second experiment. Both experiments included localization procedures with head still and controlled head rotations.

Results from the first experiment confirmed the hypothesis that head rotations result in virtual images localized close to the loudspeaker pair that is reproducing them, when rendered through the frontal and rear loudspeaker. However, the overall localization performance of the upper loudspeakers was poorer than expected, especially when head rotations were allowed.

In the second experiment dynamic sources were employed. It is observed in the results that front-

back confusions are indeed reduced with the proposed system. Yet, the overall localization decreases when using the three stereo-dipoles working together. It is believed that this is a consequence of the reduced performance observed with the upper loudspeakers when evaluating static sources. The use of regularization at high frequencies in addition to the lack of dynamic cues in the signals when head rotations are allowed, result in a rather large amount of ambiguous cues that reduces the overall localization performance. Some pending issues regarding the design of the system and further research are discussed.

1.4 General Conclusions

1.4.1 Crosstalk Cancellation Techniques

In Manuscript A it was shown that if distortions are introduced in the phase of the impulse responses from the loudspeakers to the ears, large errors can be generated in the inverse filters at specific frequencies and with some loudspeaker configurations. It is argued that if the matrix containing the impulse responses of the acoustical paths to ears is well conditioned, then a rather good performance can be obtained. Such an approach, makes the inverse filters highly dependent on loudspeaker position and less robust to changes in the acoustical paths from the loudspeakers to the ears, e.g. the use of different HRTFs or head misalignments.

Since the least square approximations do not make assumptions of the phase and instead use all the information contained in the impulse responses, the overall performance tends to be more stable with regards to design parameters and loudspeaker configuration. In one hand, the frequency domain approach is more practical and efficient with regards to filter design [18], whereas the time domain approach is more efficient with regards to filter lengths [4, 17]. Thus, the choice of an optimal method to design crosstalk cancellation filters is a matter of the available processing power and memory. When sufficient resources are at hand (e.g. memory allocation for the filters), the fast deconvolution approach is thus preferred.

On the other hand, given that a direct inversion of the acoustic paths from the loudspeakers to the ears is not feasible, the calculation of the crosstalk cancellation filters is an approximation to the optimal solution. Even though the least square methods result in less magnitude and phase errors, there are still differences in the results obtained by each numerical approximation. This is particularly observed in the results presented in Manuscript C, which show that the temporal changes with respect to head movements are not only dependent on the head position but also on the method used to calculate the filters.

Limitations of the regularization

It has been previously shown that to find the exact value of the regularization constant is not critical [18]. This was also observed in Manuscript A, where the choice of an optimal regularization value proved not to be a critical parameter in the design of optimal filters. Yet, its choice is still a strong compromise between performance error, channel separation and filter gain. To regularize only the critical frequencies (e.g. low and high frequencies) has proved to be sufficient to avoid singularities and limit the gain of the filters. However, the rather large regularization usually needed at high frequencies when using head related transfer functions (HRTF), results in a poor performance of the crosstalk cancellation filters at those frequencies. Figure 1.4 shows an example of the channel separation of a loudspeaker pair spanning 12° measured with a dummy head. The loudspeakers are placed above

the mannequin's head. The crosstalk cancellation filters were calculated with the fast deconvolution approach and the shape factor used is a band-stop filter with stop bands at 500 Hz and 6 kHz (see Manuscript A for details). The regularization constant is set to $\beta = 0.0082$.

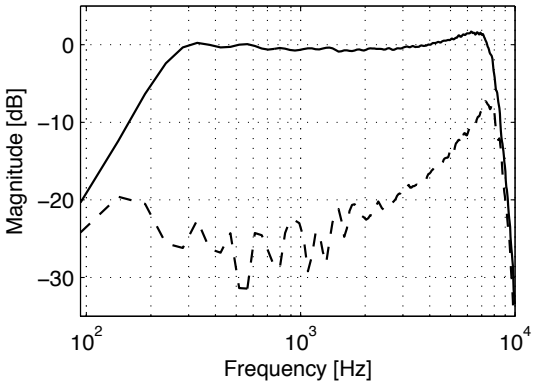


Figure 1.4: *Frequency reponse of the direct signal (solid line) and crosstalk (dashed line) measured at the left ear of a dummy head. The loudspeaker configuration used is* 12° *span angle placed above the mannequin's head. The shape factor used is a band-stop filter with stop bands at 500 Hz and 6 kHz and the regularization constant is set to* $\beta = 0.0082$.

The effect of the shape factor can be clearly seen at frequencies above 3 kHz, where the magnitude of the crosstalk begins to increase considerably. Additionally, there is also a slight increase in the magnitude of the direct signal at those frequencies due to the regularization. This degradation in performance at high frequencies can have undesirable consequences in the rendering of binaural signals. It is observed in Manuscripts D and E that insufficient channel separation at high frequencies (around and above 4 kHz) results in unclear locations of the virtual image or wrongly perceived elevation angles.

In Manuscript D is shown that crosstalk leakages are more critical at high frequencies, especially when the virtual image is at the side of the listener. It is known that the interaural level differences (ILD) are the main localization cues at frequencies above 1.6 kHz [6]. Thus, small changes of ILD at high frequencies are more audible than at lower frequencies. If there is not sufficient channel separation at frequencies above 1.6 kHz, then leakages from the contralateral paths result in ambiguous binaural cues. This casts a much stricter constraint when it comes to define the amount of regularization applied at high frequencies as is observed in Manuscript E. However, in Manuscript D it is also argued that if sufficient channel separation is provided but the time information of the virtual image is affected by for example a head misalignment, the delay differences could potentially destroy the virtual impression. The latter is especially critical with wider span angles where the path lengths variations between channels are considerably larger than with closely spaced loudspeakers.

1.4.2 Robustness to Head Misalignments

The analysis of crosstalk cancellation techniques with respect to loudspeaker position described in Manuscript A was done under ideal conditions. This is, the HRTFs used to calculate the inverse filters were also used to analyze their performance. Additionally, the frequency response of the loudspeakers were not taken into account. Therefore, in order to extend the analysis of the methods under more realistic conditions, measurements of the performance of the methods with respect to loudspeaker position and head misalignments were analyzed in Manuscript B and C.

In Manuscript A is shown that the use of four channels in binaural reproduction systems results in better channel separation when the listener is at the optimal position, as oppose to the conventional two-channel configurations. Yet, it proved to be more susceptible to errors in the acoustic paths (e.g. the head misalignments presented in Manuscript B and C) than the two-channel configurations. This is a consequence of being an overdetermined system and the large amount of redundancy between the impulse responses. Additionally, the design of crosstalk cancellation network with more channels is also found to be a computationally demanding task. Therefore, no real advantage over the conventional two-channel configurations is attained.

Closely spaced loudspeakers exhibit a relatively large sweet spot in comparison to wider span angle configurations, which is in good agreement with previous studies on this topic [2, 30]. The sweet spot with respect to head rotations increases systematically with the loudspeakers' elevation angle and is significantly wider when the loudspeakers are placed above the head if compared with other configurations as shown in Manuscripts B and C. Additionally, in Manuscript C is also observed that only when the loudspeakers are placed close to each other there is some tolerance to the temporal changes introduced by the head movements.

Elevated Loudspeakers

The large sweet spot size with respect to head rotations observed in Manuscript B and C confirms the thesis that the small variations of the HRTFs with respect to azimuth at elevated positions would result in a crosstalk cancellation system more robust to head rotations, from the objective point of view. Nevertheless, this also implies that when the loudspeakers are placed above the head and the head is rotated, no updates in the HRTFs are perceived. From the perceptual point of view, such lack of updates can also result in ambiguous cues since the auditory system might be expecting spectral and temporal changes (dynamic cues) that correlates with the location of the source [26, 33]. In Manuscript E, this proved to be rather critical in the localization performance of virtual images. The lack of dynamic cues in the virtual images rendered through the elevated loudspeakers resulted in a strong bias towards the frontal plane and images with undefined locations.

On the other hand, when the loudspeakers are placed at elevated positions, the channel separation as function of frequency exhibits a smoother pattern when compared with loudspeakers placed on the horizontal plane. This is a consequence of the inherent smoothness observed in the HRTFs at those locations. Figure 1.5(a) shows an example of the HRTFs at 6° in the horizontal plane and figure 1.5(b) shows the HRTFs at 90° azimuth 84° elevation. It is clear that the peaks and notches present in the HRTFs in the horizontal plane are not so pronounced in the HRTFs above the head. This implies that when the loudspeakers are above the head, the system is better conditioned.

In Manuscript B is suggested that when placing the loudspeakers at elevated positions, not only there is an improvement in robustness to head rotations but also there could be a potential reduction of coloration and artifacts in the virtual images. Even though the high frequency peaks and notches of the HRTFs above the head are not so pronounced as the ones observed on the horizontal plane, they still play an important roll in the elevation judgement [26]. Errors at those frequencies can easily change the coloration of the source if the accuracy of the inversion at those frequencies is low. This is especially critical due to the regularization. By regularizing the high frequencies the channel separation is considerably reduced in those frequency bands (see figure 1.4). This results in audible leakages from the contralateral paths which contribute to decrease significantly the overall performance of the crosstalk cancellation network.

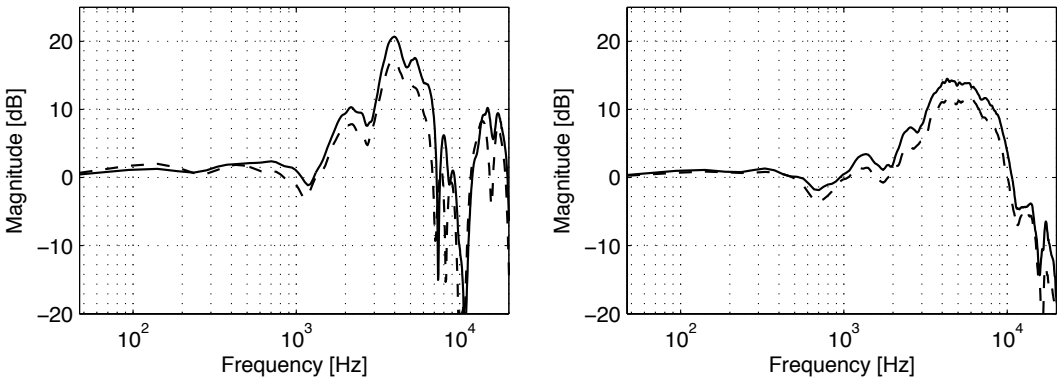


Figure 1.5: *Examples of HRTFs corresponding to* 6° *azimuth on the horizontal plane (a) and* 90° *azimuth* 84° *elevation (frontal plane) (b). Solid lines: left ear (ipsilateral), dashed lines: right ear (contralateral).*

1.4.3 The Multiple Stereo-Dipole Approach

Head rotations are known to improve the localization performance of real sources [12, 26, 33]. This is also the case when the source is moving as is observed in Manuscript E. In the case of virtual sources, it is shown that head rotations place the virtual images near the loudspeakers pair that is reproducing them. This is especially observed with loudspeakers placed in front and behind the listener. Thus, when the binaural reproduction is divided into regions and each loudspeakers pair renders only the images in its vicinity as proposed in Manuscript E, it is shown that head rotations help indeed to resolve front-back confusions. Nevertheless, there are unavoidable differences between the filters applied to each loudspeaker pair due to the nature of the impulse responses used to calculate the filters. As an example of these differences, figure 1.6 shows the measured crosstalk for the three loudspeaker pairs used in the experiments described in Manuscript E. Notice that there are significant differences of the crosstalk magnitude between the loudspeaker pairs, especially at high frequencies. This is also a consequence of the regularization at high frequencies, which means that the peaks and notches due to the pinna effect that leak into the reproduced signals result in coloration differences between loudspeaker pairs.

In Manuscript E the proposed binaural reproduction system with three stereo-dipoles is also evaluated with dynamic sources. It is shown that even though the front-back confusions are considerably reduced with such a system, the coloration changes between loudspeaker pairs degrades substantially the localization performance. In addition, the lack of dynamic cues in the binaural signals when the head is rotated and the images are reproduced by the upper loudspeakers introduces into the virtual sources more ambiguities than expected.

1.4.4 Limitations of the study and further work

The indexes used to evaluate the performance of the different crosstalk cancellation techniques and the robustness to head misalignments are averages over frequency. In consequence, they are only indicators of the overall performance of the evaluated systems. By doing an average over frequency, peaks and notches in the impulses are reduced. In other words, the values presented in Manuscripts A and B do not take into account the variations in performance with frequency. Those variations have proved to be

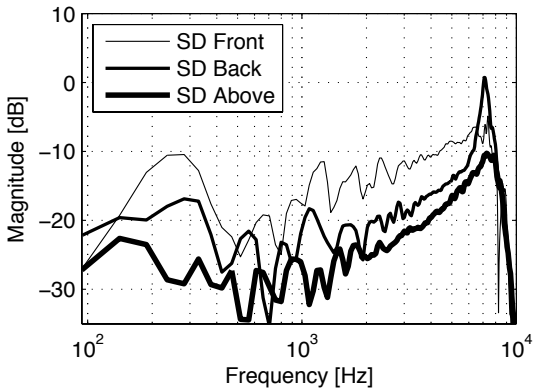


Figure 1.6: *Frequency response of the crosstalk measured at the left ear of a dummy head with the stereo-dipoles at the front (SD front), at the back (SD back) and above (SD above). The mannequin is placed at the nominal center position.*

critical in the overall performance of binaural reproduction systems through loudspeakers as shown in Manuscript E. Therefore, in order to assess properly these systems a better set of indexes that take into account variations with frequency and perceptual attributes should be developed.

It has been shown that by dividing the reproduction of binaural signals into regions, front-back confusions are considerably reduced. However, there are still some open questions regarding the problem of coloration changes between loudspeaker pairs and possible ways to tackle it. Additionally, it is necessary to investigate into which extend the lack of dynamic cues perceived in the virtual images when the head is rotated influences the perceived location.

This PhD thesis addresses the binaural reproduction systems through loudspeakers from a very controlled point of view. In real applications the binaural reproduction systems will be affected by reflexions and the room's characteristics. Thus, an extended study of the performance of such systems should take into account the influence of the room in it and possible solutions to overcome the problems that such inclusion might imply. A possible way to reduce the room influence in the binaural reproduction could be the use of loudspeaker arrays and beamforming as discussed in Appendix A.1. This is an interesting topic for further research.

Another interesting issue to look into is the necessity of using individual HRTFs to reproduced a more accurate virtual environment. This should be addressed together with the regularization effects at high frequencies and its influence in the perceived elevation of the sources.

BIBLIOGRAPHY

[1] B.S. Atal and M. R. Schroeder. Apparent Sound Source Translator. U.S. Patent 3,236,949, February 1966.

[2] Mingsian R. Bai and Chih-Chung Lee. Objective and Subjective Analysis of Effects of Listening angle on Crosstalk Cancellation in Spatial Sound Reproduction. *Journal of the Acoustic Society of America*, 120(4):1976–1989, October 2006.

[3] Mingsian R. Bai and Chih-Chung Lee. Development and Implementation of Cross-Talk Cancellation System in Spatial Audio Reproduction Based on Subband Filtering. *Journal of Sound and Vibration*, 290:1269–1289, August 2005.

[4] Mingsian R. Bai, Geng-Yu Shih, and Chih-Chung Lee. Comparative Study of Audio Spatializers for Dual-Loudspeaker Mobile Phone. *Journal of the Acoustic Society of America*, 121(1):298–309, January 2007.

[5] Jerry Bauck. A Simple Loudspeaker Array and Associated Crosstalk Canceler for Improved 3D Audio. *Journal Audio Engineering Society*, 49(1 - 2):3 – 13, 2001.

[6] Jens Blauert. *Spatial Hearing*. Hirzel-Verlag, 3rd edition, 2001.

[7] Richard David Clemow. Method of Synthesizing a Three Dimensional Sound-Field. U. S. Patent 6,577,736, June 2003.

[8] Duane H. Cooper and Jerald L. Bauck. Head Diffraction Compensated Stereo System With Optimal Equalization. U.S. Patent 4,975,954, December 1990.

[9] Duane H. Cooper and Jerald L. Bauck. Prospects for Transaural Recordings. *Audio Engineering Society*, 37(1/2):3–19, January/February 1989.

[10] P. Damaske. Head-Related Two-Channel Stereophony with Loudspeaker Reproduction. *Journal of the Acoustic Society of America*, 50 (Part 2)(4):1109–1115, November 1971.

[11] William G. Gardner. *3-D Audio Using Loudspeakers*. Kluwer Academic Publishers, 1st edition, 1998.

[12] P. A. Hill, P. A. Nelson, O. Kirkeby, and H. Hamada. Resolution of Front-Back Confusion in Virtual Acoustic Imaging Systems. *Journal of the Acoustic Society of America*, 108(6):2901–2909, December 2000.

[13] Yuvi Kahana, Philip A. Nelson, Ole Kirkeby, and Hareo Hamada. A Multiple Microphone Recording Technique for the Generation of Virtual Acoustic Images. *Journal of the Acoustic Society of America*, 105(3):1503–1516, March 1998.

[14] Y. Kim, O. Deille, and P.A. Nelson. Crosstalk Cancellation in Virtual Acoustic Imaging Systems for Multiple Listeners. *Journal of Sound and Vibration*, 297:251–266, June 2006.

[15] Ole Kirkeby and Philip A. Nelson. The "Stereo Dipole" – A Virtual Source Imaging System Using Two Closely Spaced Loudspeakers. *Audio Engineering Society*, 46(5):387–395, May 1998.

[16] Ole Kirkeby and Philip A. Nelson. Digital Filter Design for Inversion Problems in Sound Reproduction. *Audio Engineering Society*, 47(7/8):583–595, July/August 1999.

[17] Ole Kirkeby, Philip A. Nelson, and Hareo Hamada. Local Sound Field Reproduction Using Two Closely Spaced Loudspeakers. *Journal of the Acoustic Society of America*, 104(4):1973–1981, October 1998.

[18] Ole Kirkeby, Philip A. Nelson, Hareo Hamada, and Felipe Orduna-Bustamante. Fast Deconvolution of Multi-Channel Systems Using Regularization. *IEEE Transactions on Speech and Audio Processing*, 6(2):189–195, 1998.

[19] R. G. Klump and H. R. Eady. Some Measurement of Interaural Time Difference Thresholds. *The Journal Acoust. Soc. Am.*, 28(5):859 – 860, September 1956.

[20] Improved Sound Separation Using Three Loudspeakers. Jun Yang and Woon-Seng Gan and See-Ee Tan. *Acoustic Research Letters Online-ARLO*, 4(2):47–52, April 2003.

[21] Henrik Møller. Reproduction of Artificial-Head Recordings Through Loudspeakers. *Audio Engineering Society*, 37(1/2):30–33, January/February 1989.

[22] Henrik Møller. Spatial Sound Reproduction System with Loudspeakers. International Patent WO2008135049 (A1), 13th November 2008.

[23] P. A. Nelson and J. F. Rose. Errors in Two-Point Sound Reproduction. *Journal of the Acoustic Society of America*, 118(1):193–204, July 2005.

[24] Philip A. Nelson, Felipe Orduña-Bustamante, and Hareo Hamada. Inverse Filter Design and Equalization Zones in Multichannel Sound Reproduction. *IEEE Transactions on Speech and Audio Processing*, 3(3):185 – 192, May 1995.

[25] Jan Abildgaard Pedersen. Audio Processor for Narrow-Spaced loudspeaker Reproduction. International Patent WO 2006/076926 A2, July 2006.

[26] Stephen Perrett and William Noble. The Effect of Head Rotations on Vertical Plane Sound Localization. *Journal of the Acoustic Society of America*, 102(4):2325–2332, October 1997.

[27] John Rose, Philip Nelson, Boaz Rafaely, and Takashi Takeuchi. Sweet Spot Size of Virtual Acoustic Imaging Systems at Asymmetric Listener Locations. *Journal of the Acoustic Society of America*, 112(5):1992–2002, November 2002.

[28] M. R. Schroeder. Digital Simulation of Sound Transmission in Reverberant Spaces. *Journal of the Acoustic Society of America*, 47(2 (Part I)):424–431, February 1969.

[29] Takashi Takeuchi and Philip A. Nelson. Optimal Source Distribution for Binaural Synthesis Over Loudspeakers. *Journal of the Acoustic Society of America*, 112(6):2786–2797, December 2002.

[30] Takashi Takeuchi and Philip A. Nelson. Robustness to Head Misalignment of Virtual Sound Imaging Systems. *Journal of the Acoustic Society of America*, 109(3):958–971, March 2001.

[31] Darren B. Ward. Joint Least Squares Optimization for Robust Acoustic Crosstalk Cancellation. *IEEE Transactions on Speech and Audio Processing*, 8(2):211–215, February 2000.

[32] Darren B. Ward and Gary W. Elko. Effect of Louspeaker Position on the Robustness of Acoustic Crosstalk Cancellation. *IEEE Signal Processing Letters*, 6(5):105–108, May 1999.

[33] Frederic L. Wightman and Doris J. Kistler. Resolution of Front-Back Ambiguity in Spatial Hearing by Listener and Source Movement. *Journal of the Acoustic Society of America*, 105(5): 2841–2853, May 1999.

2 MANUSCRIPT A

Lacouture Parodi, Y. and Rubak, P. (2010), **Analysis of Design Parameters for Crosstalk Cancellation Filters Applied to Different Loudspeaker Configurations**, *Journal of the Audio Engineering Society*, re-submitted after first revision from reviewers.

This manuscript was presented at the 125^{th} Convention of the Audio Engineering Society, San Francisco, CA, USA, 2008 October 2-5.

Analysis of design parameters for crosstalk cancellation filters applied to different loudspeaker configurations

Yesenia Lacouture Parodi[1] and Per Rubak[1]

[1]*Aalborg University, Aalborg, DK-9220, Denmark*

Correspondence should be addressed to Yesenia Lacouture Parodi (ylp@es.aau.dk)

ABSTRACT

Several approaches to render binaural signals through loudspeakers have been proposed in the past decades. Some studies have focused on the optimum loudspeaker arrangement while others have proposed more efficient filters. However, to our knowledge, the identification of optimal parameters for crosstalk cancellation filters applied to different loudspeakers configurations has not yet been addressed systematically. In this paper, we document a study of three different inversion techniques applied to several loudspeaker arrangements. Least square approximations in frequency and time domain are evaluated along with a crosstalk canceler based on minimum-phase approximation with a frequency independent delay. The three methods were applied to loudspeaker configurations with two-channel and the least square approaches to configurations with four channels. Several different span angles and elevations were simulated for each case. In order to obtain optimum parameters, we varied the bandwidth, filter length and regularization constant for each loudspeaker configuration and each method. We present in this paper a description of the simulations performed and the obtained results. The simulation results are documented in terms of the channel separation index, optimum regularization constant and performance error.

1. INTRODUCTION

Binaural technology is based on the assumption that the sound pressures at the ears control our sound perception. One of the biggest challenges of reproducing binaural signals through loudspeakers is to be able to achieve this control. This is due to the fact that the signal which is to be heard on the left ear is also heard on the right ear and vice-versa. This problem is known as crosstalk and can be alleviated by introducing proper filters into the reproduction chain. Figure 1 shows a simplified diagram of a multi-channel binaural reproduction system. We can describe this system in matrix form in the frequency domain as follows ($z = e^{-j\omega}$):

$$\underbrace{\begin{bmatrix} W_1(z) \\ W_2(z) \end{bmatrix}}_{\mathbf{w}(z)} = \underbrace{\begin{bmatrix} H_{11}(z) & H_{12}(z) \\ H_{21}(z) & H_{22}(z) \\ \vdots & \vdots \\ H_{n1}(z) & H_{n2}(z) \end{bmatrix}^T}_{\mathbf{H}^T(z)} \cdot \underbrace{\begin{bmatrix} V_1(z) \\ V_2(z) \\ \vdots \\ V_n(z) \end{bmatrix}}_{\mathbf{v}(z)} \quad (1)$$

where $\mathbf{w}(z)$ is a vector which contains the sound pressures at the eardrums, $\mathbf{H}(z)$ is the matrix containing the acoustical transfer functions $H_{ij}(z)$ from the loudspeakers to the ears. In this paper, we will refer to this matrix and the transfer functions as the plant matrix and the plant transfer functions respectively. The vector $\mathbf{v}(z)$ contains the loudspeakers input signals which are the result of multiplying the crosstalk cancellation filters matrix with the desired signals, i.e. $\mathbf{v}(z) = \mathbf{C}(z)\mathbf{d}(z)$.

From equation 1 we have that $\mathbf{w}(z) = \mathbf{H}(z)\mathbf{C}(z)\mathbf{d}(z)$. Ideally, perfect crosstalk cancellation is obtained when $\mathbf{w}(z) = \mathbf{d}(z)$. This is:

$$\mathbf{H}(z)\mathbf{C}(z) \quad = \quad \mathbf{I} \quad (2)$$

where $\mathbf{I}$ is the identity matrix. However, due to the inherent characteristics of the plant transfer functions H_{ij}, the matrix $\mathbf{H}(z)$ is in general singular for some frequencies and therefore not invertible. Consequently, we need to model the problem in a way that it best approximates the required solution.

Some studies have addressed the problem by developing different methods to design crosstalk cancellation filters. For example, in [9] a crosstalk canceler based on

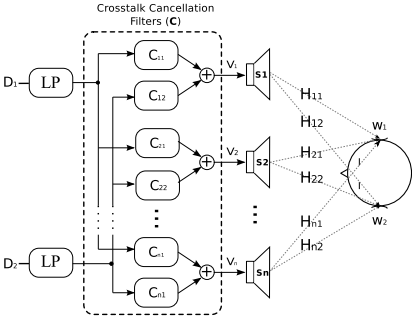


Fig. 1: Acoustic paths from loudspeakers to the ears. The functions H_{ij} represents the transfer functions between loudspeakers and the ears. The signals D_j are the desired binaural signals to be reproduced and the signals W_j are the signals which are reproduced at the ears. The blocks LP represent the low-pass filters applied to the crosstalk cancellation network.

minimum-phase approximation with a frequency independent delay is presented. A more elaborated algorithm is described by Kirkeby et. al in [14], where the crosstalk cancellation filters are obtained by means of a least square approximation using the fast fourier transform (FFT) and regularization. Another least square approach is proposed in [12] but in the time domain. In general, methods based on least square approximation are the most commonly used and a number of approaches have been introduced in the last decades [1, 23, 28].

The appropriate solution of equation 2 is highly dependent on the loudspeaker placement. Thus, other studies have proposed different span angles and loudspeaker arrangements to improve performance. Ward and Elko suggested that the loudspeaker spanning should be inversely proportional to the operating frequency in order to obtain a numerically robust crosstalk canceler [29]. Based on these results, Takeuchi and Nelson proposed the so called optimum source distribution [25]. Additonally, some have suggested the use of closely spaced loudspeakers - also known as stereo-dipoles - in order to obtain wider sweet spots [11], while others have proposed to use more loudspeaker pairs [4, 8, 10, 19].

Nevertheless, a systematic evaluation of crosstalk cancellation filter parameters applied to different loudspeakers configuration has not been found in the literature known by the authors. This is of practical importance when an optimum performance is desired but there exist constraints regarding span angle, amount of loudspeakers and computational power, e.g. in home entertainment or mobile phones applications.

In this paper we present an evaluation of three inversion techniques applied to several loudspeaker configuration. Our main purpose is to find the key characteristics of each method with respect to loudspeaker configuration. The first inversion technique consists of a crosstalk cancellation system based on a minimum-phase approximation proposed by Gardner in [9]. This system is modeled in terms of the Interaural Transfer Functions (ITF) which are calculated as the ratio between the minimum-phase components of the ipsilateral and contralateral transfer functions. The excess-phase components are approximated with a frequency independent Interaural Time Delay (ITD). We will refer to this method as the generic crosstalk canceler.

The two other methods are based on a least square approximation, one in frequency domain and the other in time domain [12, 14]. The use of a regularization factor is often suggested in order to limit the gain of the filters and to ensure that the matrix, which is to be inverted, is not singular. This factor can be either a constant or frequency dependent. We consider the latter to be more convenient given that we only want a certain range of frequencies to be weighted [15].

We evaluated the methods throughout MATLAB simulations. As mentioned before, some studies have proposed the use of more than two-channel setups. Thus, to obtain a better understanding of the performance of the methods, we evaluated in this study two- and four-channel configurations. The first was evaluated at different span and elevation angles placed in the region corresponding to the half hemisphere in front and above of the listener. We derived the crosstalk cancellation filters using the three mentioned methods and varied parameters such as filter length, regularization constant and bandwidth. We used a frequency dependent regularization in this case. The four-channel configurations were evaluated for symmetric and asymmetric arrangements at different elevations. We set a frequency independent regularization in this case. Only the least square methods were used, given that the theory used to derive the generic crosstalk canceler is only applicable for the two-channel configuration.

This paper is organized as follows: the second section provides an overview of the methods evaluated and their

respective implementation. In the third section we describe the simulations carried out, the different loudspeaker configurations and the parameters we varied. In the fourth section we document the simulation results for the different loudspeaker configurations. The results are divided into two main groups: two- and four-channel arrangements. A short discussion of the results obtained for each of these arrangements is presented. The fifth section contains the general conclusions and limitations of the study.

2. CROSSTALK CANCELLATION METHODS

We can approach the problem described in equation 2 by either trying to find the exact solution or an approximation that results in minimum errors. This section describes the three methods we simulated and analyzed in this paper. The first one is based on the exact solution. By rearranging the plant matrix $\mathbf{H}$ we attempt to solve the problem in a direct and simple manner. We will refer to this method as the generic crosstalk canceler (GCC). The second and third methods do not try to invert directly the plant matrix, but instead they seek for a set of causal finite impulse response (FIR) filters that are the best approximation to the desired solution in a least square sense.

2.1. Generic Crosstalk Canceler

Applying the exact inverse matrix definition [5, p. 68], equation 2 equals:

$$\mathbf{C}(z) = \frac{1}{\bar{D}} \begin{bmatrix} H_{22}(z) & -H_{21}(z) \\ -H_{12}(z) & H_{11}(z) \end{bmatrix} \tag{3}$$

where $\bar{D} = H_{11}(z)H_{22}(z) - H_{12}(z)H_{21}(z)$. Assuming that the loudspeakers have ideal frequency responses, we can rewrite this equation in terms of the interaural transfer functions (ITF) by dividing numerator and denominator by $H_{11}(z)H_{22}(z)$ [9, 20]:

$$\mathbf{C}(z) = \frac{1}{\hat{D}} \begin{bmatrix} \frac{1}{H_{11}(z)} & 0 \\ 0 & \frac{1}{H_{22}(z)} \end{bmatrix} \begin{bmatrix} 1 & -ITF_2(z) \\ -ITF_1(z) & 1 \end{bmatrix} \tag{4}$$

where $\hat{D} = 1 - ITF_1(z)ITF_2(z)$ and the terms $ITF_1(z) = {}^{H_{12}(z)}/_{H_{11}(z)}$ and $ITF_2(z) = {}^{H_{21}(z)}/_{H_{22}(z)}$ are the interaural transfer functions for the left and right channel respectively.

The transfer functions H_{ij} are in general non-minimum phase and therefore non-invertible. However, it is possible to obtain an approximate inverse by decomposing them into minimum-phase and excess-phase systems [24, p. 280 ff.]. According to Gardner, the phase of the excess-phase section of the plant transfer functions is approximately linear at low frequencies. Following this line, we can model the ITF as the ratio between the minimum-phase components of the ipsilateral and contralateral transfer functions, cascaded with a frequency independent delay [9]:

$$ITF_i(e^{j\omega}) \cong \frac{H_{ij}(e^{j\omega})_{minphase}}{H_{ii}(e^{j\omega})_{minphase}} e^{-j\omega ITD/T}, \quad j \neq i \tag{5}$$

where H_{ij} and H_{ii} are the contralateral and ipsilateral transfer functions respectively. The term ${}^{ITD}/_{T}$ is the frequency independent delay in samples, where ITD is the interaural time delay (in seconds) and T is the sampling period.

Notice that this approach is only possible when $\mathbf{H}$ is square, since equation 3 only holds for square matrices. Therefore, for configurations where the number of loudspeakers is larger than the number of receivers (in this case ears) the only possible solution is the Moore-Penrose generalized inverse, generally known as pseudo-inverse. This solution is based on a least square approximation [5, p. 250].

2.2. Least Square Approximation

This technique does not try to find the exact inverse of the plant matrix $\mathbf{H}(z)$, but instead it attempts to find a set of causal and finite filters $\mathbf{C}(z)$, which when convolved with $\mathbf{H}(z)$ produce an output that is the best approximation (in the least square sense) to the identity matrix. The central idea is to minimize a quadratic cost function of the type

$$J = E + \beta V = \mathbf{e}^H(z)\mathbf{e}(z) + \beta \mathbf{v}^H(z)\mathbf{v}(z) \tag{6}$$

where E is a measure of the performance error $\mathbf{e} = \mathbf{d} - \mathbf{w}$, which is the deviation of the reproduced signals $\mathbf{w}$ from the desired signals $\mathbf{d}$ (see figure 1). V is a measure of the effort $\mathbf{v}$, which are the input signals to the loudspeakers. The superscript H denotes the Hermitian transpose operator. The positive real number β is the so called regularization parameter, which determines how much weight is assigned to the effort term $\mathbf{v}$ [14]. As it is increased from zero to infinity, the solution changes gradually from minimizing E only to minimizing V only. It can be shown that in our crosstalk cancelation problem described in equation 2 the total cost function J is minimum when [12, 27]

$$\mathbf{C}_o(z) = [\mathbf{H}^H(z)\mathbf{H}(z) + \beta \mathbf{I}]^{-1}\mathbf{H}(z)^H z^{-m} \tag{7}$$

where m is a modeling delay introduced in order to ensure causality and compensate for the non-minimum-phase components of the system [21]. In other words, it ensures that the crosstalk cancellation network performs well not only in terms of amplitude, but also in terms of phase. The general and practical rule of thumb is to choose the modeling delay to be half the length of the filters. The choice of this value is not critical, since its optimal value lies in a broad range [30, pp. 238-239].

The term $\beta\mathbf{I}$ is meant to compensate for singularities and therefore ensures that the matrix is always invertible. It is convenient to consider the regularization to be a product of two components: a gain factor β and a shape factor $B(z)$ [15]. The gain factor β is a small positive number, and the shape factor $B(z)$ is the z-transform of a digital filter. The frequencies that are suppressed by $B(z)$ are not affected by the regularization (see figure 5). The phase response of $B(z)$ is irrelevant, since $C(z)$ is determined by minimizing an energy quantity.

Using this approach, equation 7 can be rewritten in the following way [15]

$$\mathbf{C}_o(z) = [\mathbf{H}^H(z)\mathbf{H}(z) + \beta\mathbf{B}^H(z)\mathbf{B}(z)]^{-1}\mathbf{H}^H(z)z^{-m} \quad (8)$$

where $\mathbf{B}(z) = B(z)\mathbf{I}$. This method can be implemented either in frequency or time domain. The frequency domain approach, known as the fast deconvolution method [14], is based on the Fast Fourier Transform (FFT). It designs a matrix of causal FIR filters which performance is optimized at a large number of discrete frequencies. Its main advantage is that it is computationally efficient with respect to filter design, especially for multi-channel deconvolution problems. However, it suffers from circular convolution effects and the optimal filters will normally be larger than the corresponding minimization in the time domain [3]. One way to counteract this issue is by making the regularization frequency dependent. In this way we can control the time response of the optimal filters in a quite profound way. The regularization parameter influences mainly the poles that are closest to the unit circle. By increasing the regularization, we push the poles further away from the unit circle, thereby shortening the length of the inverse filter. This will unfortunately increase the performance error and the accuracy of the inversion will be degraded. Thus, we have to choose an appropriate compromise between filter length and accuracy of the inversion.

The time domain approach requires a formulation more complicated that the one presented in the z-domain in equation 8, even though the approach is practically the same [13]. It is computationally more cumbersome than the frequency domain deconvolution for calculating the filters, but it has shown to make efficient use of the available filter coefficients [12]. Additionally, it avoids the artifacts introduced by the circular convolution effect, which can be an advantage when short filters are needed [3].

2.2.1. Frequency Domain Approach

The implementation of this approach is quite straightforward. We use the FFTs of the plant transfer functions to get into the frequency domain and then we invert the system for each single frequency. Thus, equation 8 will be equivalent to:

$$\mathbf{C}_o(k) = [\mathbf{H}^H(k)\mathbf{H}(k) + \beta\mathbf{B}^H(k)\mathbf{B}(k)]^{-1}\mathbf{H}^H(k)e^{\frac{-j2\pi(k-1)m}{N_c}} \quad (9)$$

for $k = 1,\ldots,N_c$, where N_c is the filter length and m is the modeling delay in samples. The matrix $\mathbf{B}(k)$ is defined as the shape factor $B(k)$ multiplied by the identity matrix $\mathbf{I}$. As mentioned before, this shape factor is a digital filter used to limit the gain of the filters at selected frequencies.

2.2.2. Time Domain Approach

In order to derive an expression similar to equation 8 but in the time domain, we need to use a set of matrices composed by convolution matrices (see Appendix). First, we define a convolution matrix $\mathbf{h}_{ij}$ that contains a digital FIR filter with length N_h, which describes the electroacoustic impulse responses h_{ij}. Here we use low case bold characters to denote a convolution matrix.

Notice that the impulse responses h_{ij} have to be convolved with the inverse filters c_{ij}. Thus, the size of the matrix $\mathbf{h}_{ij}$ should be $N_c \times (N_h + N_c - 1)$, where N_c is the length of the inverse filters c_{ij} (see Appendix) . Now, the plant matrix will be a matrix composed by the convolution matrices of each impulse response and has the form

$$\mathbf{H}_t = \begin{bmatrix} \mathbf{h}_{11} & 0 & \mathbf{h}_{21} & 0 & \ldots & \mathbf{h}_{n1} & 0 \\ 0 & \mathbf{h}_{11} & 0 & \mathbf{h}_{21} & \ldots & 0 & \mathbf{h}_{n1} \\ \mathbf{h}_{12} & 0 & \mathbf{h}_{22} & 0 & \ldots & \mathbf{h}_{n2} & 0 \\ 0 & \mathbf{h}_{12} & 0 & \mathbf{h}_{22} & \ldots & 0 & \mathbf{h}_{n2} \end{bmatrix} \quad (10)$$

where n represents the number of loudspeakers.

The shape factor can be defined in a similar fashion. In the time domain, the effort term V of equation 6 is equivalent to the frequency domain case (i.e. $\mathbf{v_t}^T \cdot \mathbf{v_t}$). In this case, $\mathbf{v}_t$ is a column vector composed of the vectors describing the sum of the energy of the n loudspeakers. Thus, it can be

shown that the shape factor will be equivalent to a $nN_c \times nN_c$ block-diagonal matrix with the form [12]:

$$\mathbf{B}_t = \begin{bmatrix} \hat{\mathbf{b}} & 0 & \dots & 0 \\ 0 & \hat{\mathbf{b}} & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & \hat{\mathbf{b}} \end{bmatrix} \tag{11}$$

where $\hat{\mathbf{b}}$ is a convolution matrix containing the impulse response of the shape factor. Similar to equation 9, the optimal filters in time domain are given by

$$\mathbf{c}_t = [\mathbf{H}_t^T \mathbf{H}_t + \beta \mathbf{B}_t^T \mathbf{B}_t]^{-1} \cdot \mathbf{H}_t^T d_m \tag{12}$$

The result is then a composed column vector $\mathbf{c}_t$ containing the inverse filters c_{ij} of length N_c, i.e. $\mathbf{c}_t = [c_{11}, c_{12}, ..., c_{ij}]^T$ for $i = 1, \dots, n$ and $j = 1, 2$. The vector $d_m = [\underline{\delta} \quad 0 \dots 0 \quad \underline{\delta}]^T$ with length nN_c, represents the modeling delay. $\underline{\delta}$ is a column vector of length N_c, that corresponds to the delayed unit sample sequence $\delta(t-m)$.

3. SIMULATIONS

We implemented these inverse methods in MATLAB. We calculated crosstalk cancellation filters for several loudspeaker configurations. For each configuration, we varied the low-pass cut-off frequency, the filter length and the regularization constant. We evaluated the performance of each case by cascading the crosstalk cancellation filters $\mathbf{C}$ with the plant matrix $\mathbf{H}$:

$$\mathbf{R}(z) = \mathbf{H}(z)\mathbf{C}(z) \tag{13}$$

where $\mathbf{R}(z)$ is a 2×2 matrix. The diagonal elements represent the direct signals and the off-diagonal elements the crosstalk. The plant matrix $\mathbf{H}$ is composed by the transfer functions from the loudspeaker to the ears. In this study, we made use of the head related transfer functions (HRTF) measured on the artificial head designed at Aalborg University [7, 6][1]. For simplicity, we did not take into account the loudspeakers transfer function and assumed they were ideal and symmetric.

Ideally, $\mathbf{R}(z) = \mathbf{I}z^{-m}$, but since exact inversion is not possible we need to measure how close to ideal this approximation is. Thus, we used two indexes as performance indicators for each setup. The first index we used, is the channel separation index defined by Bai et al. in [2]. The channel separation is defined as the magnitude ratio between the direct signal and the crosstalk. This can be written in terms of the discrete frequency index k as follows:

$$CHSP_1(k) = \frac{R_{12}(k)}{R_{11}(k)}, \quad CHSP_2(k) = \frac{R_{21}(k)}{R_{22}(k)} \tag{14}$$

where $k = 1, \dots, N_c$. The channel separation index is then defined as the average over the frequency k of equation 14:

$$\overline{CHSP}_i = \frac{1}{n_f - n_j + 1} \sum_{k=n_j}^{n_f} 20\log_{10}(|CHSP_i(k)|) \quad [dB] \tag{15}$$

where n_j and n_f define the frequency range of interest. The selected frequencies k are distributed over a logarithmic frequency scale [2].

The second index we used in this study is the performance error which is a measure of the equalization performance. It is defined as the magnitude ratio between the output of the cascaded systems and the desired signals, which in this case are equivalent to unit sample functions delayed m samples:

$$PE_i(z) = \frac{z^{-m}}{R_{ii}(z) + R_{ji}(z)} \tag{16}$$

where R_{ii} represents the i^{th} diagonal components of the result matrix $\mathbf{R}$ (i.e. R_{11} and R_{22}) and R_{ji} are the crossterms (i.e. R_{21} and R_{12}). This is evaluated in the frequency domain with $z = e^{j\frac{2\pi}{N_c}k}$. Using the same approach as with the channel separation, we define the performance error index as an average over frequency k of the power of the performance error:

$$\overline{PE}_i = \frac{1}{n_f - n_j + 1} \sum_{k=n_j}^{n_f} 10\log_{10}\left(|PE_i(k)|^2\right) \quad [dB] \tag{17}$$

The frequencies k are also distributed over a logarithmic scale.

3.1. Loudspeaker Configurations and Simulations Setup

We divided the simulations into two main groups: two- and four-channel configurations. For each of these groups, we varied the span angle between the loudspeakers and the elevation angle with respect to the listener's head.

Figure 2 shows a simplified diagram of the two-channel configurations, where the span angle θ and the elevation

[1] The HRTFs are measured at the blocked ear canal.

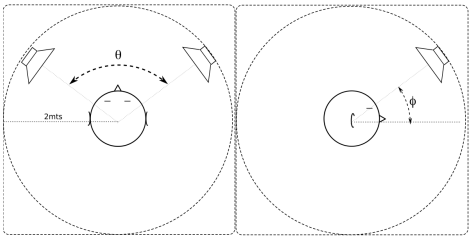


Fig. 2: Scheme of the two-channel configuration. The span angle θ and elevation angle ϕ are varied.

angle ϕ are varied. Table 1 describes the loudspeaker span angles and elevations evaluated in this case. We selected these positions so that all the evaluated points were approximately evenly spaced in the region corresponding to the half hemisphere in front and above of the listener's head. Additionally, only positions available in the HRTF database were used in order to avoid any further approximations such as interpolation.

Elevation (ϕ) [°]	Span (θ) [°]				
0	12	36	60	120	180
30	12	36	60	120	180
60	12	36	60	120	180
72	12	36	60	120	180
80	–	20	60	120	180
86	–	36	–	108	180

Table 1: Loudspeaker configurations evaluated for the two-channel case

Figure 3 shows the general diagram of the loudspeakers placement in the four-channel case. To ease the analysis, we separated these simulations into symmetric and asymmetric arrangements. This is, for the symmetric case $\theta_1 = \theta_2$ and for the asymmetric case $\theta_1 \neq \theta_2$.

Table 2 describes the loudspeakers positions evaluated for this configuration. Notice that for the asymmetric case, the values of θ_1 and θ_2 are equivalent to all the possible asymmetric combinations of the azimuth angles while the elevation angle remains symmetric. This result in a total of 212 different loudspeaker configurations including the two-channel case.

Since methods like the generic crosstalk cancellation are based on low-frequency approximations, it becomes nec-

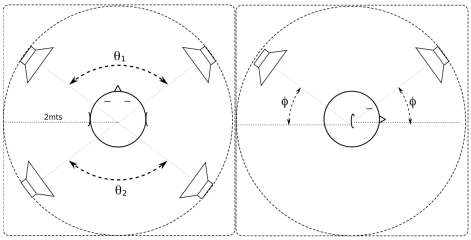


Fig. 3: Scheme of the four-channel configurations for symmetric ($\theta_1 = \theta_2$) and asymmetric ($\theta_1 \neq \theta_2$) arrangements. The elevation angle ϕ is kept symmetric.

Elevation (ϕ) [°]	Span (θ_1, θ_2) [°]					
0	12	36	60	88	120	160
30	12	36	60	88	120	160
60	12	36	60	90	120	156
72	12	36	60	84	120	156
80	–	20	60	80	120	160
86	–	36	72	–	108	144

Table 2: Loudspeakers span angles and elevations evaluated with the four-channel configurations. Symmetric case: $\theta_1 = \theta_2$, Asymmetric case: $\theta_1 \neq \theta_2$. The asymmetric case includes all possible combinations of θ_1 and θ_2 at each elevation angle.

essary to band limit the inverse filter. Therefore, to be able to compare the methods directly we applied a 16^{th} order low-pass FIR filter to all the crosstalk cancellation filters (i.e. $C_{11} \dots C_{ij}$). The filter was calculated using the linear phase windowing method. The causality achieved by introducing a modeling delay in the crosstalk cancellation filters can be destroyed if the delay of the low-pass filter is comparable to the modeling delay, i.e. $N_c/2$. Thus, the length of the low-pass filter was chosen such that the total delay introduced is negligible in comparison with the modeling delay. Figure 4 shows the magnitude and phase response of the low-pass filter, with a cut-off frequency of 8 kHz.

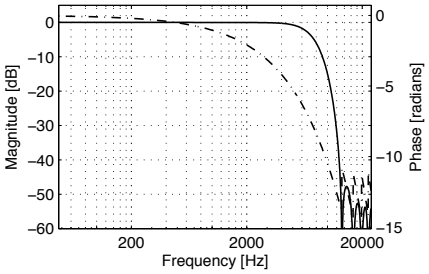


Fig. 4: Magnitude (solid line) and phase response (dashed line) of the FIR low-pass filter convolved with all the crosstalk cancellation filters.

In the two-channel case we used a digital FIR high-pass filter of order 16 as shape factor. The filter was also calculated with the linear phase windowing method and the cut-off frequency was set to 6 kHz. We chose this shape, because it is in this region where most of the singularities occur when inverting a two-channel system. Figure 5 depicts the shape factor used in this study.

In the four-channel case, we chose a frequency independent regularization. The reason for this is that we found that when introducing more than two channels in the plant matrix **H**, singularities occur all over the frequency range. To derive a proper frequency dependent regularization in this case, a careful analysis of the condition of the plant matrix is necessary for each simulated configuration. Thus, considering the large amount of configurations and variables, we decided to used a frequency independent regularization.

We evaluated the performance of the three inverse methods, for each of the 212 loudspeaker configurations, by

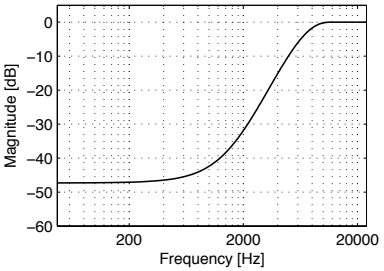


Fig. 5: Magnitude response of the 16^{th} order FIR high-pass filter used as shape factor. The linear-phase windowing method is used and the cut-off frequency is set to 6 kHz.

varying different design parameters. The parameters varied were the low-pass filter cutoff frequency f_o for the two-channel configuration, the filter length N_c and the regularization constant β. All possible combinations were evaluated in order to be able to analyze the interaction between the parameters and their influence on the system performance with respect to loudspeaker position. We did not vary the low-pass cut-off frequency for the four-channel case, given that preliminary results showed us a similar trend to the one observed with the two-channel case. Table 3 summarizes the parameters varied for each configuration and the values used.

Configuration	**2-channel**	**4-channel**
Inverse Method	GCC[a]	
	LS_f[b], LS_t[c]	LS_f, LS_t
f_o **[kHz]**[d]	6, 8,10,14	14
$\mathbf{N_c}$	128, 256, 512, 1024	128, 256,512
Shape Factor	16^{th}-order FIR[e]	Constant
(B)	High-pass, 6 kHz	
β^f	$10^{[-10:1:0]}$	$10^{[-8:1:0]}$

[a] Generic Crosstalk Canceler
[b] Least Square Approach in Frequency Domain
[c] Least Square Approach in Time Domain
[d] Cut-off frequency of the low-pass filter
[e] Calculated with linear phase windowing method (see figure 5)
[f] Vectors are displayed in MATLAB notation

Table 3: Summary of the methods evaluated and the parameters varied for each loudspeaker configuration.

3.2. Choice of Optimal Regularization

Here we define the optimum regularization constant β_{opt} as the minimum value at which the channel separation index is not greater than the minimum channel separation index plus 2 dB:

$$\beta_{opt} = min\Big\{\beta : \overline{CHSP}(\beta) \leq min\{\overline{CHSP}(\beta)\} + 2\Big\} \quad (18)$$

Figure 6 shows an example of the channel separation index as a function of regularization constant. The value of β that results in the minimum channel separation is 0.01. However, for a value of $\beta = 10^{-10}$ the result is not significantly larger than the minimum. Thus, instead of choosing the value of β that results in the minimum channel separation index, we choose the smallest β that is within 2 dB from the best performance. The reason for this is that the larger the β the more errors are obtained at the frequencies where the regularization has effect. This is why we consider the defined β_{opt} to be a better compromised between performance error and channel separation.

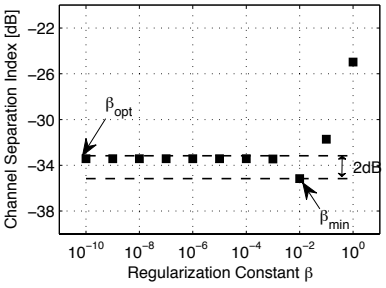


Fig. 6: Example of the channel separation index as a function of the regularization constant β evaluated for the left ear (GCC at $(30^\circ, 0^\circ)$, $N_c = 256$, $f_o = 8$ kHz)

From the figure it could be argued that the value of β is not critical, since with $\beta = 1$ the channel-separation index is still below -20 dB. Notice however that this analysis is based on ideal conditions and the same HRTFs used to calculate the filters are used in their evaluation. Thus, under such ideal conditions a channel separation index of -20dB should not be considered sufficient, since in a real situation this value is expected to increase considerably.

Notice also that the criterion used to choose β does not take into account the gain of the filters and a much more optimal value could be obtained by iterating between filter gain, performance error and channel separation. That is a topic for further research.

4. RESULTS

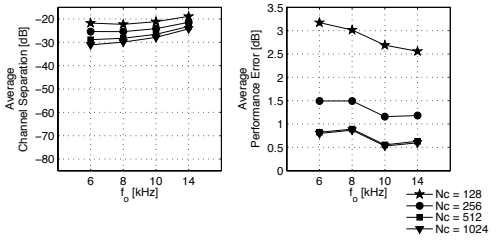
We calculated the channel separation index and the performance error index for the three aforementioned methods applied to the different loudspeaker configurations and filter design parameters. The indexes were calculated in the frequency range between $n_i = 100$ Hz and $n_f = f_o$ (see equations 17 and 15). At very low frequencies the crosstalk cancellation filters require special attention regarding allocated dynamic range and gain [26]. Thus, for convenience we set the lower frequency limit to 100 Hz.

4.1. Two-channel Configuration

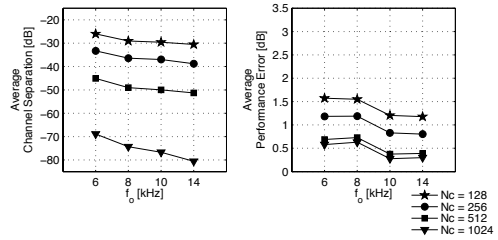
First, we evaluated the average performance of the three different methods when varying the low-pass filter cutoff frequency. This was done by estimating the overall means of the indexes defined before for each method and filter length. Figure 7 shows the average channel separation and the performance error indexes between all the two-channel loudspeaker configurations. Here we present the calculated values at the left ear. We can observe a clear difference in performance between the generic crosstalk canceler and the least square methods. The average channel separation index levels for the generic crosstalk canceler lie above -30 dB for all filter lengths (figure 7(a)), whereas the least square methods achieve values down to -80 dB (figures 7(b) and 7(c)). Additionally, we can notice that for the generic crosstalk canceler there is a slight decrease in channel separation when a bandwidth larger than 8 kHz is used. In contrast, the least square methods show an improvement when increasing the bandwidth. However, this improvement is insignificant for filters shorter than 512 and cut-off frequencies larger than 8 kHz.

All the methods show an improvement on channel separation index when increasing the filter length. Though the improvements observed in least square methods are dramatically larger than the ones observed with the generic crosstalk canceler.

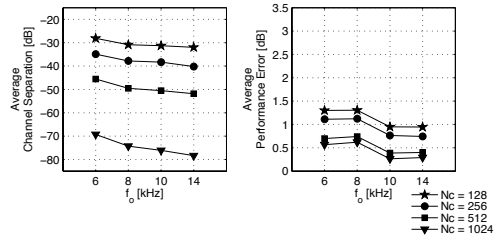
Regarding the performance error, the generic crosstalk canceler presents larger errors than the least square methods with short filters ($N_c = 128$). However, the three methods follow a similar trend: a slight improvement when increasing bandwidth from 8 kHz to 10 kHz. We presume this is due to interaction between the shape factor and the low-pass filter in this frequency range, which is more pronounced at low-pass frequencies below 8 kHz.



(a) Generic Crosstalk Canceler



(b) Least Square in the Frequency Domain



(c) Least Square in the Time Domain

Fig. 7: Estimated means over all loudspeakers positions at the left ear for each method simulated with different low-pass filter cutoff frequency and filter lengths.

Notice that the performance error is a measure of the magnitude squared of errors in the direct signal. This means that for example a performance error of 3 dB indicates that the direct signal has approximately twice as many errors as a signal with a performance error below 1 dB. However, bear in mind that the index presented here is an average over frequency, so it is thus just an indicator of the energy of the errors in the signal.

Even though there are variation of the results when increasing bandwidth, we observed in preliminary simulations that the three methods follow a similar pattern for all bandwidths with regards to loudspeaker position. Additionally, we noticed in the results presented in figure 7 that at frequencies larger than 8 kHz the performance of the generic crosstalk canceler decreases. Thus, in order to make a fair comparison of the methods, we will only present results obtained with a low-pass filter with cutoff frequency of 8 kHz for the two-channel configurations.

The following describes the detailed results for each inverse method and loudspeaker configuration. During the data analysis, we observed some variations in the results obtained with the different simulated elevation angles. However, given the large amount of configurations evaluated, we are only presenting the results obtained with three different elevations ($\phi = 0°, 60°, 86°$). We found these elevation angles to be the most representative of the observed patterns.

4.1.1. Generic Crosstalk Canceler

Figure 8 shows the general performance of the generic crosstalk canceler method as a function of loudspeaker position. To ease the analysis, the plots are divided into three columns each one corresponding to the selected elevation angles (ϕ). The labels on the x axis corresponds to the span angles θ.

The channel separation for the generic crosstalk canceler is above -30 dB for most positions. However, at an elevation angle of $\phi = 60°$ the generic crosstalk canceler achieves channel separation values below -30 dB for span angles of $\theta = 60°$ and $\theta = 120°$ and filter lengths larger and equal to 256. What's more, we observe channel separation indexes below -30 dB for the span angle of $\theta = 60°$ in most of the evaluated elevations. In general, the channel separation increases with longer filters.

The crosstalk cancellation filters with lengths 512 and 1024 result in nearly the same performance error. Short filters (128 taps) exhibit the largest performance errors

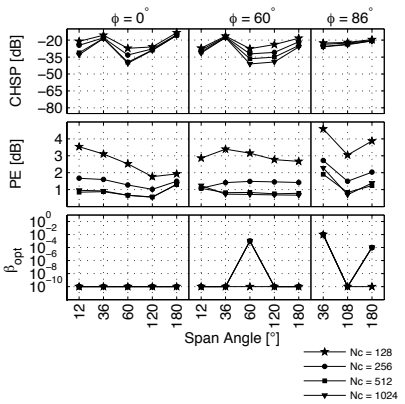


Fig. 8: Performance of the generic crosstalk canceler at the left ear as a function of loudspeakers position ($f_0 = 8$ kHz). The values of the channel separation index (CHSP), the performance error index (PE) and the optimum regularization β_{opt} are displayed.

($\overline{PE} \geq 2$ dB). Especially at an elevation of $\phi = 86°$ the performance error of the 128 taps filters reach values above 4dB.

We also observed a large variation of the regularization constant with respect to position. When the loudspeakers are located on the horizontal plane the regularization needed is $\beta_{opt} = 10^{-10}$ for all filter lengths. At elevated positions more regularization is needed for most span angles. However, there is no clear pattern of the optimum regularization constant as a function of loudspeaker configuration or filter length.

4.1.2. Least Square Methods

Figures 9 and 10 show the simulation results obtained with the least square methods in the frequency and time domain respectively.

As we expected, both methods result in similar values. Both methods result in channel separation indexes below -30 dB at all positions when the filter length is larger and equal to 256 taps. The span angle of $\theta = 120°$ in the horizontal plane exhibit a significantly better channel separation in comparison to the other evaluated span angles with both methods. We can also observe a decrease in channel separation of about 10 dB with the span angles of $\theta = 12°$ and $\theta = 36°$ placed in the horizontal plane compared to elevated positions.

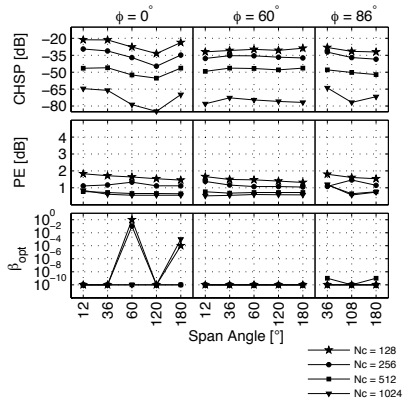


Fig. 9: Performance of the least square method in frequency domain at the left ear as a function of loudspeakers position applied to the two-channel configurations ($f_0 = 8$ kHz). The values of the channel separation index (CHSP), the performance error index (PE) and the optimum regularization β_{opt} are displayed.

When the filter length is set to 1024 taps, we notice a large increase in the channel separation. The increase in channel separation when increasing the filter length from 512 to 1024 taps, is almost twice the increase in channel separation when increasing the filter length from 256 to 512. In other words, the channel separation increases proportionally to the filter length.

Similarly to the generic crosstalk canceler, the performance error does not vary with position but in contrast it lies in most instances below 2 dB. Particularly, the least square in the frequency domain shows errors below 2 dB at all positions and for all filter lengths. Additionally, the required regularization is almost constant for most loudspeakers positions and filter lengths. A regularization constant of 10^{-10} is in general sufficient to obtain the channel separation and performance error indexes displayed in the figures.

4.1.3. Discussion

We could observed that the least square methods outperform the generic crosstalk cancellation. The least

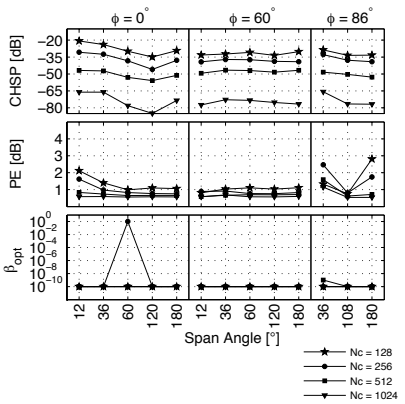


Fig. 10: Performance of the least square method in time domain at the left ear as a function of loudspeakers position applied to the two-channel configurations ($f_0 = 8$ kHz) The values of the channel separation index (CHSP), the performance error index (PE) and the optimum regularization β_{opt} (see equation 18) are displayed.

square approaches displayed a rather stable performance when varying the position of the loudspeakers whereas the generic crosstalk canceler revealed a large variance with respect to loudspeaker position. This was expected because the least square methods uses all the characteristics of the transfer functions (in this case HRTFs) to find a result that best approximates the ideal response. In contrast, the generic crosstalk canceler makes use of assumptions that do not take into account the phase characteristics of the plant transfer functions. The principle of crosstalk cancellation is that sound waves cancel each other at the contralateral ear of the listener. Therefore any approximation of the phase can easily result in waves not canceling each other properly. By doing a minimum-phase approximation, we are changing the phase of the plant transfer functions considerably. When we approximate the excess-phase components of the transfer functions to a single frequency independent delay and only invert the minimum-phase section, we are introducing phase distortion into the inverse system which can manifest in poor performance at specific frequencies and positions.

On the other hand, at span angles of 60° the generic crosstalk canceler exhibits channel separation indexes below -30 dB. Previous studies have shown that at this particular angle the plant matrix $\mathbf{H}$ is generally robust to small errors for most of evaluated the frequency range [2, 22]. One hypothesis is that if the plant matrix is well conditioned in most of the working frequency range, the generic crosstalk canceler can perform similarly to the least square methods. We should also point out that the channels separation indexes obtained with the generic crosstalk canceler are around and below -20 dB for most loudspeaker configurations, which have been found to be an acceptable level [16]. Nevertheless, the results discussed in this paper were obtained under ideal conditions. In a real application, we expect the channel separation levels to be much larger than the ones presented here.

We also observed that the performance of the generic crosstalk canceler decreases with respect to bandwidth. This was expected since this method is based on a low-frequency approximation and its accuracy is then bandlimited [9]. In contrast, the least square methods showed an improvement in performance with bandwidth.

The regularization constant did not show any clear dependance on filter length for the generic crosstalk canceler, yet it showed a high variance with respect to loudspeaker position and filter length. In general, positions located on the horizontal plane require little regularization, whereas at elevated positions larger regularization values are needed for most span angles. In contrast, the least square methods showed little variance with respect to loudspeaker positions and we observed that rather small regularization values were sufficient to avoid singularities and obtain a large channel separation. Yet, the optimum regularization was chosen such that it yields an optimum channel separation but the gain of the inverse filters was disregarded.

4.2. Four-channel Configuration

We observed earlier in the two-channel case an increase in performance when increasing bandwidth for the least square methods. Preliminary tests showed that both methods follow a similar pattern when varying the bandwidth. Therefore, in order to include the high-frequency characteristics of both methods, we present results obtained with filters low-passed at 14 kHz.

As mentioned before, the regularization was made frequency independent and its lower limit was set to 10^{-8}. We found this value to be the minimum regularization value needed in order to counteract the singularities of the plant matrix. Furthermore, given that the size of the plant matrix used for the least square method in time

domain (see equation 10) becomes extremely large with four-channel configurations, the maximum filter length was set to 512 taps due to processing and memory limitations.

4.2.1. Symmetric Arrangement

Figures 11 shows the results obtained with the least square method in the frequency domain applied to the four-channel configurations in symmetric arrangements. Just like with the two-channel configurations, we are only presenting here the results obtained at elevation angles $\phi = 0°, 60°$ and $86°$.

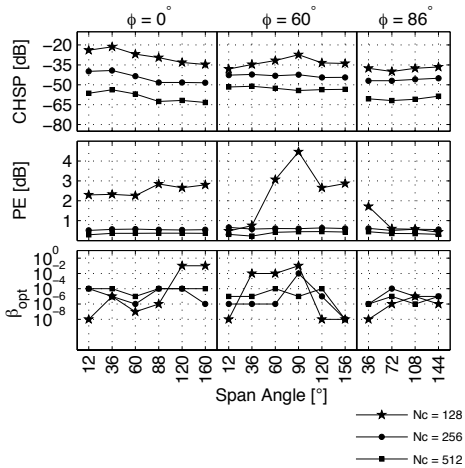


Fig. 11: Performance of the least square method in frequency domain at the left ear with respect to loudspeaker position applied to four-channel configurations in symmetric arrangement ($f_0 = 14$ kHz) The values of the channel separation index (CHSP), the performance error index (PE) and the optimum regularization β_{opt} are displayed.

In the horizontal plane, the channel separation index lies above -30dB for span angles up to $\theta = 88°$ when using short filters ($N_c = 128$) and improves to approximately -35dB with span angles of $\theta = 120°$ and $\theta = 160°$.

When the loudspeakers with span angles of $\theta = 12°$ and $\theta = 36°$ are placed at elevated positions, there is an improvement in channel separation. However, this does not occur with the 512 taps filters. Interestingly, at an elevation angle of $60°$ and span angles larger than $\theta = 36°$, the channel separation indexes obtained with the 256 and 512 taps filters are larger in comparison with the indexes observed with other elevation angles.

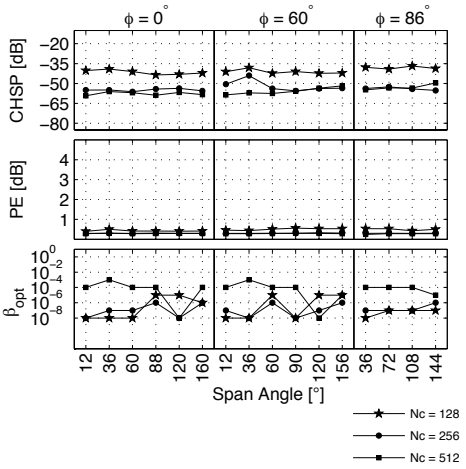


Fig. 12: Performance of the least square method in time domain at the left ear with respect to loudspeaker position applied to four-channel configurations in symmetric arrangement ($f_0 = 14$ kHz) The values of the channel separation index (CHSP), the performance error index (PE) and the optimum regularization β_{opt} are displayed.

Figure 12 illustrates the results obtained with the least square in the time domain with the four-channel configurations in symmetric arrangements. We can see that in general, the least square in the time domain results in larger channel separation than the frequency domain, especially with short filters. Additionally, the filters with 256 and 512 taps result in approximately the same performance for most configurations. This method shows a rather small variance with respect to loudspeaker position in comparison to the frequency domain approach.

As expected, with short filters ($N_c = 128$) the frequency domain approach shows performance errors significantly larger than the time domain approach. The latter presents errors below 1 dB, approximately constant for all positions and filter length.

Regarding the optimum regularization constant β_{opt}, there is a clear variation with respect to loudspeaker position

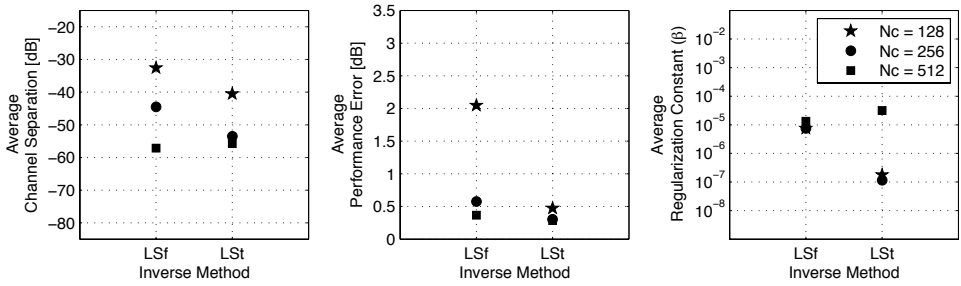


Fig. 13: Average over all symmetric loudspeaker arrangement evaluated at the left ear for the least square method in the frequency domain (LSf) and in the time domain (LSt) ($f_0 = 14$ kHz)

for both methods. Nevertheless, we do not observe any particular trend for the frequency domain approach. Yet, we can see that the time domain approach requires regularization values of $\beta_{opt} = 10^{-4}$ for most loudspeaker arrangements with filter lengths of 512 taps.

Figure 13 displays the average over all positions of the channel separation, the performance error and regularization constant for both methods. We can observe that with a filter length of 512 taps both methods perform similarly. Nevertheless, with shorter filters the time domain approach shows a better performance than the frequency domain, particularly with 128 taps filters.

The time domain approach requires in general less regularization than the frequency domain approach with filter lengths of 256 and 128 taps. However, we can see that with 512 taps filters the time domain approach needs larger regularization values than with shorter filter.

4.2.2. Asymmetric Arrangement

With the asymmetric arrangements we evaluated in total 152 loudspeakers positions including all elevations. The results obtained with the different asymmetric configurations follow similar patterns. Thus, for convenience, we are only presenting the results obtained when the frontal loudspeakers span angle is 12°. In the case of the 86° elevation angle, we are presenting the results obtained with a frontal span angle of 36°.[2]

Figure 14 shows the results obtained with the least square in frequency domain. Each plot is divided into columns,

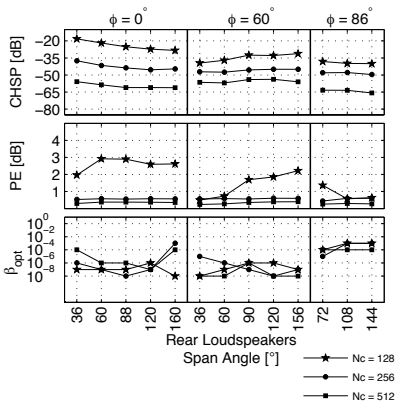


Fig. 14: Performance of the least square method in the frequency domain at the left ear with respect to loudspeaker positions applied to asymmetric configurations at different elevation angles ($f_0 = 14$ kHz). Left panel: $\theta_1 = 12°, \phi = 0°$ (Horizontal plane). Central panel: $\theta_1 = 12°, \phi = 60°$. Right panel $\theta_1 = 36°, \phi = 86°$. The values of the channel separation index (CHSP), the performance error index (PE) and the optimum regularization β_{opt} are displayed.

[2]The choice of this span angle is due to the resolution of the HRTF database used (see table 2).

each one corresponding to the selected elevation angles ($\phi = 0^\circ, 60^\circ$ and 86°). The x axis corresponds to the span angle of the rear loudspeakers (θ_2).

In the horizontal plane, the frequency domain approach achieves generally a better channel separation when the rear loudspeakers are positioned at wide angles ($\theta_2 \geq 80^\circ$). However, when increasing elevation the channel separation becomes rather constant for all arrangements. With filter lengths of 128 taps, the performance improves significantly when the loudspeakers are placed at elevated positions. This improvement is of approximately 10 dB for most configurations.

The performance error is below 1 dB for the filters with 256 and 512 taps. With respect to the regularization constant, we can see similar characteristics to the symmetric case. This is, there is a large variation with respect to loudspeaker position, but no clear pattern is observed.

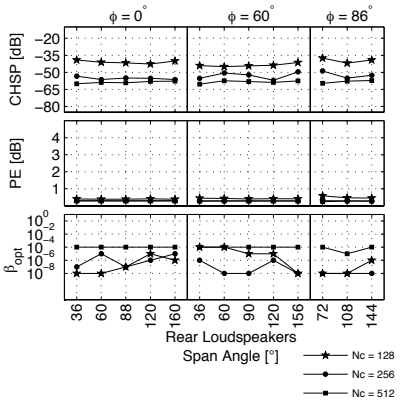


Fig. 15: Performance of the least square method in the time domain at the left ear with respect to loudspeaker position applied to asymmetric configurations at different elevation angles ($f_0 = 14$ kHz). Left panel: $\theta_1 = 12^\circ, \phi = 0^\circ$ (Horizontal plane). Central panel: $\theta_1 = 12^\circ, \phi = 60^\circ$. Right panel $\theta_1 = 36^\circ, \phi = 86^\circ$. The values of the channel separation index (CHSP), the performance error index (PE) and the optimum regularization β_{opt} are displayed.

Figure 15 presents the results obtained with the time domain approach. Similar to the symmetric configuration, there is little variance with respect to loudspeaker arrangement at all elevations. In general, the channel separation index lies below -40 dB and the performance error is smaller than 1 dB for all configurations and filter lengths. Then again, the channel separation index difference between filter lengths of 256 and 512 is not significant in some of the cases.

The regularization constant presents a large variation with respect to arrangement. Likewise the symmetric arrangements, the 512 taps filters requires in general regularization values of 10^{-4} for most positions and elevations.

Figure 16 shows the average channel separation, performance error and regularization over all arrangement for both methods. The results are very similar to the ones obtained with the symmetric case. In general, the time domain approach outperforms the frequency domain method.

4.2.3. Discussion

The least square method in the time domain exhibit a better performance than the frequency domain approach when short filters are used. These results agree with the ones presented by Bai et. al in [3]. In their study, they evaluated the channel separation of both methods for different filter lengths for two closely spaced loudspeakers. However, our results show that even though this is the case for most positions on the horizontal plane, for elevated positions the channel separation of the frequency domain approach improves significantly. This is especially noticeable at asymmetric configurations, were the least square in the frequency domain approach yields values similar to the values obtained with the time domain method.

In contrast to the frequency domain approach, the performance of the least square method in the time domain does not vary significantly with loudspeaker arrangement. As a whole, it shows channel separation indexes below -40 dB and performance errors smaller than 1 dB for most angles and filter lengths.

On the other hand, the regularization constant showed a large variability with respect to arrangement on both methods. Yet, with the time domain approach we observed a particular trend for the 512 taps filters: it remains approximately constant with position and is in most cases larger than the required values for shorter filters. The latter is expected since longer filters are more accurate at high frequencies, which make the system more ill-conditioned.

5. CONCLUSIONS

This paper documents a study of three different crosstalk

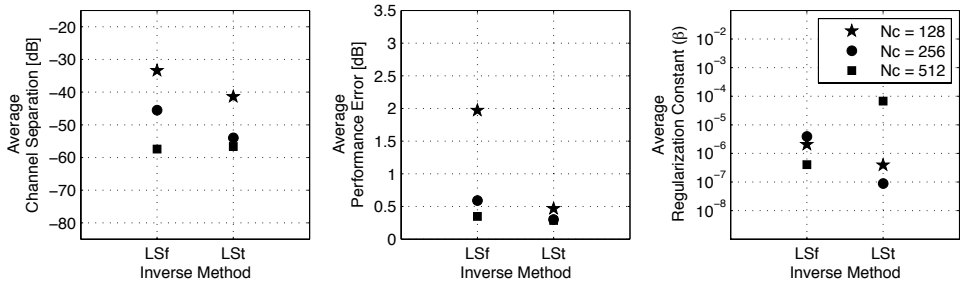


Fig. 16: Average over all asymmetric loudspeaker arrangement evaluated at the left ear for the least square method in the frequency domain (LSf) and in the time domain (LSt) ($f_0 = 14$ kHz)

cancellation methods applied to 212 different loudspeaker configurations, including two- and four-channel arrangements. The first method consists of a generic crosstalk canceler that is described in terms of the interaural transfer functions (ITF). The ITF is modeled as the ratio of the minimum-phase components of the ipsilateral and contralateral transfer function. The excess-phase components are modeled as a frequency independent delay. The second and third methods are based on least square approximations in the frequency and time domain respectively. A frequency dependent regularization was applied to all the methods with the two-channel configurations, while a frequency independent regularization was used with the four-channel arrangements. We implemented the methods in MATLAB and varied the filter length and the regularization constant for all configurations. We also varied the bandwidth of the filters for the two-channel case. In this paper we documented the results obtained with some of the evaluated loudspeaker positions. The results are presented in terms of the channel separation index, performance error index and optimum regularization values.

The results presented here suggest that the phase distortion introduced by the minimum-phase approximation generates large errors at specific frequencies and loudspeaker arrangements. However, we observed that for span angles of 60° the generic crosstalk canceler result in channel separation below -30 dB. Additionally, we also noticed that at these specific positions less regularization is needed. It has been shown that the plant matrix is usually well conditioned in the frequency range evaluated when the loudspeakers span angle is 60° [2, 22]. This infers that the phase distortion introduced by the minimum-phase approximation has less influence on the inverse system when the plant matrix is well-conditioned.

Both least square methods outperformed the generic crosstalk canceler. We expected the frequency domain approach to have a poorer performance with short filters in comparison with the time domain approach. However, with the two-channel configurations we found no significant differences between both least square methods. Both methods showed performance errors smaller than 2 dB and channel separation indexes below -30 dB for most arrangements. The channel separation indexes increased with filter length and remained almost constant with respect to loudspeaker position.

In general, we observed that little regularization is required for most loudspeaker positions and filter lengths when applying the least square methods to the two channel-cases. Nevertheless, the analysis we presented in this paper does not take into account the gain of the filters. The shape factor used as regularization is mainly limiting the peaks and notches present above and around 4 kHz on the frequency response of the filters. In other words, by increasing the regularization we are decreasing the performance of the crosstalk cancellation filters at those frequencies. When small regularization values are used, the performance at high frequencies improves but at the cost of large gains at high frequencies in the inverse filters. This can also have some perceptual consequences, in the sense that with large gains audible distortions or coloration changes could be introduced in the reproduced signals. This was confirmed in a later experiment, which results will be presented in the near future.

In another study carried out by the authors, it was shown that insufficient channel separation can result in lateralization of the virtual image [16]. Additionally, these peaks and notches make the filters less robust to errors and misplacement of the listener. The choice of optimum regularization is thus a tradeoff between performance error, channel separation and filter gain.

There are significant differences between the least square methods with the four-channel configurations for both symmetric and asymmetric setups. In general, the time domain approach outperformed the frequency domain. While the first exhibited a rather constant behavior with respect to loudspeaker arrangement, the latter showed differences in performance with respect to elevation. With the frequency domain approach, the channel separation improves with elevation, especially when short filters ($N_c = 128$) are applied. Another interesting observation is that the performance of the time domain approach does not improve significantly when we increase the filter length from 256 taps to 512 taps in most cases. This can be a consequence of the efficient use of filter coefficients the time domain approach makes. Thus, we presume that the least square in the time domain already reaches its maximum performance at frequencies above 100 Hz with a filter length of 256 taps.

We found it necessary to use a frequency independent regularization with the four-channel configurations. We observed that in this case the singularities in the plant matrix are spread all over the evaluated frequency range. Thus, in order to derive a proper shape factor, a thorough analysis of the condition of the plant matrix is necessary for each configuration. Given the large amount of configurations we evaluated in this study, we considered convenient to keep the regularization frequency independent. We also found that the minimum regularization value needed to avoid singularities is 10^{-8} for this case.

With the four-channel configurations we also observed that the optimum regularization constant varied with loudspeaker arrangement. These values are between 10^{-4} and 10^{-8} for most loudspeaker positions, but for most configurations with a 512 taps filter length, the value remained almost constant ($\beta_{opt} \approx 10^{-4}$). Then again, the gain of the filters was disregarded in the analysis.

The results presented in this paper are averages over frequency. This means that peaks and notches present at specific frequencies could have been reduced by averaging. However, the indexes used here are meant to give us a general idea of the performance of the methods and their variations with respect to loudspeaker configuration.

6. PENDING ISSUES

In this paper we only analyzed the optimum parameters for the best-case scenario. In other words, the results we presented are valid as long as the HRTFs used to design the filters correspond to the listener's HRTFs and the exact position of the listener. Additionally, we analyzed the inverse methods under ideal conditions and the loudspeaker transfer functions were neglected. This is, we assumed zero-phase and flat transfer functions for the loudspeakers and only used the HRTFs to calculate the crosstalk cancellation filters. In a real situation the loudspeakers characteristics would play an important role given that their transfer functions are neither flat nor symmetric. In order to extend the results obtained in this study, we evaluated the robustness to errors of the three methods in different conditions. Results of this study can be found in [17] and [18].

7. REFERENCES

[1] Mingsian R. Bai and Chih-Chung Lee. Development and Implementation of Cross-Talk Cancellation System in Spatial Audio Reproduction Based on Subband Filtering. *Journal of Sound and Vibration*, 290:1269–1289, August 2005.

[2] Mingsian R. Bai and Chih-Chung Lee. Objective and Subjective Analysis of Effects of Listening angle on Crosstalk Cancellation in Spatial Sound Reproduction. *Journal of the Acoustic Society of America*, 120(4):1976–1989, October 2006.

[3] Mingsian R. Bai, Geng-Yu Shih, and Chih-Chung Lee. Comparative Study of Audio Spatializers for Dual-Loudspeaker Mobile Phone. *Journal of the Acoustic Society of America*, 121(1):298–309, January 2007.

[4] Mingsian R. Bai, Chih-Wei Tung, and Chih-Chung Lee. Optimal Design of Loudspeaker Arrays for Robust Cross-talk Cancellation Using the Taguchi Method and the Genetic Algorithm. *Journal Acoust. Soc. Am.*, 117(5):2802 – 2813, 2005.

[5] Stephen Barnett. *Matrices: Methods and Applications*. Oxford University Press, 1st edition, 1990.

[6] Bjarke P. Bovbjerg, Flemming Christensen, Pauli Minnaar, and Xiaoping Chen. Measuring the Head-Related Transfer Functions of an Artificial Head

with a High Directional Resolution. In *109th Convention of The Audio Engineering Society*, page 5264, Los Angeles, CA, September 22-25 2000.

[7] Flemming Christensen, Clemen Boje Jensen, and Henrik Møller. The Design of VALDEMAR - An Artificial Head for Binaural Recordings Purposes. In *109th Convention of The Audio Engineering Society*, page 5253, Los Angeles, CA, September 22-25 2000.

[8] Richard David Clemow. Method of Synthesizing a Three Dimensional Sound-Field. U. S. Patent 6,577,736, June 2003.

[9] William G. Gardner. *3-D Audio Using Loudspeakers*. Kluwer Academic Publishers, 1st edition, 1998.

[10] P. A. Hill, P. A. Nelson, O. Kirkeby, and H. Hamada. Resolution of Front-Back Confusion in Virtual Acoustic Imaging Systems. *Journal of the Acoustic Society of America*, 108(6):2901–2909, December 2000.

[11] Ole Kirkeby and Philip A. Nelson. The "Stereo Dipole" – A Virtual Source Imaging System Using Two Closely Spaced Loudspeakers. *Audio Engineering Society*, 46(5):387–395, May 1998.

[12] Ole Kirkeby and Philip A. Nelson. Digital Filter Design for Inversion Problems in Sound Reproduction. *Audio Engineering Society*, 47(7/8):583–595, July/August 1999.

[13] Ole Kirkeby, Philip A. Nelson, and Hareo Hamada. Local Sound Field Reproduction Using Two Closely Spaced Loudspeakers. *Journal of the Acoustic Society of America*, 104(4):1973–1981, October 1998.

[14] Ole Kirkeby, Philip A. Nelson, Hareo Hamada, and Felipe Orduna-Bustamante. Fast Deconvolution of Multi-Channel Systems Using Regularization. *IEEE Transactions on Speech and Audio Processing*, 6(2):189–195, 1998.

[15] Ole Kirkeby, Per Rubak, and Angelo Farina. Analysis of Ill-Conditioning of Multi-Channel Deconvolution Problems. In *Proceedings 1999 IEEE Workshop on Applications of Signal Processing to Audio*, New Paltz, New York, October 17-20 1999.

[16] Yesenia Lacouture Parodi and Per Rubak. A Subjective Evaluation of the Minimum Audible Channel Separation in Binaural Reproduction Systems through Loudspeakers. In *128th Convention of the Audio Engineerin Society*, London, UK, May 22 - 25 2010.

[17] Yesenia Lacouture Parodi and Per Rubak. Objective Evaluation of the Sweet Spot Size in Spatial Sound Reproduction Using Elevated Loudspeakers. *Journal of the Acoustical Society of America*, 128(3), September 2010.

[18] Yesenia Lacouture Parodi and Per Rubak. Sweet Spot Size in Virtual Sound Reproduction: A Temporal Analysis. In *Principles and Applications of Spatial Hearing*. World Scientific, in press.

[19] Improved Sound Separation Using Three Loudspeakers. Jun Yang and Woon-Seng Gan and See-Ee Tan. *Acoustic Research Letters Online-ARLO*, 4(2):47–52, April 2003.

[20] Henrik Møller. Reproduction of Artificial-Head Recordings Through Loudspeakers. *Audio Engineering Society*, 37(1/2):30–33, January/February 1989.

[21] John N. Mourjopoulos. Digital Equalization of Room Acoustics. *Journal of the Audio Engineering Society*, 42(11):884–900, November 1994.

[22] P. A. Nelson and J. F. Rose. Errors in Two-Point Sound Reproduction. *Journal of the Acoustic Society of America*, 118(1):193–204, July 2005.

[23] Philip A. Nelson, Felipe Orduña-Bustamante, and Hareo Hamada. Inverse Filter Design and Equalization Zones in Multichannel Sound Reproduction. *IEEE Transactions on Speech and Audio Processing*, 3(3):185 – 192, May 1995.

[24] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing*. Prentice-Hall, Inc., 1999.

[25] Takashi Takeuchi and Philip A. Nelson. Optimal Source Distribution for Binaural Synthesis Over Loudspeakers. *Journal of the Acoustic Society of America*, 112(6):2786–2797, December 2002.

[26] Takashi Takeuchi and Philip A. Nelson. Subjective and Objective Evaluation of the Optimal Source Distribution for Virtual Acoustic Imaging. *Journal of the Audio Engineering Society*, 55(11):981–997, November 2007.

[27] A. N. Tikhonov. Solution of Incorrectly Formulated Problems and the Regularization Method. *Soviet Mathematics DOKL*, 4:1035 – 1038, 1963.

[28] Darren B. Ward. Joint Least Squares Optimization for Robust Acoustic Crosstalk Cancellation. *IEEE Transactions on Speech and Audio Processing*, 8(2):211–215, February 2000.

[29] Darren B. Ward and Gary W. Elko. Effect of Louspeaker Position on the Robustness of Acoustic Crosstalk Cancellation. *IEEE Signal Processing Letters*, 6(5):105–108, May 1999.

[30] Bernard Widrow. *Adaptive Signal Processing*. Prentice-Hall, Inc., 1st edition, 1985.

APPENDIX

Linear Convolution Matrix

The convolution between two vectors can be calculated by a matrix-vector multiplication. First, let us convolve the two column vectors $\mathbf{a}$ and $\mathbf{b}$ with lengths N_a and N_b respectively. This convolution can be calculated as follows

$$\mathbf{a} * \mathbf{b} = \mathbf{A} \cdot \mathbf{b} \tag{19}$$

where $*$ denotes a discrete time linear convolution, $\mathbf{A}$ is an $N_b \times (N_a + N_b - 1)$ matrix of Toeplitz form, i.e. elements are identical along the diagonals, and it is known as linear convolution matrix:

$$\mathbf{A} = \begin{bmatrix} a(0) & 0 & & 0 \\ \vdots & a(0) & \ddots & \vdots \\ a(N_a-1) & \vdots & \ddots & 0 \\ 0 & a(N_a-1) & \ddots & a(0) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & a(N_a-1) \end{bmatrix} \tag{20}$$

3 MANUSCRIPT B

Lacouture Parodi, Y. and Rubak, P. (2010), **Objective Evaluation of the Sweet Spot Size in Spatial Sound Reproduction Using Elevated Loudspeakers**, *Journal of the Acoustic Society of America*, 128(3), 1045 - 1055, September, 2010.

Portions of this manuscript were presented in "Preliminary Evaluation of the Sweet Spot Size in Virtual Sound Reproduction Using Dipoles", Proceedings of the 126^{th} Convention of the Audio Engineering Society 2009, 7-10 May, Munich, Germany.

Objective evaluation of the sweet spot size in spatial sound reproduction using elevated loudspeakers[a)]

Yesenia Lacouture Parodi[b)] and Per Rubak
Section of Acoustics, Department of Electronic Systems, Aalborg University, Frederik Bajers Vej 7B5, 9220 Aalborg, Denmark

(Received 17 August 2009; revised 25 March 2010; accepted 1 July 2010)

In a previous study, three crosstalk cancellation techniques were evaluated and compared under different conditions. Least-squares approximations in the frequency and time domain were evaluated along with a method based on minimum-phase decomposition and a frequency independent delay. In general, the least-squares methods outperformed the method based on the minimum-phase decomposition. However, the evaluation was only done for the best-case scenario, where the transfer functions used to design the filters correspond to the listener's transfer functions and his/her location and orientation relative to the loudspeakers. This paper presents a follow-up evaluation of the performance of the three inversion techniques when the above mentioned conditions are relaxed. A setup to measure the sweet spot of different loudspeaker arrangements is described. The sweet spot was measured for 21 different loudspeaker configurations, including two- and four-channel setups. Lateral and frontal displacement were measured along with head rotations. The setups were evaluated at different elevation angles. The results suggest that when the loudspeakers are placed at elevated positions, a wider effective area is obtained. Additionally, the two-channel configurations showed to be more robust to head misalignments than the four-channel configurations. © *2010 Acoustical Society of America.* [DOI: 10.1121/1.3467763]

PACS number(s): 43.38.Md, 43.66.Pn, 43.60.Pt [MAH] Pages: 1045–1055

I. INTRODUCTION

With binaural technology it is possible to simulate a virtual environment where the listener perceives an acoustic source located at a position where no physical source exists. Based on the assumption that the sound pressures at the ears control the perception of the sound, we can in theory simulate any virtual source by employing appropriate filters. Headphones are often used for reproduction due to the perfect channel separation they provide. However, reproduction through headphones might not be practical or comfortable and in some cases can even reduce the virtual reality perception. Additionally, if head rotations are to be allowed, a head-tracker should be included so that the location of the source relative to the listener is updated properly. This can be an expensive solution and it requires extra computational power.

When reproducing binaural signals through loudspeakers, the signals that are to be heard in one ear are also heard in the other ear. This is known as crosstalk and it is possible to reduce it by adding the appropriate inverse filters into the reproduction chain. These techniques are known as crosstalk cancellation. However, the optimal filters are often designed for a static position. Thus, the overall effective area—also known as sweet spot—is still limited to a narrow region.

In the past decades, different crosstalk cancellation techniques have been developed.[1–5] The most commonly used are based on least-squares approximations. Some studies have also proposed different span angles and loudspeaker arrangements in order to improve the performance of the binaural reproduction by widening the sweet spot.[6–10] However, most of them have mainly focused on loudspeakers placed in the horizontal plane.

In a previous study carried out by the authors, three crosstalk cancellation techniques were evaluated.[11] The first one consisted of a generic crosstalk cancellation system based on a minimum-phase decomposition proposed by Gardner.[12] The other two methods are based on least-squares inversion, one in the frequency domain and the other in the time domain.[1,2]

However, the methods evaluated in that investigation were analyzed for the best-case conditions. In other words, the obtained results apply only when the head related transfer functions (HRTF) used to design the filters correspond to the listener's HRTFs and his/her location and orientation relative to the loudspeakers. Therefore, it is the intention of this paper to present a follow-up evaluation of the performance of the three inversion techniques when the head is displaced or rotated from its original position.

Previous studies have shown that closely spaced loudspeakers—the so-called stereo-dipole[7]—lead generally to a wider sweet spot than the standard stereo configuration.[6,13] One of the reasons for obtaining a wider sweet spot with closely spaced loudspeakers is that the variations of the distance between the listener and the loudspeakers are reduced when the loudspeakers are set close

[a)]Portions of this work were presented in "Preliminary evaluation of sweet spot size in virtual sound reproduction using dipoles," Proceedings of the 126th Convention of the Audio Engineering Society, Munich, Germany, 7–10 May 2009.

[b)]Author to whom correspondence should be addressed. Electronic mail: ylp@es.aau.dk

Author's complimentary copy

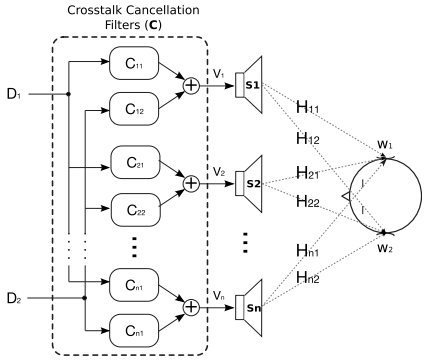


FIG. 1. Diagram of a binaural reproduction system through loudspeakers. The functions H_{ij} represents the transfer functions between loudspeakers and the ears (plant transfer functions). The signals D_j are the desired binaural signal to be reproduced, the functions C_{ij} are the crosstalk cancellation filters, V_i are the input signals to the loudspeakers and W_j are the signals which are reproduced at the listener's eardrums.

together.[10] If we now place the loudspeakers above the head, these distance variations would be even smaller for certain kind of movements, such as head rotations. However, those differences are also highly dependent on frequency due to the diffractions of the head, therefore we need experimental data to support such an argument.

Most of the investigations of the robustness to head misalignments of binaural reproduction systems through loudspeakers had focused on two-channel configurations placed in the horizontal plane. We present in this paper a study of two- and four-channel configurations, including the standard stereo configuration and closely spaced loudspeakers, placed at four different elevation angles. In total, 21 different loudspeaker configurations were evaluated. We measured the channel separation as a function of lateral displacement, frontal displacement and head rotation for each configuration. The three crosstalk cancellation techniques evaluated in our previous study were used to calculate the corresponding crosstalk cancellation filters.

This paper is organized as follows: Section II presents a short description of the methods used in this study; Sec. III describes the setup installed to carry out the measurements of the sweet spot of 21 different loudspeaker configurations; in Sec. IV we present and discuss the results obtained from the measurements, divided into lateral displacement, head rotation and frontal displacement; Sec. V contains a summary of the results and the conclusions drawn from this study.

II. CROSSTALK CANCELLATION TECHNIQUES

Figure 1 illustrates the transmission paths for a binaural reproduction system using loudspeakers, where the signals D_j are the desired binaural signals to be reproduced at the ears.

This system can be described in matrix form in the frequency domain ($z=e^{j\omega}$) as follows:

$$\underbrace{\begin{bmatrix} W_1(z) \\ W_2(z) \end{bmatrix}}_{\mathbf{w}(z)} = \underbrace{\begin{bmatrix} H_{11}(z) & H_{21}(z) & \dots & H_{n1}(z) \\ H_{12}(z) & H_{22}(z) & \dots & H_{n2}(z) \end{bmatrix}}_{\mathbf{H}(z)} \cdot \underbrace{\begin{bmatrix} V_1(z) \\ V_2(z) \\ \vdots \\ V_n(z) \end{bmatrix}}_{\mathbf{v}(z)}, \tag{1}$$

where $\mathbf{w}(z)$ is a vector which contains the sound pressures at the eardrums, $\mathbf{H}(z)$ is the matrix containing functions $H_{ij}(z)$ and $\mathbf{v}(z)$ is a vector containing the loudspeakers input signals $V_i(z)$. In this paper we will refer to the matrix $\mathbf{H}$ and the functions H_{ij} as the plant matrix and the plant transfer functions respectively. The loudspeakers input signals can be described as $\mathbf{v}(z)=\mathbf{C}(z)\mathbf{d}(z)$, where $\mathbf{C}$ is the matrix containing the crosstalk cancellation filters C_{ij} and $\mathbf{d}$ is a vector containing the desired binaural signals D_j. Note that the notations C and H used in this paper are opposite of those used in some other works on the subject. We used this notation given that it seemed more natural to us to use the letter C to denote the crosstalk cancellation filters and H to denote the transfer functions to the head.

Ideally, perfect crosstalk cancellation is obtained when $\mathbf{w}(z)=\mathbf{d}(z)$. This is

$$\mathbf{H}(z)\mathbf{C}(z)=\mathbf{I}, \tag{2}$$

where $\mathbf{I}$ is the identity matrix. A direct solution of Eq. (2) is usually not feasible, given that the matrix $\mathbf{H}$ is in general non-invertible at some frequencies due to inherent singularities. Furthermore, when reproducing binaural signals through more than two channels, the equation system becomes over determined. Consequently, we need to model the problem in a way that it best approximates the required solution.

There exists a number of methods to solve the equation system (2). The optimal solution is often obtained for a static position, e.g., for a listener placed at the nominal center position. Let us therefore first define the result matrix for the "best-case" scenario

$$\mathbf{R}_o(z)=\mathbf{H}_o(z)\mathbf{C}_o(z), \tag{3}$$

where $\mathbf{C}_o$ corresponds to the crosstalk cancellation filters and $\mathbf{H}_o$ corresponds to the plant matrix used to calculate $\mathbf{C}_o$ (e.g., transfer functions to the nominal center position). Ideally, $\mathbf{R}_o$ will be close to the identity matrix $\mathbf{I}$. However, in a real situation, the plant matrix usually differs from $\mathbf{H}_o$. A typical situation is when the listener moves the head. In this case, the larger the displacement or rotation the more the plant matrix differs from $\mathbf{H}_o$.

In this paper we evaluated the matrix $\hat{\mathbf{R}}$, which is the result of a plant matrix $\hat{\mathbf{H}}$ multiplied by the crosstalk cancellation filters $\mathbf{C}_o$. The matrix $\hat{\mathbf{H}}$ corresponds to a displaced or rotated position with respect to $\mathbf{H}_o$

$$\hat{\mathbf{R}}(z)=\hat{\mathbf{H}}(z)\mathbf{C}_o(z). \tag{4}$$

Previously, we carried out a study in which three crosstalk cancellation methods were evaluated under ideal conditions.[11] The first one, which we will refer to as the generic crosstalk canceler, is proposed by Gardner in Ref. 12

Author's complimentary copy

and consists of an exact solution approach. The second and third methods are based on least-squares approximation, one in the frequency domain and the other in the time domain. In the following we present a short introduction to these techniques; a more detailed description can be found in Ref. 11.

A. Generic crosstalk canceler

The generic crosstalk canceler (GCC) is based on the exact matrix inversion. Applying the standard method to calculate the direct inverse of $\mathbf{H}$ to solve Eq. (2) and rewriting the expression in terms of the interaural transfer functions (ITF),[12,14] we obtain

$$\mathbf{C}(z) = \frac{1}{\hat{D}} \begin{bmatrix} \frac{1}{H_{11}(z)} & 0 \\ 0 & \frac{1}{H_{22}(z)} \end{bmatrix} \begin{bmatrix} 1 & -ITF_2(z) \\ -ITF_1(z) & 1 \end{bmatrix}, \tag{5}$$

where $\hat{D}=1-ITF_1(z)ITF_2(z)$ and the terms $ITF_1(z)=H_{12}(z)/H_{11}(z)$ and $ITF_2(z)=H_{21}(z)/H_{22}(z)$ are the interaural transfer functions for the left and right channel, respectively. Since the functions H_{ii} are in general non-minimum phase, the direct implementation of Eq. (5) is not possible. Gardner proposes to model the ITF as[12]

$$ITF_i(z) \cong \frac{H_{ij}(z)_{minphase}}{H_{ii}(z)_{minphase}} z^{-ITD/T}, \quad j \neq i, \tag{6}$$

where H_{ij} and H_{ii} are the contralateral and ipsilateral transfer functions respectively. The ITD is the frequency independent interaural time delay and T is the sampling period. This frequency independent delay is calculated based on the assumption that the all-pass section of the transfer functions are approximately linear phase below 6 kHz. Note that this method is only feasible when $\mathbf{H}$ is square since an exact matrix inversion only holds for square matrices. Thus, this method is only applicable for two-channel configurations.

B. Least-squares approximations

The least-squares approximations are the most commonly used methods for designing crosstalk cancellation filters. In contrast to the generic crosstalk cancellation, these method do not try to find the exact solution but the best approximation which results in minimum errors (in the least-squares sense). Employing Tikhonov regularization,[15] the central idea of this filter design algorithm is to minimize a quadratic cost function of the type

$$J = E + \beta V = \mathbf{e}^H(z)\mathbf{e}(z) + \beta \mathbf{v}^H(z)\mathbf{v}(z), \tag{7}$$

where E is a measure of the performance error $\mathbf{e}=\mathbf{d}-\mathbf{w}$ and V is a measure of the power of the signals $\mathbf{v}$, which are the input signals to the loudspeakers. The superscript H is the Hermitian transpose operator. The positive real number β is the so-called regularization parameter that determines how much weight is assigned to the term V.[1,15] For the crosstalk cancellation problem described in Eq. (2) the total cost function J is minimum when

$$\mathbf{C}(z) = [\mathbf{H}^H(z)\mathbf{H}(z) + \beta \mathbf{I}]^{-1}\mathbf{H}^H(z)z^{-m}, \tag{8}$$

where m is the modeling delay used to ensure that the crosstalk cancellation network is causal and performs well not only in terms of amplitude, but also in terms of phase.

If the regularization parameter β is frequency dependent, it is convenient to express it as the product of a gain factor β and shape factor $B(z)$. The shape factor is a digital filter, which suppresses the frequencies we do not want to alter with the regularization. Frequency dependent regularization is used mainly to prevent sharp peaks in the magnitude response of the optimal filters. In this way, it is possible to control the dynamic range of the filters. The algorithm works optimal when a gain factor and a shape factor are set appropriately. Using this approach, Eq. (8) can be rewritten as follows:

$$\mathbf{C}(z) = [\mathbf{H}^H(z)\mathbf{H}(z) + \beta \mathbf{B}^H(z)\mathbf{B}(z)]^{-1}\mathbf{H}^H(z)z^{-m} \tag{9}$$

We implemented this method in the frequency and the time domain (in the analysis referred to as LS_f and LS_t respectively). The first one is also known as the fast deconvolution method[1] and is based on the Fast Fourier Transform (FFT). It is simple to implement and very efficient for both single- and multi-channel deconvolution. However, given that circular convolution artifacts are not easily avoided when dividing two FFTs sequences,[16] the designed filters will normally be longer than the corresponding filter obtained with the time domain approach.[17] Yet, by using frequency dependent regularization we can control the time response of the optimal filters. The regularization parameter will mainly influence the poles closest to the unit circle. Thus, by increasing the regularization we can push the poles further away from the unit circle, shortening the length of the inverse filters.

The least-squares method in the time domain approaches the problem in a similar way as Eq. (9), though instead of calculating the inverse filters by inverting each single frequency it uses a set of matrices composed of convolution matrices.[11,18] These matrices can be quite large. The time domain approach is thus more cumbersome than the frequency domain approach with regard to filter calculation. Nevertheless, this method is known to make efficient use of the available filter coefficients,[2] and therefore it is usually convenient when short filters are needed.

III. OBJECTIVE MEASUREMENTS

We carried out the objective measurements in an anechoic chamber at the acoustics laboratory at Aalborg University. The measurements were done on Valdemar, the artificial head designed at Aalborg University.[19]

Valdemar was placed in the center of two half circular arcs of 170 cm radius. The arcs were mounted on two supporting poles and it was possible to rotate them [see Figs. 2(a) and 2(b)]. Thus, we were able to cover the area corresponding to the hemisphere above Valdemar's head.

Four loudspeakers with 70 mm Vifa M10MD-39 drivers mounted in 155 mm diameter hard plastic balls, were mounted onto sliding devices connected to the arcs. This made it possible to place the loudspeakers in different span

Author's complimentary copy

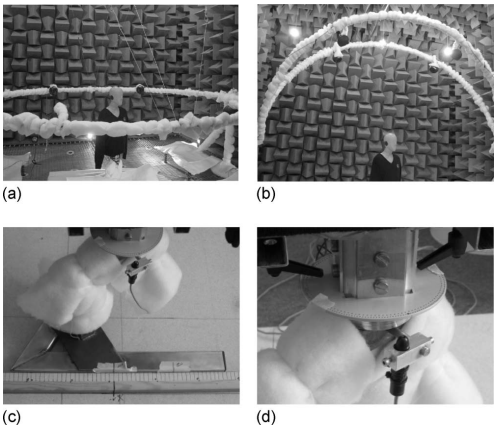


FIG. 2. Setup used to measure the sweet spot size of different loudspeaker configurations. Four ball loudspeakers are mounted on two half circular arcs. (a) Loudspeakers placed at 0° elevation (horizontal plane), (b) Loudspeakers placed at 60° elevation, (c) Valdemar's base and wooden bar used to control lateral and frontal displacement, (d) Rotating base used to control rotations.

angles. The physical center of the loudspeakers was placed at 150 cm distance from the center of the circumference.

The base of Valdemar's stand has a triangular shape. This was mounted on a wooden plate and put together to a straight wooden bar [Fig. 2(c)]. In that way we could constrain movements to one direction and measure lateral and frontal displacements. The manikin was also connected to a turning device. This device has a plastic disk mounted onto it, with holes every 2 degrees which allowed us to control rotations with a 2° resolution [Fig. 2(d)]. Laser pointers were used to control the amount of rotation.

We measured the impulse responses from the loudspeakers to Valdemar's ears using a 14^{th} order MLS signal at a sampling frequency of 48 kHz. These measurements were done with Valdemar being placed in the geometrical center of the two semicircular arcs, and oriented toward the middle point between the two frontal loudspeakers. We will refer to this position as the nominal center position. The loudspeakers were placed in three different span angles ($\theta_s = \{12°, 28°, 60°\}$) and at four different elevations ($\phi = \{0°, 30°, 60°, 90°\}$) Only two loudspeakers were measured at an elevation angle of 90°. Note that the mentioned loudspeaker positions are not expressed in spherical coordinates, strictly speaking. The angle θ_s refers to the span between loudspeakers in the front and in the back. ϕ is the angle between the horizontal plane and the plane formed by each pair of loudspeakers and the center of the manikin's head.

The impulse responses to the manikin's ears obtained for each of the aforementioned configurations (with Valdemar placed at the nominal center position) were used to calculate the crosstalk cancellation filters. Filters with 512 coefficients were calculated in Matlab using the three crosstalk cancellation techniques described earlier. In our previous study we observed that the performance of the crosstalk cancellation techniques improved when increasing the filter length. This was observed in particular with the generic crosstalk canceler and the least-squares in the frequency domain. Thus, we chose 512 filter coefficients so that the evaluation of the methods was done under optimal conditions. We calculated inverse filters for two- and four-channel setups. This resulted in a total of 21 different configurations.

A strong boost of high frequencies is generally undesirable. It is particularly important to be aware of this when working with HRTFs. The HRTFs present steep notches at frequencies around and above 8 kHz. When trying to invert such a transfer function, the solution will contain large peaks around these frequencies. Additionally, the analog antialiasing filter in the data acquisition equipment will cause also a strong boost just below the Nyquist frequency after inversion. This is also the case at the low frequency end due to the measurement equipment limitations. It is thus necessary to apply a band-pass filter to the inverse filters. In order not to overdrive the loudspeakers, we decided to set the upper frequency limit to 8 kHz. The lower limit was set to 200 Hz due to the physical limitations of the loudspeakers.

Due to the inherent high gain characteristics at low frequencies when using closely spaced loudspeakers, we additionally considered it necessary not only to regularize the high frequencies but also the frequencies below 500 Hz. Therefore, we chose a band-stop filter as shape factor ($B(z)$). Using the equiripple method, we calculated an FIR filter of order 112 with stop-band frequencies at 500 and 6000 Hz, and 60 dB of attenuation at the stop-band. The frequencies in the stop-band are not affected by the regularization.

In order to choose an appropriate regularization value, methods such as the L-curve or the generalized cross-validation (GCV) had been proposed previously in the literature.[20] However, the determination of a proper choice of regularization parameter is not straightforward. For instance, it has been shown that the L-curve is more robust in the presence of errors than the GCV.[21] Another drawback of the GCV is that it can have a very flat minimum and the minimum itself might be difficult to localize numerically.[20] On the other hand, it is also known that if the problem is relatively well-conditioned the L-curve fails to converge.[22]

This lead us to believe that in the case of crosstalk cancellation problems, calculating the optimum regularization values by using these methods might be unpractical. First, the matrix $\mathbf{H}$ is ill-conditioned only at few frequencies, thus the L-curve method will fail to converge for a large range of frequencies. Second, calculating the optimum regularization value for each single frequency using GCV method becomes a numerically demanding problem when the number of channels is increased and the minimum in the GCV function is not clearly defined. Besides, it is also well known that the exact value of β is usually not critical.[1]

Since we know that most singularities occur at low and high frequencies, it seems more reasonable to design a shape factor that accounts for this—such as the shape factor described before. In this work, the regularization constant β was set such that the power of the crosstalk cancellation filters was approximately the same for all the configurations.

Author's complimentary copy

TABLE I. Summary of the different configurations measured in this study. A 14th order MLS signal with sampling frequency of 48 kHz was employed to measure the channel separation at both ears for each configuration and head misalignment.

Configurations	Two-channel	Four-channel
Elevation angles (ϕ)	0°,30°,60°,90°	0°,30°,60°
Span angles (θ_s)	12°,28°,60°	
Lateral displacement (cm)	$x=-20,-18,\ldots,18,20$	
Frontal displacement (cm)	$y=-20,-18,\ldots,18,20$	
Head rotation (°)	$\theta_d=-30,-28,\ldots,28,30$	

We chose this criterion in order be able to compare the different configurations directly. The power of the crosstalk cancellation filters is defined as[6]

$$P_i=\frac{1}{N}\sum_{k=0}^{N-1}[|C_{i1}(k)|^2+|C_{i2}(k)|^2], \tag{10}$$

where C_{ij} is the j^{th} component of the i^{th} row of the matrix $\mathbf{C}$ and N is the number of frequency samples. When choosing the proper regularization constant, it was also ensured that the gain of the filters did not exceed 12 dB, in order not to overdrive the loudspeakers. Initial experiments showed that sufficient crosstalk cancellation could be obtained with that limitation.

A. Sweet spot measurements

For each of the 21 different configurations, we measured the channel separation for different lateral displacements, frontal displacements and head rotations of Valdemar. The channel separation is defined as the ratio of the cross-term to the direct signal,[11]

$$CHSP_i(k)=\frac{\hat{R}_{ij}(k)}{\hat{R}_{ii}(k)}, \tag{11}$$

where $\hat{\mathbf{R}}$ in this case is the result matrix corresponding to a displaced position or rotation with respect to the nominal center position [see Eq. (4)]. The lateral displacements x and frontal displacements y were measured between −20 cm to 20 cm with steps of 2 cm, where $x=0$ and $y=0$ correspond to the nominal center position. The head rotations θ_d were measured between −30° to 30° with a resolution of 2°, where $\theta_d=0°$ corresponds to the nominal center position and negative angles to clockwise rotations. Table I summarizes the measurements carried out.

To measure the channel separation we measured the impulse responses to the ipsilateral and contralateral ears, when the sound was intended to be reproduced at only one ear. This means that one of the binaural input signals to the crosstalk cancellation network was an MLS signal, while the other binaural signal was set to zeros. A more detailed description of the channel separation measurements can be found in Ref. 23.

In order to calculate the sweet spot size we used the channel separation index, which is the average over frequency of Eq. (11):[6]

$$\overline{CHSP}_i=\frac{1}{n_f-n_j+1}\sum_{k=n_j}^{n_f}20\log_{10}(|CHSP_i(k)|)\quad[dB], \tag{12}$$

where n_j and n_f define the frequency range of interest and the selected frequencies k are distributed over a logarithmic scale. In this study, we calculated the channel separation index in the frequency range from 200 to 8000 Hz.

To get a better overview of the results, we made use of the two sweet spot definitions described by Bai *et al.*[6] The first one is the absolute sweet spot, which is defined as two times the maximum leftward displacement from the nominal center position that results in a channel separation index below −12 dB. The second one is the relative sweet spot, which is two times the maximum leftward displacement from the nominal center position that results in a channel separation index degradation of 12 dB with respect to the nominal center position. Though in this paper we did not assume a symmetric sweet spot. Therefore, we re-defined the absolute and relative sweet spot as the maximum left-to-right/back-to-front displacement or left-to-right rotations such that the aforementioned criteria are met.

The relative sweet spot can be regarded as a measure of robustness, whereas the absolute sweet spot can be considered as a measure of the area in which the performance of the crosstalk cancellation system is effective.

IV. MEASUREMENTS RESULTS

The analysis is divided into the three different head misalignments. For each movement, we present the absolute and relative sweet spot defined before, as a function of loudspeakers placement. Only two examples of the channel separation as a function of frequency and displacement are presented. One for the lateral displacement and one for the frontal displacements. We chose these two movements given that complementary features can be observed in those graphs.

A. Lateral displacement

Figure 3 shows the channel separation as a function of lateral displacement and frequency with the two-channel configurations using the least-squares method in the frequency domain. The columns correspond to the span angles $\theta_s=12°$, 28° and 60°. The rows correspond to the elevation angles $\phi=0°$, 60° and 90°.

We can observe a good agreement with results presented in previous studies in which the sweet spot was simulated for loudspeakers placed in the horizontal plane.[6,13] In general, closely spaced loudspeakers ($\theta_s=12°$) result in a wider effective area than the typical stereo setup ($\theta_s=60°$). We can also see that the region of ringing frequencies shifts to higher frequencies as the loudspeakers get closer to each other. With the 12° span angle configuration, no ringing is observed in the evaluated frequency range. Interestingly, as the loudspeakers go up in elevation ($\phi=60°,90°$), we observe that the channel separation becomes considerably smoother for all span angles.

Author's complimentary copy

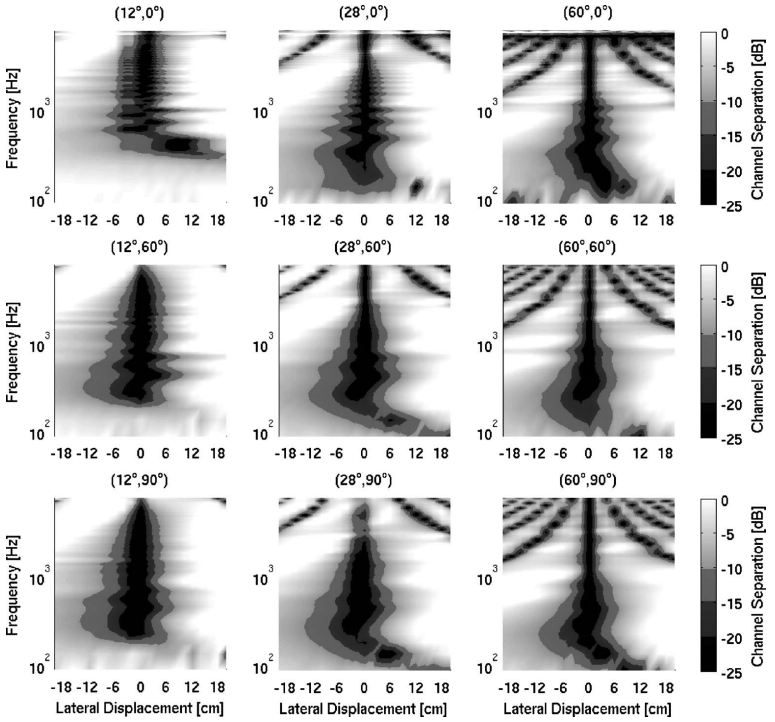


FIG. 3. Channel separation at the left ear as a function of lateral displacement and frequency measured on nine different loudspeaker configurations in two-channel arrangements. The results measured with the least-squares method in the frequency domain (LS_f) are presented. Left panel: 12° span angle at 0° (horizontal plane), 60° and 90° elevation. Center panel: 28° span angle at 0° (horizontal plane), 60° and 90° elevation. Right panel: 60° span angle at 0° (horizontal plane), 60° and 90° elevation.

The calculated absolute and relative sweet spot with respect to lateral displacement are presented in Fig. 4. To ease the analysis of the results, we divided each plot into three columns, each one corresponding to a different span angle. The *x*-axis represents the different elevation angles.

As observed previously, closely spaced loudspeakers present a larger controlled area in comparison to the typical stereo setup. The absolute sweet spot decreases slightly as the span angle becomes wider, but we do not observe any significant changes with elevation. All the methods result in similar values. Only with the two-channel configurations [Fig. 4(a)], when using the generic crosstalk canceler with the 12° span angle placed at 30° and 60° elevation, a narrower effective area is obtained. The four-channel configurations show a slightly larger absolute sweet spot than the two-channel configurations.

These result might seem contradicting with the results presented in Ref. 6, in which the absolute sweet spot increased with wider span angles. In that study, the channel separation index was calculated for different bandwidths in which the lower bound was 100 Hz and the regularization was set constant with respect to frequency. Additionally, that analysis was done under ideal conditions where only the HRTFs were taken into account. In this paper the channel separation index is calculated from 200 Hz, which is the lower frequency limit imposed by the loudspeakers. We used in addition a frequency dependent regularization that penalizes low and high frequencies. It is well known that at low frequencies closely spaced loudspeakers present a rather poor performance. However, given the natural roll-off at low frequencies of the loudspeakers impulse responses and the employed shape factor, the differences at low frequencies between the configurations are not as dramatic as the ones observed in Ref. 6 (see Fig. 3). Thus, the poor performance at low frequency of the closely spaced loudspeakers does not influence significantly the channel separation indexes with the used bandwidth.

Looking at the relative sweet spot we can see that there is a large variance in results among the two-channel configurations. Generally, at an elevation angle of $\phi=30°$ the relative sweet spot is smaller than in the horizontal plane ($\phi=0°$) and right above the manikin's head ($\phi=90°$), especially with the 12° and 28° span angle configurations. The 60° span angle configuration is in general less robust than the closely spaced loudspeakers.

Although the generic crosstalk canceler results in narrower controlled areas with a span angle of 12°, it shows to

Author's complimentary copy

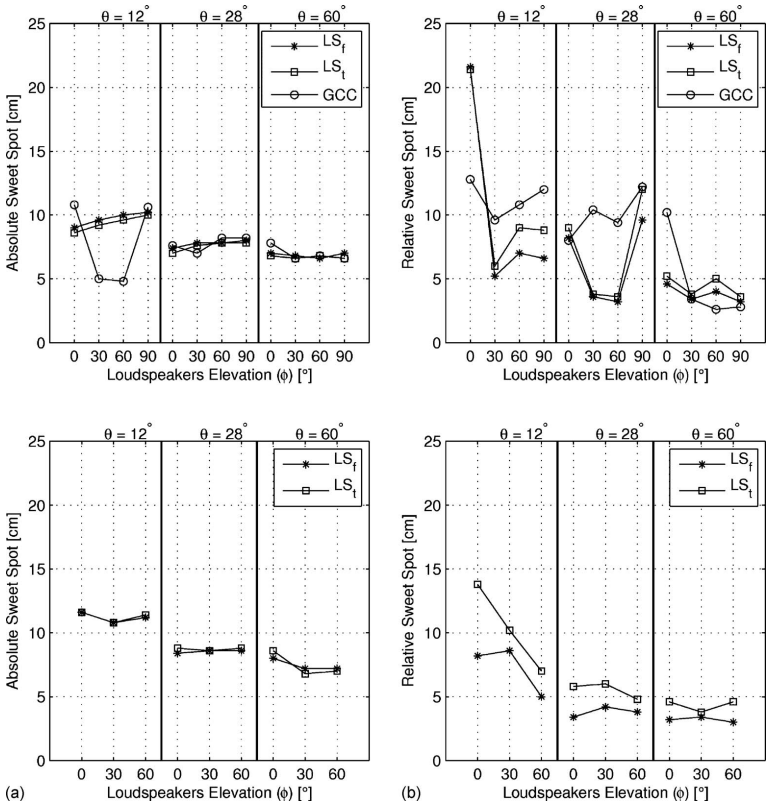


FIG. 4. Absolute and relative sweet spot at the left ear as a function of lateral displacement for each measured loudspeaker configuration. (a) Two-channel configurations, (b) Four-channel configurations. The generic crosstalk canceler (GCC), the least-squares methods in the frequency domain (LS_f) and the time domain (LS_t) are compared. The columns correspond to the loudspeakers span angle (θ_s) and the x-axis to the loudspeakers elevation angle (ϕ).

be more robust than the least-squares methods when the loudspeakers are placed above 30°. We observed a similar behavior with the 28° span angle placed at 30° and 60° elevation. This can be surprising, since in our previous study the least-squares methods outperformed the generic crosstalk canceler. In general the least-squares methods result in larger channel separation than the generic crosstalk canceler. That is still the case if we evaluate only the channel separation measured at the nominal center position. Figure 5 shows the channel separation index calculated for each method as a function of lateral displacement for a loudspeaker span of 28° placed at 30° elevation. We can observe that the three methods result in similar values at all displaced positions. Only at the nominal center position, the channel separation index obtained with the generic crosstalk canceler is around 10 dB larger than the indices obtained with the least-squares methods. Consequently, in spite of the fact that the generic crosstalk canceler is less effective at the nominal center position than the least-squares methods, the channel separation index does not vary as dramatically with lateral displacement as it does with the least-squares methods at these specific configurations.

Regarding the relative sweet spot of the four-channel configurations illustrated in Fig. 4(b), we can see that the least-squares method in the time domain is slightly more robust than the frequency domain method. Additionally, when the 12° and 28° span angle configurations are placed in the horizontal plane, the four-channel configuration shows to be less robust to lateral displacement than the two-channel configuration.

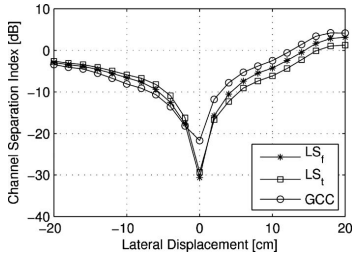


FIG. 5. Channel separation index as a function of lateral displacement. The loudspeakers span is 28° and are placed at 30° elevation

Author's complimentary copy

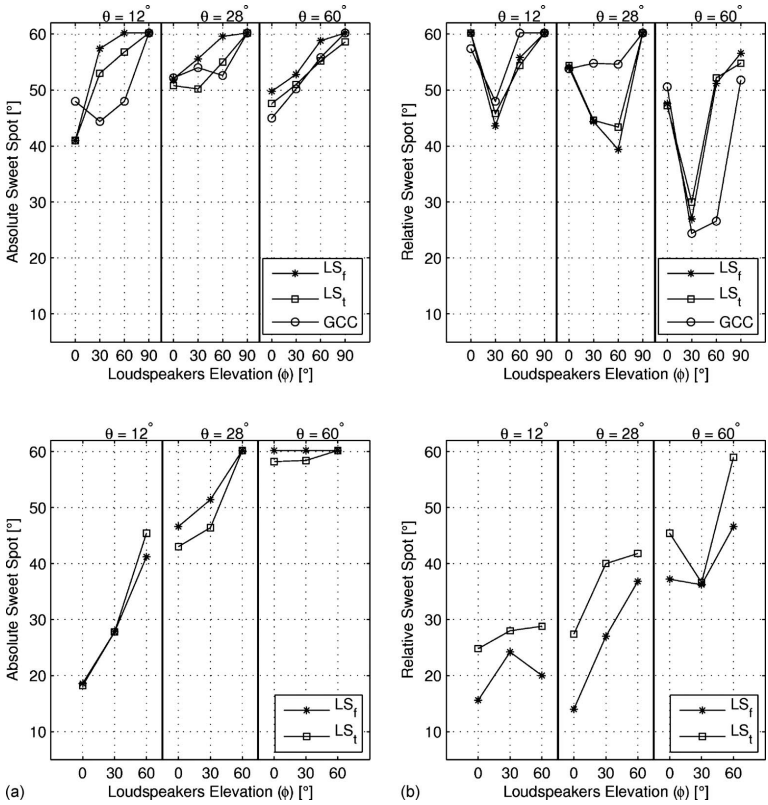


FIG. 6. Absolute and relative sweet spot at the left ear as a function of head rotation for each measured loudspeaker configuration. (a) Two-channel configurations. (b) Four-channel configurations. The generic crosstalk canceler (GCC), the least-squares methods in the frequency domain (LS_f) and the time domain (LS_t) are compared. The columns correspond to the loudspeakers span angle (θ_s) and the x-axis to the loudspeakers elevation angle (ϕ).

B. Head rotation

The absolute and relative sweet spot with respect to head rotation for the two-channel case is presented in Fig. 6(a). The absolute sweet spot is larger than 40° for all configurations and there is a systematic improvement when increasing the elevation angles. There is generally little variance between the methods.

When the loudspeakers are placed at elevation angles of 30° and 60° the relative sweet spot is smaller compared with the loudspeakers placed at 0° and 90°. Looking at the absolute sweet spot values, we can imply that at 0° elevation the channel separation index is in general larger than at elevated positions. Due to the head shadowing effect, which provides natural separation at high frequencies, the channel separation index does not change dramatically with head rotations in the direction of the contralateral ear when the loudspeakers are placed in the horizontal plane. This results in a larger relative sweet spot than when the loudspeakers are placed at 30° elevation, where the head shadowing effect is considerably reduced and the variations of the channel separation index are more marked. Yet, as the loudspeakers go higher up in elevation, the variations of the transfer functions to the ears with respect to azimuth become substantially smaller. As a result, the relative sweet spot increases with increasing elevation.

The generic crosstalk canceler shows again a larger relative sweet spot than the least-squares methods with a span angle of 28°. At 90° elevation, the 12° and 28° span angle configurations show maximum robustness with all crosstalk cancellation methods.

As oppose to the two-channel configurations, the absolute sweet spot of the four-channel configurations increases with wider span [Fig. 6(b)]. Interestingly, the relative sweet spot of the four-channel configuration shows that the 60° span angle configuration is more robust to head rotation than closely spaced loudspeakers. It shows to be in most instances less robust than the two-channel configurations with the 12° and 28° span angles.

C. Frontal displacement

Unlike the results obtained with the lateral displacements and head rotations, we observed significant differences

Author's complimentary copy

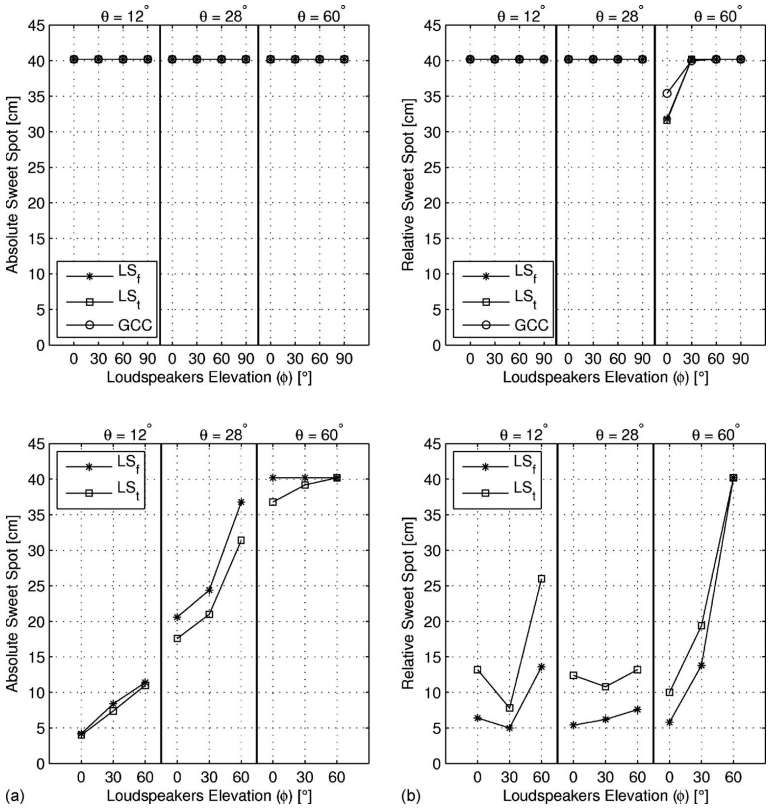


FIG. 7. Absolute and relative sweet spot at the left ear as a function of frontal displacement for each measured loudspeaker configuration. (a) Two-channel configurations. (b) Four-channel configurations. The generic crosstalk canceler (GCC), the least-squares methods in the frequency domain (LS_f) and the time domain (LS_t) are compared. The columns correspond to the loudspeakers span angle (θ_s) and the x-axis to the loudspeakers elevation angle (ϕ).

between the channel separation of the two- and four-channel configurations with respect to frontal displacements. All the two-channel configurations result in a maximum effective area independent of span and elevation angle [Fig. 7(a)], whereas the controlled area of the four-channel configurations improves with elevation and wider span angles, reaching its maximum with the 60° span angle configuration [Fig. 7(b)]. Note that the measurements carried out in this work were done in a range of ±20 cm from the nominal center position. Thus, the results observed in Fig. 7(a) suggest that the absolute sweet spot with respect to frontal displacements for the two-channel configurations is wider than the measured region.

The relative sweet spot of the two-channel configurations is also maximum for most arrangements. Merely the 60° span angle configuration in the horizontal plane shows to be less robust to this movement. On the other hand, the four-channel configurations is significantly less robust to frontal displacements than the two-channel configuration. Yet, the 60° span angle configuration placed at 60° elevation, equals the performance of the two-channel configurations.

Figure 8 complements these observations with the four-channel configuration. We can clearly see an interaction between the frontal and rear loudspeakers, where a region of ringing frequencies is present above 1 kHz for all configurations.

V. DISCUSSION AND CONCLUSIONS

The results of the measurements presented in this study are in good agreement with those of previous studies, in which the sweet spot size was found to become wider as the loudspeakers get closer to each other.[6,13] This applies in particular to the lateral displacement. However, those studies evaluated the sweet spot size mostly for loudspeakers placed in the horizontal plane (ϕ=0°) and in two-channel arrangements. Yet, the results presented in this paper show that when using four-channel configurations a rather unexpected behavior occurs: the absolute sweet spot with respect to head rotation and frontal displacement increases with wider angles.

One particular observation with the two-channel configurations, is that the generic crosstalk canceler is more robust than the least-squares methods when the loudspeakers with 28° span angle are placed at 30° and 60° elevation [Fig.

Author's complimentary copy

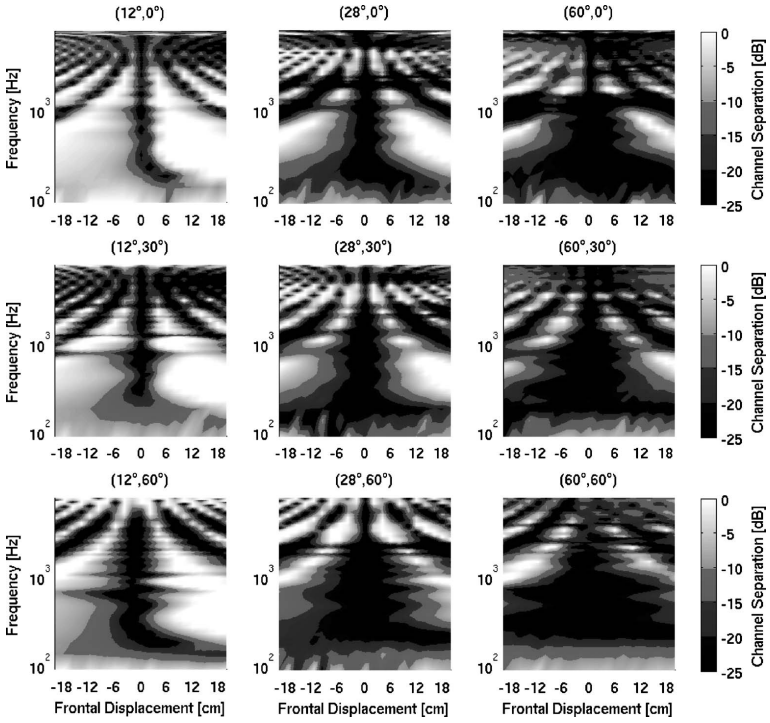


FIG. 8. Channel separation at the left ear as a function of frontal displacement and frequency measured on nine different loudspeaker configurations in four-channel arrangements. Results obtained with the least-squares method in the frequency domain (LS_f) are presented. Left panel: 12° span angle at 0° (horizontal plane), 30° and 60° elevation. Center panel: 28° span angle at 0° (horizontal plane), 30° and 60° elevation. Right panel: 60° span angle at 0° (horizontal plane), 30° and 60° elevation.

4(a)]. In our previous study,[11] the generic crosstalk canceler showed a poor performance in comparison to the least-squares methods. However, we observed that the channel separation was generally larger when the loudspeakers were placed at 30° elevation, which suggested us that this method can be especially robust with those configurations. Even though the generic crosstalk canceler might result in a smaller channel separation index than the least-squares methods, the channel separation index does not change dramatically with movements as it does with the least-squares methods.

Each of the three methods evaluated in this paper approximates the phase of the system differently. The influence of such phase approximations in the crosstalk cancellation performance variates with loudspeaker placement. This implies that at some particular configurations the phase differences between methods could have a larger influence in the overall performance of the crosstalk cancellation than at other configurations. This could be the reason for the particular good performance of the generic crosstalk canceler with loudspeakers placed at 30° and 60° elevation angles.

With respect to head rotations, all the configurations showed to be rather robust. We observed in general that the higher the loudspeakers are placed the wider the obtained absolute sweet spot. This is not a surprising observation, given that the variations of the HRFTs with respect to azimuth at elevated positions are considerably smaller than the variations of the HRTFs with respect to azimuth in the horizontal plane. Thus, when the listener turns the head and the loudspeakers are placed above the head, there are less errors introduced into the system than when the loudspeakers are placed in the horizontal plane.

Whereas there are no significant changes with span angle in the absolute sweet spot with respect to head rotations with the two-channel configurations [Fig. 6(a)], the absolute sweet spot increases significantly with wider span angles with the four-channel configurations [Fig. 6(b)]. The four-channel configurations showed to be more robust to head rotation with the 60° span angles.

Most of the two-channel configurations appear to be practically immune to frontal displacement [Fig. 7(a)]. This was expected, since with the fore-aft movement the variations of the HRFTs are less critical than with the lateral displacements, especially with closely spaced loudspeakers. Additionally, we believe that symmetry plays an important role here. In the case of lateral displacements, the variations of the path lengths between the ears and the loudspeakers are asymmetric, whereas with the frontal displacements, the

Author's complimentary copy

TABLE II. Values of the absolute sweet spot as a function of lateral displacements and head rotations for the two-channel configurations measured with the least square method in the frequency domain. The rows are the loudspeakers elevation angle (ϕ) and the columns the loudspeakers span angle (θ). The values of the relative sweet spot for each case are presented in parenthesis. The bold values highlight the best compromise between absolute and relative sweet spot.

	θ=12°		θ=28°		θ=60°	
	LD[a] (cm)	HR[b] (°)	LD (cm)	HR (°)	LD (cm)	HR (°)
ϕ=0°	9.0 (21.6)	41.0 (60.2)	7.4 (8.2)	52.0 (54.0)	7.0 (4.6)	49.8 (47.6)
ϕ=30°	9.6 (5.2)	57.4 (43.6)	7.8 (3.6)	55.6 (44.4)	6.8 (3.4)	52.8 (27.0)
ϕ=60°	10.0 (7.0)	60.2 (55.8)	7.8 (3.2)	59.6 (39.4)	6.6 (4.0)	58.8 (51.2)
ϕ=90°	**10.2 (6.6)**	**60.2 (60.2)**	**8.0 (9.6)**	**60.2 (60.2)**	**7.0 (3.2)**	**60.2 (56.6)**

[a]Lateral displacement.
[b]Head rotation.

paths lengths change symmetrically. In other words, for the lateral displacements the direct signal and the cross-terms sum up out of phase, but for the frontal displacement the phase differences between them are kept practically constant.

This hypothesis can be supported with the results of the four-channel configurations, where the path lengths change asymmetrically between the frontal and rear channels. These configurations exhibit a significantly narrower effective area with respect to frontal displacement [Fig. 7(b)] than the two-channel configurations. We observed a clear interaction between the rear and frontal loudspeakers, and ringing frequencies were present above 1 kHz for all configurations (Fig. 8). Then again, there is a clear tendency of increasing controlled area when widening the span angle and at elevated positions with the four-channel configurations.

The optimal loudspeaker configuration could be defined as the configuration that results in a best compromise between effective area (absolute sweet spot) and robustness (relative sweet spot). Table II summarizes the results of the absolute sweet spot and relative sweet spot (in parenthesis) for the two-channel configurations. For clarity reasons, only the values obtained with the least square method in the frequency domain are presented. We can conclude from the table that the optimum performance is obtained when the loudspeakers are placed right above the listener's head (ϕ =90°). Moreover, results presented in this paper show also that channel separation becomes considerably smoother when the elevation angle is increased. This might have a perceptual advantage, given that less coloration or artifacts would be introduced into the reproduction. We can also conclude that in general the two-channel configurations result in wider controlled area and are more robust to head rotation and frontal displacement than the four-channel configurations.

[1]O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante, "Fast deconvolution of multi-channel systems using regularization," IEEE Trans. Speech Audio Process. **6**, 189–194 (1998).
[2]O. Kirkeby and P. A. Nelson, "Digital filter design for inversion problems in sound reproduction," J. Audio Eng. Soc. **47**, 583–595 (1999).
[3]P. A. Nelson, F. Orduna-Bustamante, and H. Hamada, "Inverse filter design and equalization zones in multichannel sound reproduction," IEEE Trans. Speech Audio Process. **3**, 185–192 (1995).
[4]D. B. Ward, "Joint least squares optimization for robust acoustic crosstalk cancellation," IEEE Trans. Speech Audio Process. **8**, 211–215 (2000).
[5]M. R. Bai and C.-C. Lee, "Development and implementation of cross-talk cancellation system in spatial audio reproduction based on subband filtering," J. Sound Vib. **290**, 1269–1289 (2006).
[6]M. R. Bai and C.-C. Lee, "Objective and subjective analysis of effects of listening angle on crosstalk cancellation in spatial sound reproduction," J. Acoust. Soc. Am. **120**, 1976–1989 (2006).
[7]O. Kirkeby and P. A. Nelson, "The 'stereo dipole'—A virtual source imaging system using two closely spaced loudspeakers," J. Audio Eng. Soc. **46**, 387–395 (1998).
[8]D. B. Ward and G. W. Elko, "Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation," IEEE Signal Process. Lett. **6**, 106–108 (1999).
[9]T. Takeuchi and P. A. Nelson, "Optimal source distribution for binaural synthesis over loudspeakers," J. Acoust. Soc. Am. **112**, 2786–2797 (2002).
[10]J. Bauck, "A simple loudspeaker array and associated crosstalk canceler for improved 3D audio," J. Audio Eng. Soc. **49**, 3–13 (2001).
[11]Y. L. Parodi, "Analysis of design parameters for crosstalk cancellation filter applied to different loudspeaker configurations," in 125th Convention of the Audio Engineering Society, San Francisco, CA (2008).
[12]W. G. Gardner, *3-D Audio Using Loudspeakers*, 1st ed. (Kluwer Academic, Dordrecht, 1998).
[13]T. Takeuchi and P. A. Nelson, "Robustness to head misalignment of virtual sound imaging systems," J. Acoust. Soc. Am. **109**, 958–971 (2001).
[14]H. Møller, "Reproduction of artificial-head recordings through loudspeakers," J. Audio Eng. Soc. **37**, 30–33 (1989).
[15]A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," Sov. Math. Dokl. **4**, 1035–1038 (1963).
[16]J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*, 3rd ed. (Prentice-Hall, Englewood Cliffs, NJ, 1996).
[17]M. R. Bai, G. Shih, and C. Lee, "Comparative study of audio spatializers for dual-loudspeaker mobile phone," J. Acoust. Soc. Am. **121**, 298–309 (2007).
[18]O. Kirkeby, P. A. Nelson, and H. Hamada, "Local sound field reproduction using two closely spaced loudspeakers," J. Acoust. Soc. Am. **104**, 1973–1981 (1998).
[19]F. Christensen, C. B. Jensen, and H. Møller, "The design of VALDEMAR—An artificial head for binaural recordings purposes," in 109th Convention of the Audio Engineering Society, Los Angeles, CA (2000), p. 5253.
[20]P. Hansen and D. O'Leary, "The use of the L-curve in the regularization of discrete ill-posed problems," SIAM J. Sci. Comput. (USA) **14**, 1487–1503 (1993).
[21]Y. Kim and P. A. Nelson, "Optimal regularisation for acoustic source reconstruction by inverse methods," J. Sound Vib. **275**, 463–487 (2004).
[22]C. R. Vogel, "Non-convergence of the L-curve regularization parameter selection method," Inverse Probl. **12**, 535–547 (1996).
[23]Y. Lacouture Parodi and P. Rubak, "Preliminary evaluation of sweet spot size in virtual sound reproduction using dipoles," in 126th Convention of the Audio Engineering Society, Munich, Germany (2009).

Author's complimentary copy

4 MANUSCRIPT C

Lacouture Parodi, Y. and Rubak, P. (2010), **Sweet Spot Size in Virtual Sound Reproduction: A Temporal Analysis**, chapter in *Principles and Applications of Spatial Hearing*, World Scientific, in press.

Portions of this manuscript were in "Sweet Spot Size in Virtual Sound Reproduction: A Temporal Analysis", Proceedings of the International Workshop on the Principles and Applications of Spatial Hearing, Miyagi, Zao, Japan, November 11 - 13, 2009.

SWEET SPOT SIZE IN VIRTUAL SOUND REPRODUCTION: A TEMPORAL ANALYSIS

Y. LACOUTURE PARODI* and P. RUBAK

Department of Electronic Systems, Section of Acoustics
Aalborg University Aalborg East, 9220, Denmark
**E-mail: ylp@es.aau.dk*
http://www.es.aau.dk/sections/acoustics/

The influence of head misalignments on the performance of binaural reproduction systems through loudspeakers is often evaluated based on the amplitude ratio between the crosstalk and the direct signals. The changes in magnitude give us an idea of how much of the crosstalk is leaked into the direct signal and therefore a sweet spot performance can be estimated. However, as we move our heads, the time information of the binaural signals is also affected. This can result in ambiguous cues that can destroy the virtual experience. In this paper, we present an analysis in the time domain of the influence of head misalignments. Using the interaural cross-correlation we estimated the interaural time delay and defined a sweet spot. The analysis is based on measurements carried out on 21 different loudspeaker configurations, including two- and four-channel arrangements. Results show that closely spaced loudspeakers are more robust to lateral displacements than wider span angles. Additionally, the sweet spot as a function of head rotations increases systematically when the loudspeakers are placed at elevated positions.

Keywords: Virtual acoustics; Crosstalk cancellation; Sweet spot; Interaural time delay; Stereo dipoles.

1. Introduction

The reproduction of an authentic auditory event is possible if the sound signals at the ears matches the sound pressures of the real environment. This is the central idea of binaural techniques and it is based on the assumption that the sound pressures at the ears control our perception of any auditory event. Virtual auditory events can be rendered through headphones or loudspeakers.

One of the biggest challenges of binaural reproduction through loudspeakers is to avoid that the signals that are to be hear in one ear are also hear in the other. This problem can be solved by introducing the appropriate filters into the reproduction chain. These filters are usually designed for a fixed position.

Head movements are known to add important dynamic cues to the localization

of sound sources. However, in binaural reproduction systems through loudspeakers, when the listener moves the head, the transfer functions of the acoustical paths from the loudspeakers to the ears do not longer correspond to the transfer functions used to design the filters. This can result in leakages from the contralateral path into the ipsilateral path and thus, the virtual image can be destroyed.

The maximum amount of displacement allowed such that the errors introduced do not significantly affect the virtual reality effect had been the focus of different studies in the past.[1–4] In most instances, these analysis were conducted based on magnitude ratios between the crosstalk and the direct signals, given that with the spectral information we can easily observe how much of the signal from the contralateral path leaks into the ipsilateral path. However, spectral information might not be sufficient to assess the space region in which the errors introduced are negligible. As we move our heads, the time information of the signals changes accordingly, introducing delay errors into the desired binaural signal. This can also produce conflicting cues and therefore destroy the spatial perception.

The robustness of temporal cues were discussed by Takeuchi *et al.* in Ref. 1. In their analysis, they employed a free field model and the head related transfer functions (HRTF) from a head and torso simulator. They only analyzed two loudspeaker configurations: two-channel configurations with 10° and 60° span angles placed on the horizontal plane. In this paper we present a temporal analysis of the sweet spot for 21 different loudspeaker configurations, including two- and four-channel arrangements placed at different elevations. The analysis is based on measurements carried out at the acoustical laboratories at Aalborg university and is intended as a complement of the channel separation analysis presented in Ref. 4.

2. Crosstalk Cancellation

The purpose of a crosstalk cancellation network is to cancel the signals that arrive from the contralateral path, so that the binaural signals are reproduced at the ears in the same way they would be reproduced through headphones. Figure 1 illustrates a simplified diagram of a binaural reproduction system through loudspeakers. The blocks C_{ji} represents a set of optimal filters and the functions h_{ji} describes the acoustical paths from the j^{th} loudspeakers to the i^{th} ear. The signal d_i contains the binaural signal that is to be reproduced at the i^{th} ear and v_i is the signal that is actually reproduced at the i^{th} ear.

Perfect crosstalk cancellation is achieved when $d_i = v_i$. In other words, we need a set of filters C_{ji} such that $\mathbf{H} \cdot \mathbf{C} = \mathbf{I}$. Here $\mathbf{C}$ is an $n \times 2$ matrix containing the crosstalk cancellation filters, where n is the number of loudspeakers. $\mathbf{H}$ is an $2 \times n$ matrix which contains the transfer functions describing the acoustical paths from the sources to the ears and which we will refer to as the plant matrix. $\mathbf{I}$ is the

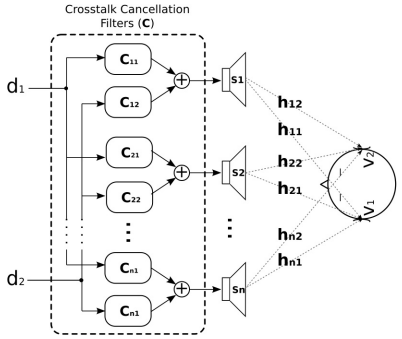


Fig. 1. Simplified diagram of a crosstalk cancellation system.

identity matrix. The problem is basically to find the inverse of **H**.

The plant matrix **H** is generally singular and therefore not invertible. Besides, when the reproduction system consists on more than two loudspeakers, the equation system becomes overdetermined and a direct inversion is not feasible. Thus, it is necessary to model the system such that we can obtain an approximation that is closed enough to the required solution.

There exists a number of methods to obtain the optimal inverse filters **C**. In this study, we implemented three different crosstalk cancellation techniques. The first one, which we will refer as the generic crosstalk canceler (GCC), applies the exact matrix inverse definition. It obtains the filters by inverting directly a matrix composed of the minimum-phase section of the plant transfer functions h_{ji}. It models the interaural transfer functions (ITF) as the ratio between the minimum phase component of the ipsilateral and contralateral transfer functions. The all-pass section of the transfer functions are approximated to a frequency independent delay. This is based on the assumption that the phase of the all-pass section is approximately linear at low frequencies. Given that it is based on a direct matrix inverse, this methods is only applicable for two-channel arrangements.

The other two methods are based on least square approximations. These methods do not try to invert the plant matrix directly but seek for the best approximation which result in minimum errors. One of the methods we implemented is the so called fast deconvolution method, which is based on the fast fourier transform. We will refer to it in the results as LS_f. The other method calculates the optimal filters in the time domain, using matrices that contain digital FIR filters. We will refer to this method as LS_t. Frequency dependent regularization was incorporated. A detailed description of the methods and the implementation can be found in Refs. 5 and 4.

3. Changes in the Interaural Time Delay

In Ref. 6 we presented an analysis of the sweet spot size as function of lateral displacements and head rotations. This analysis was done for 21 different loudspeaker configurations, including two- and four-channel arrangements. In that study, we made use of the absolute and relative sweet spot definitions described in Ref. 3. The first one is defined as the maximum displacement from the nominal center position that results in an average channel separation index larger than -12dB. The second one is defined as the maximum displacement from the nominal center position that results in a channel separation index degradation of 12dB with respect to the nominal center position. This is,

$$\text{Absolute sweet spot} := max\left\{d|\ \overline{CHSP}_d \leq -12dB\right\} \tag{1}$$

$$\text{Relative sweet spot} := max\left\{d|\ \overline{CHSP}_d \leq \overline{CHSP}_o + 12dB\right\}, \tag{2}$$

where the variable d corresponds to either a lateral displacement or a head rotation from the nominal center position. $\overline{CHSP}_d$ is the channel separation index at the position d and $\overline{CHSP}_o$ is the channel separation index at the nominal center position. The channel separation index is defined as the magnitude ratio between the contralateral and ipsilateral signals.[4]

In general, we observed that a wider control area is obtained when the loudspeakers are closely spaced and at elevated positions. Additionally, results showed us that the two-channel configurations tend to be more robust and result in wider control area than when using the four-channel configurations. However, questions were still open on whether such movements influence the phase information in a similar manner and whether this phase changes depend only on the loudspeakers positions or also on the method used to calculate the filters.

In an ideal situation, if we send a pair of impulses at the same time through our binaural reproduction system illustrated in Fig. 1, the reproduced signals v_i will be the same impulses with the same delay. Thus, the interaural time delay (ITD) of the binaural reproduction system will equal $0\mu s$. When the listener moves the head this ITD changes accordingly to the movement. This time difference is introduced into the desired binaural signal, changing the original ITD and therefore generating ambiguous cues.

Several methods to determine the ITD can be found in the literature. In Ref. 7, Minnaar *et al.* described and compared different methods to calculate the ITD. It is suggested that determining the ITD by calculating the group delay of the excess phase components evaluated at 0Hz is numerically more robust and consistent than other methods proposed in the literature. However, if the group delay at 0Hz is incorrect due to for example a high-pass filtering, results from this methods are not longer consistent. In Ref. 1, Takeuchi *et al.* propose to use the interaural

cross-correlation (IACC) to determine the ITD and analyze the temporal changes. According to Minnaar's results,[7] the IACC consistently overestimated the ITD values. Nevertheless, the relative variation with respect to angle was similar to the variations observed with the other methods. One advantage of the IACC method is that it is less sensitive to noise as opposed to methods such as the leading-edge. Furthermore, there are indications that the nervous system calculates the ITD by means of a cross-correlation.[8] Thus, it can be considered a better approximation of the auditory system. Based on these arguments and given that the analysis presented here was done with high-pass filtered signals, we decided to use the IACC to estimate the ITD changes of our reproduction system as a function of head movements. The discrete time IACC is defined as:

$$\Psi(m) = \begin{cases} \sum_{n=0}^{N-m-1} p_1(n)p_2(n+m) & m \geq 0 \\ \\ \Psi^*(-m) & m < 0 \end{cases} \tag{3}$$

where $p_i(t) = R_i^{contra}(t) + R_i^{ipsi}(t)$ is the linear combination of the ipsilateral and contralateral signals, when an impulse is sent to the i^{th} ear and N is the length of the signals.[1] Here the ITD of the binaural reproduction system can be estimated as the delay corresponding to the maximum in the cross-correlation function.

Now, we need to define a threshold for the ITD discrimination in order to assess the sweet spot quantitatively. Experiments presented in Ref. 9 suggest that the audibility threshold for the ITD is $10\mu s$. In Ref. 10 larger values were obtained when evaluated with naïve listeners. However, these results showed a large variance between subjects. Thus, $10\mu s$ can be considered a strict but safe limit. Following this line, we defined the sweet spot as the maximum head misalignment allowed, such that the ITD difference between the nominal center position and the new position does not exceed $10\mu s$. In other words,

$$\text{ITD sweet spot} := max\{d|\ |ITD_o - ITD_d| \leq 10\mu s\}, \tag{4}$$

where ITD_o and ITD_d are the ITDs at the nominal center position and at a laterally displaced or rotated position respectively. In order to distinguish from the other sweet spot definitions mentioned above, we will refer to this definition as the ITD sweet spot. Since we are not assuming symmetry, the total sweet spot is calculated by adding the maximum leftward misalignment and the maximum rightward misalignment.

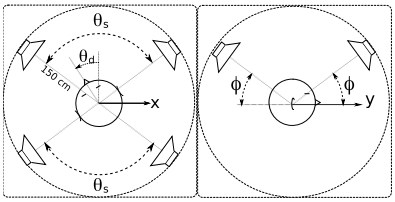


Fig. 2. General diagram of the measured loudspeaker configurations. θ_s corresponds to the span angle and ϕ corresponds to the elevation angle. Only the loudspeakers in front of the listener were used to evaluate the two-channels configuration.

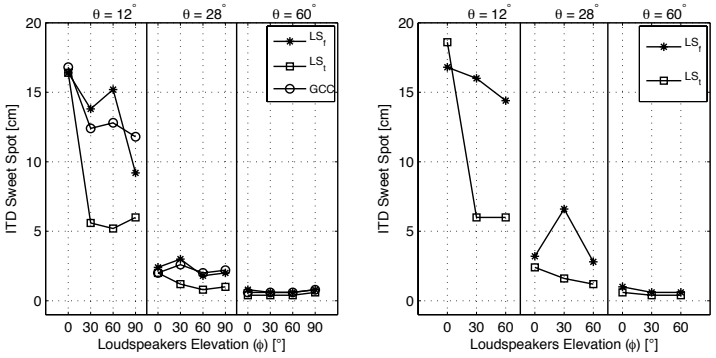
4. Measurements

We carried out the measurements in an anechoic chamber at the acoustics laboratories at Aalborg University. We measured 21 different loudspeaker configurations, including two- and four-channel arrangements. The loudspeakers were placed at three different span angles: $\theta_s = \{12^\circ, 28^\circ, 60^\circ\}$. Each of these configurations were measured at four different elevations: $\phi = \{0^\circ, 30^\circ, 60^\circ, 90^\circ\}$. Only two loudspeakers were measured at an elevation angle of 90°. Note that the mentioned loudspeaker positions are not expressed in spherical coordinates, strictly speaking. The angle θ_s refers to the span between loudspeakers in the front and in the back. ϕ is the angle between the horizontal plane and the plane formed by each pair of loudspeakers and the center of the manikin's head. Figure 2 illustrates the general diagram of the measured configurations.

To design the filters, we used the transfer functions of each of the aforementioned loudspeaker configurations measured with the artificial head Valdemar designed at Aalborg university.[11] For this purpose, the manikin was placed in the geometrical center of the arcs formed by the frontal and rear loudspeakers, facing towards the middle point between the two frontal loudspeakers. We refer to this position as the nominal center position. The filters were calculated using the three different crosstalk cancellation techniques mentioned before. Details of the measurement setup can be found in Ref. 4.

To evaluate the effect of head misalignments, we measured the channel separation when the manikin was placed at positions corresponding to lateral displacements, frontal displacements and head rotations. The lateral displacements x and frontal displacements y were measured from -20 cm to 20 cm with a resolution of 2 cm, where $x = 0$ and $y = 0$ correspond to the nominal center position. The head rotations θ_d were measured between -30° and 30° with a resolution of 2°, where $\theta_d = 0^\circ$ corresponds to the nominal center position and the negative angles to clockwise rotations.

4.1. *Results*



(a) Two-channel configurations (b) Four-channel configuration

Fig. 3. ITD sweet spot size for lateral displacements as function of loudspeaker configuration. Each column corresponds to each measured span angle θ_s and the x–axis to each measured elevation ϕ. Only the two-channel configuration was measured at 90° elevation.

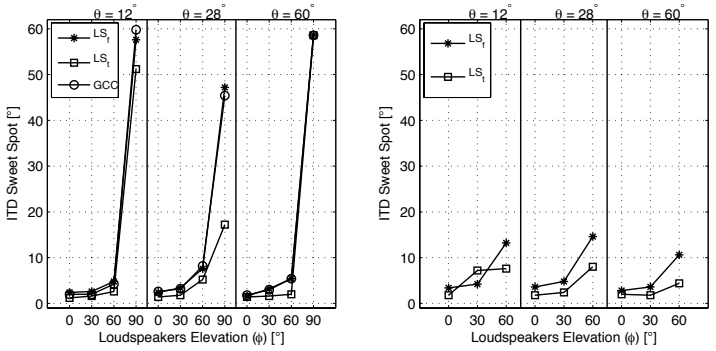
The ITD does not change significantly with frontal displacements, hence we only present the results obtained with lateral displacements and head rotations. Figure 3 shows the ITD sweet spot size for the two- and four-channel configurations. The plots are split into three sections, each one corresponding to one of the different span angles. The x-axis corresponds to the different elevation angles ϕ.

We can notice in Fig. 3(a) that the loudspeakers set with 28° and 60° span angles in two-channel configurations are significantly less robust to lateral displacements than the 12° span angle configuration. In our previous study, we found that the absolute sweet spot decreases gradually with wider span angles (see App. A Fig. 1(a)). However, the ITD sweet spot suggests us that only the 12° span angle is robust to lateral displacements, especially when it is placed on the horizontal plane.

We can also observe that the three different methods yield different results. In general, the LS_t results in narrower controlled areas than the LS_f and the GCC approaches. Even though there is no redundancy when inverting a two-channel system, the different numerical approximations used yield different filter coefficients. Thus, different phase errors are introduced by each method in the signals that reach the ears.

Figure 3(b) shows the ITD sweet spot size for the four-channel configurations as a function of lateral displacements. The results follow a trend similar to the two-channel case. The LS_f shows to be more robust to lateral displacements than

the LS_t, especially at 30° elevation. We can also observe a slight improvement in controlled area when compared with the two-channel configuration.



(a) Two-channel configurations (b) Four-channel configurations

Fig. 4. ITD sweet spot size for head rotations as function of loudspeaker configuration. Each column corresponds to each measured span angle θ_s and the x–axis to each measured elevation ϕ. Only the two-channel configuration was measured at 90° elevation.

Regarding head rotations, we can see in Fig. 4(a) that there is a dramatic improvement in sweet spot when the loudspeakers are placed at 90° elevation for the two-channel case. In general, the ITD sweet spot increases with elevation. We obtained a similar trend with the absolute sweet spot (see App. A Fig. 1(c)). Nevertheless, the ITD sweet spot shows to be narrower for elevations angles below 60°.

Looking at the results obtained with the four-channel configurations we can see that it follows the same tendency as the two-channel configurations (see Fig. 4(b)). Then again, there is an improvement in sweet spot size if compared with the two-channel configurations, especially at 60° elevation.

5. Conclusions

Different studies have evaluated the effect of head misalignments in binaural reproduction systems. Yet, these evaluations are often based on magnitude ratios between the contralateral and ipsilateral signals. In this paper, we intended to extend the results described in a previous analysis presented in Ref. 4 in which the absolute and relative sweet spot as a function of lateral displacements, frontal displacements and head rotations was evaluated for 21 different loudspeaker configurations. Here, we presented an evaluation of the robustness to movements from the time domain perspective. We defined the ITD sweet spot as the maximum

movement such that the ITD difference between the nominal center position and the new position does not exceed $10\mu s$. This is a rather strict limit, but based on some studies of the minimum audible ITD, we consider it a safe criteria.[9,10]

We observed that when evaluating the sweet spot in the time domain, a narrower control area is usually obtained in comparison to the results obtained in the sweet spot definitions based on magnitude ratios (see App. A Fig. A1). Only the 12° span angle showed to be sufficiently robust with respect to lateral displacements and actually resulted in larger values than those observed with the absolute sweet spot.

The controlled area with respect to head rotations of the two-channel configurations increases with elevation. Especially at 90° elevation, the ITD does not vary significantly with large rotations. In contrast, when we place the loudspeakers on the horizontal plane, small head rotations result in ITD changes larger than $10\mu s$. This is expected, since the slope of the ITD as a function of head rotations decreases with elevation.

The four-channel configurations showed to be more robust to head rotations than the two-channel case. This is surprising, since when analyzing the absolute sweet spot, the performance of the four-channel configurations showed to be poorer than the two-channel setups (Fig. A1). Another aspect we noted in the results obtained with the two- and four-channel cases, is that the ITD sweet spot does not vary significantly with the different span angles. Only elevation shows an improvement in the controlled area.

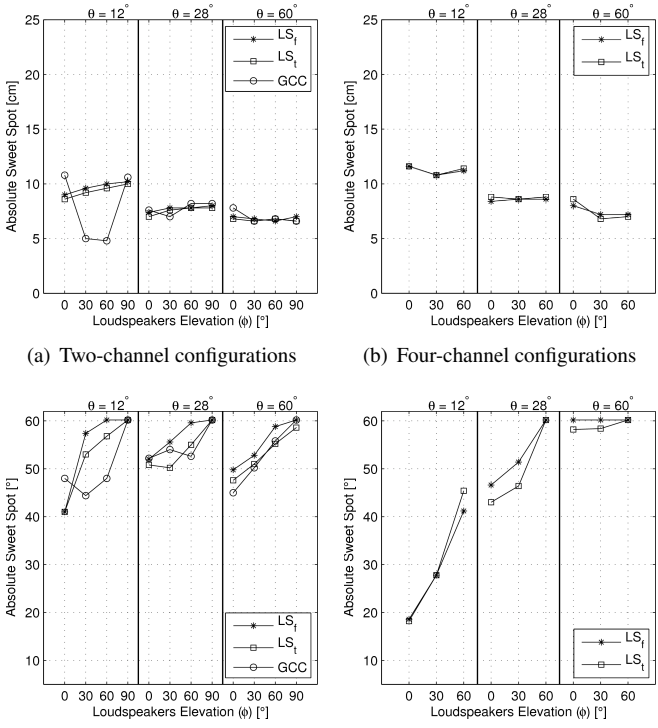
In this paper we also evaluated three different crosstalk cancellation techniques. We expected that the variations of the ITD depend only on the placement of the loudspeakers. Yet, the methods not always yielded the same results. In general, the LS_t resulted in narrower controlled area than the LS_f and the GCC. That is especially noticeable with the 12° span angle configuration and lateral displacements. Even though the LS_t is known to make an efficient use of the available coefficients,[12] these results suggest that it is less robust to phase errors than the LS_f. The results presented here show that not only the loudspeaker placement influences the phase variations of the binaural system, but also the numerical errors introduced by the different approximations used by each method. This should be taken into consideration when designing optimal crosstalk cancellation filters.

It is well known that the human auditory system employs two mechanism to discriminate the location of a sound source. The first one extracts the interaural time differences and it is known to work up to 1.6kHz. The second one uses the interaural sound pressure level differences and it is known to be dominant for signals with frequencies above 1.6kHz. Even though they function independently up to a certain extend, there is evidence that they interact with each other. For exam-

ple, trading experiments had shown that up to a certain extend an auditory event can be displaced by either a time or a level difference.[8] So far, we have analyzed the influence of head misalignment in the frequency and time domain separately and that give us a pretty good idea of how the performance of the different configurations change when varying specific parameters. However, it should be possible to find a model that combines both results and predicts the controlled area more accurately. In this way, the results can be better understood from the human localization point of view. This question is the topic of future research.

Appendix A. Absolute Sweet Spot

In order to give the reader a better understanding of the results described in this paper, we include here the results of the absolute sweet spot presented in Ref. 4.



(a) Two-channel configurations (b) Four-channel configurations

(c) Two-channel configurations (d) Four-channel configurations

Fig. A1. Absolute sweet spot at the left ear for lateral displacements [(a) and (b)] and head rotations [(c) and (d)] as a function of loudspeaker configuration. Each column corresponds to each measured span angle θ_s and the x–axis to each measured elevation ϕ. Only the two-channel configuration was measured at 90° elevation.

References

1. T. Takeuchi and P. A. Nelson, Robustness to Head Misalignment of Virtual Sound Imaging Systems, *Journal of the Acoustic Society of America* **109**, 958(March 2001).
2. J. Rose, P. Nelson, B. Rafaely and T. Takeuchi, Sweet Spot Size of Virtual Acoustic Imaging Systems at Asymmetric Listener Locations, *Journal of the Acoustic Society of America* **112**, 1992(November 2002).
3. M. R. Bai and C. Lee, Objective and Subjective Analysis of Effects of Listening angle on Crosstalk Cancellation in Spatial Sound Reproduction, *Journal of the Acoustic Society of America* **120**, 1976(October 2006).
4. Y. Lacouture Parodi and P. Rubak, Preliminary Evaluation of Sweet Spot Size in Virtual Sound Reproduction Using Dipoles, in *126th Convention of the Audio Engineering Society*, (Munich, Germany, 2009).
5. Y. Lacouture Parodi, Analysis of Design Parameters for Crosstalk Cancellation Filter Applied to Different Loudspeaker Configurations, in *125th Convention of The Audio Engineering Society*, (San Francisco, C.A, 2008).
6. Y. Lacouture Parodi and P. Rubak, Objective Evaluation of the Sweet Spot Size in Spatial Sound Reproduction Using Elevated Loudspeakers, *Journal of the Acoustical Society of America* **128**, 1045 (September 2010).
7. P. Minaar, J. Plogsties, S. Krarup, F. Christensen and H. Møller, The Interaural Time Difference in Binaural Synthesis, in *108th Convention of The Audio Engineering Society*, (Paris, France, 2000).
8. J. Blauert, *Spatial Hearing*, 3rd edn. (Hirzel-Verlag, 2001).
9. R. G. Klump and H. R. Eady, Some Measurement of Interaural Time Difference Thresholds, *The Journal Acoust. Soc. Am.* **28**, 859 (September 1956).
10. P. F. Hoffmann and H. Møller, Audibility of Differences in Adjacent Head-Related Transfer Functions, *Acta Acustica united with Acustica* **94**, 945 (2008).
11. F. Christensen, C. B. Jensen and H. Møller, The Design of VALDEMAR - An Artificial Head for Binaural Recordings Purposes, in *109th Convention of The Audio Engineering Society*, (Los Angeles, CA, 2000).
12. O. Kirkeby and P. A. Nelson, Digital Filter Design for Inversion Problems in Sound Reproduction, *Audio Engineering Society* **47**, 583(July/August 1999).

5 MANUSCRIPT D

Lacouture Parodi, Y. and Rubak, P. (2010), **A Subjective Evaluation of the Minimum Audible Channel Separation in Binaural Reproduction Systems Through Loudspeakers**, *Journal of the Audio Engineering Society*, submitted.

This manuscript was presented at the 128^{th} Convention of the Audio Engineering Society 2010, May 22-25, London, UK and received *the AES 128^{th} Convention Student Technical Paper Award*.

A Subjective Evaluation of the Minimum Audible Channel Separation in Binaural Reproduction Systems Through Loudspeakers

Yesenia Lacouture Parodi[1], and Per Rubak[1]

[1]*Aalborg University, Aalborg, DK-9220, Denmark*

Correspondence should be addressed to Yesenia Lacouture Parodi (ylp@es.aau.dk)

ABSTRACT

To evaluate the performance of crosstalk cancellation systems the channel separation is usually used as parameter. However, no systematic evaluation of the minimum audible channel separation has been found in the literature known by the authors. This paper describes a set of subjective experiments carried out to evaluate the minimum amount of channel separation needed such the binaural signals with crosstalk are perceived to be equal to the binaural signals reproduced without crosstalk. A three alternative-forced-choice discrimination experiment, with a simple adaptive algorithm with weighed up-down method was used. The minimum audible channel separation was evaluated for the listeners placed at symmetric and asymmetric positions with respect to the loudspeakers. Eight different stimuli placed a two different locations were evaluated. Span angles of 12 and 60 degrees were also simulated. Results indicate that in order to avoid lateralization the channel separation should be below -15dB for most of the stimuli and around -20dB for broad-band noise.

1. INTRODUCTION

To reproduce binaural signals through loudspeakers it is necessary to invert the acoustic paths from the loudspeakers to the ears. This is not only in order to equalize the loudspeakers response, but also to counteract the crosstalk. This is, the signals that should be heard in the left ear are also heard in the right ear and vice-versa. This process is known as crosstalk cancellation and there exists a number of different methods to calculate the optimal inverse filters [1, 11, 16, 18].

To evaluate the effective performance of a crosstalk cancellation system, the channel separation is usually used as parameter. The channel separation is defined as the magnitude ratio of the cross-terms to the direct signal. In other words, it is a measure of how much of the crosstalk is leaked into the desired signal. Additionally, to assess the sweet spot size of a binaural reproduction system through loudspeakers a limit for the channel separation should be defined. In [2] and [17] it is suggested to set the maximum acceptable level of crosstalk relative to the desired signal to -12 dB and -10 dB respectively. However, those values are based on personal experiences and no systematic evaluation of the audibility of the crosstalk has been found in the literature known by the authors.

This paper presents a set of subjective experiments carried out, which purpose was to measure the minimum audible channel separation. We thus define the minimum audible channel separation as the maximum level of crosstalk at which the binaural signals with crosstalk are perceived to be equal to the binaural signals reproduced without crosstalk.

The minimum audible channel separation was evaluated using a three-alternative forced-choice (3AFC) discrimination experiment. A simple adaptive up-down method was implemented, with the rule 1-down-2-up which converges to the 66,6% of the psychometric function [9].

Two subjective experiments were carried out using headphones. The first experiment simulated the listener symmetrically located with respect to the loudspeakers and pointing towards the middle point between the two loudspeakers. We refer to this location as the nominal center position. The second experiment simulated the listener placed at 5 cm to the left/right away from the nominal center position. In total, sixty four stimuli were assessed, among which were band-pass noise, 1-octave-band noises and speech placed at two different locations ($\pm 40^\circ$ and $\pm 90^\circ$) and simulating two different span angles (12° and 60°).

To simulate the crosstalk, we added to the raw binaural signals the cross-terms multiplied by a gain factor with a delay corresponding to the path differences between loudspeakers. With the gain factor we simulated the channel separation and with the delays the loudspeaker's span angle as well as the listener's location with respect to the loudspeakers.

This paper is organized as follows: in section 2 we introduce the concept of channel separation and describe the simplified model we used in order to carry out the subjective experiments. In section 3 we describe the psychometric method used to evaluate the minimum audible channel separation. Section 4 and 5 present the setup and results of the experiment I and II respectively. The conclusions drawn from this study are summarized in section 7.

2. SIMULATION OF THE CHANNEL SEPARATION

Figure 1 depicts the acoustic paths from the loudspeakers to the ears with a two-channel system. The functions H_{ij} represent the transfer functions between the i^{th} loudspeaker and the j^{th} ear. The signals d_j are the desired binaural signals to be reproduced and the signals v_j are the reproduced signals. The matrix $\mathbf{C}$ contains the crosstalk cancellation filters.

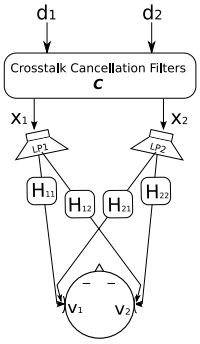


Fig. 1: Acoustic paths from the loudspeakers to ears in a real situation with a two-channel binaural reproduction system.

From this system we have that $\mathbf{HC} = \mathbf{R}$, where $\mathbf{H}$ is a matrix containing the transfer functions H_{ij} and $\mathbf{R}$ is a 2×2 matrix of which the diagonal elements represent the direct signals and the off-diagonal elements the crosstalk.

The channel separation (CHSP) of the aforementioned system is thus defined as the magnitude ratio of the cross-terms to the direct signal [2, 17]:

$$CHSP_i = 20\log_{10}\left(\frac{R_{ij}}{R_{ii}}\right) \quad [dB] \tag{1}$$

where R_{ii} and R_{ij} are the diagonal and off-diagonal elements of the matrix $\mathbf{R}$ respectively.

Now the minimum audible channel separation can be defined as the maximum relative level of crosstalk (R_{ij}/R_{ii}) at which the binaural signals v_i are still perceived to be equal to the desired binaural signals d_i.

To be able to evaluate the minimum audible channel separation, we need to have complete control over the amount of crosstalk that leaks into the desired signals. This can be simulated with headphones by adding the contralateral signals to the ipsilateral signals. Thus, in a simplified fashion, we can model the channel separation as a gain factor and a delay as shown in figure 2.

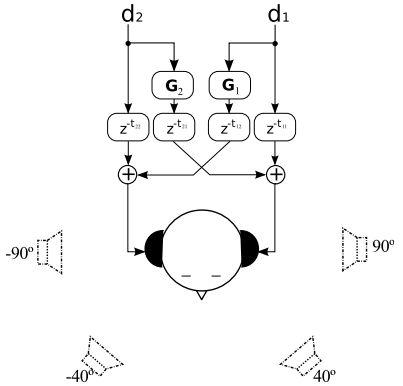


Fig. 2: Simplified model of simulated crosstalk through headphones. The amount of crosstalk is controlled with the gain factors G_i. The span angle and listener location are controlled with the delays τ_{ij}. The signals d_i are the desired binaural signals and the loudspeakers represent the simulated vitual sources.

Here $G_i = 10^{CHSP_i/20}$ is the gain factor, in which the CHSP level is given in dB. The delays τ_{ij} correspond to the de-

lays between channels. Note that this is a rather simplified approach, in which the transfer functions H_{ij} from the loudspeakers to the ears are not taken into account and thus a perfect equalization is assumed.

In an ideal case, when the listener is looking towards the center point between the two loudspeakers illustrated in figure 1, the channel separation $CHSP_i$ is symmetrical as well as the delays τ_{ij}. When the listener is not located at a symmetrical position there is a channel separation difference between the ears and the delays are not symmetrical either.

With the proposed model we can set the gains G_i and the delays τ_i in such a way that they correspond to the delay and gain factor difference of the desired span angle and the listener placement with respect to the virtual loudspeakers. In this manner, different span angles and listener's positions can be simulated.

3. METHODS

3.1. Stimuli and Binaural Synthesis

Eight different signals were assessed among which there were band-pass noise, 1-octave-band noises and speech. Each signal was convolved with head related transfer functions (HRTF) corresponding to sources located at $\pm 40^\circ$ and $\pm 90^\circ$ on the horizontal plane, where the negative angles correspond to the right side of the listener (see figure 2). The sign ($\pm$) of the source location was set randomly among stimuli and subjects. The HRTFs were obtained from a database containing artificial-head HRTFs measured with a 2° resolution [4]. These angles were chosen given that sound sources at those locations are easily externalized when reproducing them through headphones.

Additionally, two different span angles were simulated: 12° and 60°. This was done by setting the delays τ_{ij} and the gains G_i accordingly. Table 1 summarizes the stimuli used.

The reference stimuli were the binaural signals without crosstalk, i.e. $G_i = 0$ and $\tau_{ii} = 0$. Two different scenarios were simulated: one with the listener placed at the nominal center position and the other with the listener laterally displaced 5 cm to left or to the right away from the nominal center position. The direction of the displacement was set randomly among stimuli and subjects, under the assumption that results are symmetrical with respect to lateral displacements.

Stimulus	Frequency Band [Hz]	Duration [s]
Band-pass Noise	200 - 8000	1.2
1-Octave-band Noises	251.1 [a]	1.2
	501.1	1.2
	1000	1.2
	1995.3	1.2
	3981.1	1.2
	7943.3	1.2
Speech (Female voice saying the numbers "seven eight")	200 - 4000	1.4

[a] Center frequencies of the band-pass filters according to the IEC 1260:1995 and the ANSI S1.11:2004 standards.

Table 1: Summary of the stimuli assessed. All the stimuli were convolved with HRTFs corresponding to $\pm 40^\circ$ and $\pm 90^\circ$.

With these two scenarios, two span angles and two source positions we obtain in total sixty four different stimuli. All the stimuli were set to equal loudness using as a reference the 60 Phon curve defined in the international standard ISO-226. Before starting the test, subjects were instructed to adjust the reproduction volume to an audible and comfortable level.

3.2. Subjects

Thirty two subjects with normal hearing participated in these experiments (12 females and 20 males between 22 and 37 years old). Most of the subjects have some experience with listening test and discrimination procedures.

Assuming that the thresholds are normally distributed, to ensure that the 95% confidence interval is not larger than $\pm$3dB around the means[1], the minimum sample size should be 7 (see equation (3) in [6]). The subjects were randomly divided into four groups. Each group evaluated different stimuli, giving as a result of at least eight thresholds per stimulus. The order of presentation of the stimuli followed a Latin square design, in order to account for carry over effects [5].

3.3. Psychophysical Method

Thresholds were determined using an adaptive three-alternative forced-choice (3AFC) procedure. In each trial, three test stimuli were presented with 0.1s intervals between the stimuli. The reference stimulus was the raw

[1] This is assuming a standard deviation of 4dB observed in a pilot experiment

binaural signal without crosstalk. All the possible combinations of reference stimuli and signal with crosstalk were reproduced randomly. Each stimulus had a duration between 1.2s and 1.4s. For each trial, the order of the stimulus presentation was randomized and the subjects were asked to discriminate which of the three signals was the different one.

To measure the thresholds we used the simple adaptive testing algorithm with weighted up-down method proposed by Kaernbach in [9]. The advantage of this method is that it converges to any desired point of the psychometric function and is rather simple to implement. Note that with this method is not possible to draw a complete psychometric function because most of the observations are placed very close to the target level [14].

For the nAFC methods the threshold is often defined as the signal level at which the probability of correct responses is half way between perfect performance (100% correct answers) and probability of guessing (33,3% correct answers in the case of the 3AFC) [10]. With the rule 1-down-2-up, Kaernbach's method converges to the 66,6% of the psychometric function, which will correspond to the threshold of the 3AFC. This means that for each correct answer the channel separation level goes down one step and for each incorrect answer the channel separation level goes up two steps.

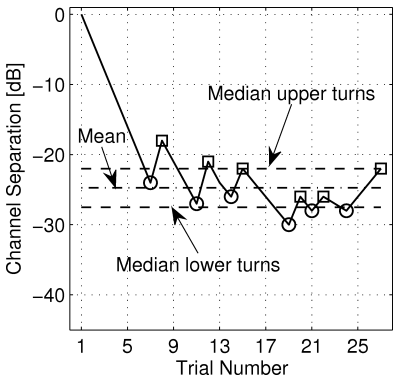


Fig. 3: Example of typical data obtained with an adaptive staircase algorithm. The upper turns are marked with a square and the lower turns are marked with a circle. The applied rule is 1-down-2-up.

There exists a number of ways to estimate the threshold from an adaptive up-down method. In [14] it is suggested that the mid-run estimates are rather robust, relatively efficient and result in low bias. A run is defined as a series of steps in on direction. A min-run estimate is calculated by taking the mean between the lower and upper turnarounds. However, we consider the medians to be more robust than means with respect to luck and lack of attention. Thus, we defined the threshold as the mean between the median of the lower and upper turnarounds. Figure 3 shows an example of typical data obtained from an adaptive staircase algorithm and the calculated threshold using the proposed estimate.

In order to ensure a faster convergence to the target level, we reduced the step size along the test [14]. The initial step size was set to 3 dB, after the third run it was reduced to 2 dB and after the sixth run it was finally reduced to 1 dB. The algorithm was stopped after 12 runs.

3.4. Procedures and Experimental Design

Part of these experiments were conducted at the Sound and Music Innovation Technology (SMIT) laboratories at the National Chiao-Tung University in Taiwan and the rest was conducted at the acoustic laboratories at Aalborg University. For the part conducted at the SMIT laboratories, we used a pair of dynamic headphones Sennheiser HD595 and for the part conducted at Aalborg University, we used a pair of open-cup headphones Beyerdynamic DT990. The impulse response of both headphones pairs were measured with a mannequin and equalized during the test.

The experiments were carried out in sound isolated listening rooms. Subjects had access to a graphical user interface in which for each sequence, as the stimuli were reproduced, the screen displayed buttons with the labels 1, 2 and 3 corresponding to each played sound. Subjects were instructed to click on the button associated with the sound they perceived to be different.

At the beginning of the experiment, subjects had a short training session in which they were introduced to the procedures of the test. During this training session the differences were made clearly audible (i.e. $G_i = 1$) and feedback was provided to the subjects. This was done to familiarize the subjects with the differences and to improve concentration. After the training session, the feedback was removed and they had the opportunity to repeat each sequence once to account for lack of attention. Subjects were encourage to repeat the sequence only when they had heard a difference but had forgotten which one of the

three was the different one. They were also instructed to focus on the location of the source.

The adaptive staircase algorithm, the graphical user interface, the playback and all data collection were implemented in Matlab.

The experiments were divided into four sessions per day. Each session consisted of two blocks corresponding to different stimuli each. There was a short break of about two minutes between blocks and a longer break of about ten minutes between sessions. After the test, the subjects were asked to describe what they thought the difference was.

4. EXPERIMENT I: LISTENER AT THE NOMINAL CENTER POSITION

In this experiment the minimum audible channel separation was measured for a listener placed at the center position. This means that we assumed $G_1 = G_2$, the delays $\tau_{ii} = 0$ and the delay $\tau_{12} = \tau_{21}$. Here the delays τ_{ij} were set to simulate span angles of 12° and 60° [13]. These values are based on measurements carried out at the acoustic laboratories at Aalborg University [12].

4.1. Results

In order to analyze the effects of all the possible factors (i.e. stimuli, span angle and source position) and their interaction, we carried out a multivariate analysis of variance (MANOVA) of the channel separation thresholds.

When the location of the listener was simulated at the nominal center position, the span angles did not show significant effects in the thresholds. On the other hand, there is a significant effect of the source location in the CHSP thresholds ($F_{(1,251)} = 25.43, p < 0.001$). The effect of the stimuli was also found to be significant at the 0.05 level ($F_{(7,245)} = 18.06, p < 0.001$).

Figure 4 shows the mean thresholds obtained for each stimuli at the two different simulated source positions: 40° and 90°. Since no significant differences were observed with respect to the span angles, the data is sorted by the location of the source. The mean thresholds for each source position are offset horizontally for visual clarity. The error bars correspond to the 95% confidence intervals.

We can observe in figure 4 a very clear trend: the mean thresholds for the speech (Sp), the band-pass noise (Bp) and the narrow-band noises centered at 251 Hz (Nn_{251}) and 501 Hz (Nn_{501}) are rather homogeneous and lie close to -20 dB. In contrast, the thresholds obtained with the narrow-band noises centered at 1 kHz (Nn_{1000}) and 2 kHz (Nn_{1995}) are larger than the rest of the stimuli and lie around -15 dB. At the higher frequency bands (Nn_{3981} and Nn_{7943}) the mean thresholds go down to approximately -25 dB when the sound source is at 90°.

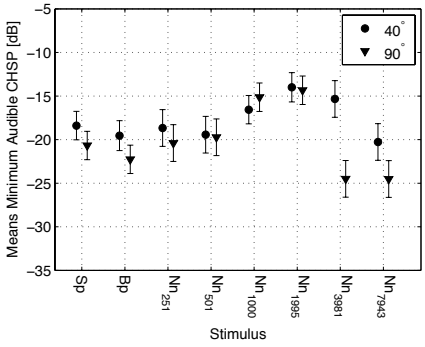


Fig. 4: Mean minimum audible channel separation (CHSP) as function of stimuli and source location. The crosstalk is simulated for the subject placed at the nominal center position. Thresholds for the stimuli placed at 40° (circle) and 90° (triangle) are offset horizontally for visual clarity. The error bars indicate the 95% confidence interval.

To support the observed tendency, we carried a pairwise comparisons between stimuli. We found the mean thresholds for the narrow-band noises centered at 1 kHz and 2 kHz to be significantly different than the thresholds observed with the other stimuli ($p < 0.001$). The narrow-band noise centered at 8 kHz is also significantly different to the other narrow-band noises and the speech signal ($p \leq 0.021$). The latter observation can be a consequence of the strong dependance on frequency of the interaural level differences (ILD) [3]. Above 1.6 kHz, the ILD increases systematically with frequency. When adding crosstalk to the binaural signals, we are directly affecting their natural ILD. This changes could be more audible at higher frequencies where a larger ILD is expected and this could be the reason why lower thresholds are observed with the narrow-band noises center at 4 kHz and 8 kHz.

Regarding the differences observed in the middle frequen-

cies (Nn_{1000} and Nn_{1995}), one thesis for the observed pattern is that in that frequency band our accuracy to discriminate angle differences decreases substantially [15]. This is due to the different mechanisms used by humans to localize sound sources. In the low frequencies region our auditory system makes use of phase and time differences to localize sounds, while in the high frequencies region it uses mainly the level differences. In [15] it is suggested that in the middle frequencies region, neither the phase nor the level differences are effective enough for localization. This could be the reason why the minimum audible channel separation is considerably smaller in this region.

There is also a clear difference between the mean thresholds obtained with the narrow-band noise centered at 4 kHz: the threshold obtained with the stimulus placed at 90° is significantly smaller (-24 dB) than the threshold obtained with the stimulus placed at 40° (-15.3 dB). It is well known that above 2 kHz the ILD is remarkably larger for sound sources placed at 90° than for sound sources placed at 40°, due to the head shadowing effect. Thus, when adding crosstalk to the narrow-band noise centered at 4 kHz the ILD for the source placed at 90° is significantly reduced. We can also observe - in a lesser extend - a similar pattern with the narrow-band noise centered at 8 kHz. However, we can notice that this effect does not occur so dramatically with the band-pass noise and the speech signal which also contain those frequency bands. This suggest us, that the changes in ILD at low frequencies - which do not vary much with source location - and the additionally low frequency cues present in the signals, mask somehow the changes in the ILD at higher frequencies.

Furthermore, there is a general tendency of larger thresholds when the source is placed at 40° than when it is placed at 90°. This tendency is clearly shown in figure 5 where the average of the thresholds as function of span angle and source position is plotted. Yet, we can notice that the differences are more pronounced with the 60° span angle. The mean difference between source location with the 60° span angle was found to be statistically significant at the 0.05 level ($F_{(1,251)} = 4.3, p = 0.039$).

5. EXPERIMENT II: LISTENER AT A LATERALLY DISPLACED POSITION

In this experiment the minimum channel separation was measured for crosstalk corresponding to a listener laterally displaced from the nominal center position. The

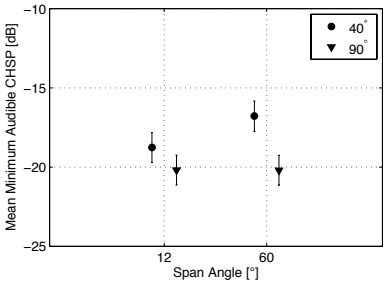


Fig. 5: Total average of the minimum audible CHSP level as function of span angle and source position. The mean thresholds for each source position are offset horizontally for visual clarity. The error bars indicate the 95% confidence interval.

delays and gain differences between channels were simulated for head positions corresponding to 5 cm to the left/right from the center position. The gain differences between G_1 and G_2 and the delays τ_{ij} were modeled based on the measurements of the channel separation of different loudspeaker arrangements [12, 13]. These values were set such that they correspond to the gain differences and the delays obtained with the 12° and 60° span angle configurations placed on the horizontal plane.

5.1. Results

In this scenario, we observed a particular pattern in the CHSP thresholds: some subjects consistently showed thresholds below -40 dB for some of the stimuli when the simulated span angle was 60°. The stimuli that showed this pattern more markedly were the band-pass noise and the narrow-band noise with center frequency at 501 Hz (Nn_{501}). This lead us to believe that with such a large "signal to noise ratio" (i.e. channel separation), what these subjects were discriminating was not the channel separation but the delay differences added to the signal with crosstalk. Based on this argument and in order to reduce the standard deviation of the data, we decided to divide the results for this scenario into two groups. The first group, which we will arbitrarily refer to as the "delay sensitive listeners", contains the thresholds obtained below -35 dB. The second group, which we will refer to as the "normal listeners", contains the thresholds obtained above -35 dB.

Figure 6 shows the mean CHSP thresholds obtained with the "delay sensitive listeners". We can not observed any

dependency with source position. Yet, as mentioned before, we can see that the stimuli that present the largest groups of "delay sensitive listeners" are the narrow-band noise centered at 501 Hz and the band-pass noise with five subjects each.

It is well known that at frequencies below 1.6 kHz, the interaural time difference (ITD) is the main mechanism that the human auditory system uses to localize a sound source. Since the delay differences were kept constant in this experiment, this supports the hypothesis that this particular group of subjects were discriminating the delay differences and not the channel separation differences.

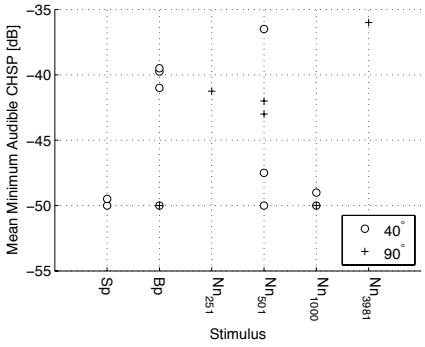


Fig. 6: Scatter plot of the CHSP thresholds obtained with the "delay sensitive listeners" as a function of the different stimuli and source location. The thresholds correspond to a simulated span angle of 60°. Thresholds for the stimuli placed at 40° (circle) and 90° (cross) are plotted.

In [13] the sweet spot of different loudspeaker configurations was evaluated as a function of changes in ITD. The results presented in that study show that the 60° span angle configuration is not robust to lateral displacements when looking at the temporal changes of the binaural signals. That analysis was done assuming a minimum audible ITD of 10μs. However, in [8] a large variance between subjects was observed, when evaluating the just noticeable differences in ITD. This could explain why some subjects were able to hear the delay differences whereas others were not.

Another possible explanation of this phenomenon, is that the "delay sensitive listeners" could have been able to discriminate the stimuli with crosstalk due to coloration introduced by a comb-filter effect. When superimposing the direct signal and a delayed version of itself we are creating what is known as comb-filter effect. Coloration due to comb-filter distortion can be audible if the delay is below 30 ms, depending on the signal content and the level differences [7]. However, when interviewed, none of the subjects reported to have perceived coloration or pitch differences between the stimuli.

A MANOVA of the CHSP thresholds obtained with the "normal listeners" was carried out. When excluding the "delay sensitive listeners" from the results, no significant differences were observed between span angles. Similarly to the results obtained in the experiment I, the effect of the source position on the CHSP thresholds was found to be significant ($F_{(7,224)} = 19.13, p < 0.001$) as well as the effect of the stimuli ($F_{(1,230)} = 14.52, p < 0.001$). The interaction between source position and stimuli showed to be also significant at the 0.05 level ($F_{(7,224)} = 2.34, p = 0.025$).

Figure 7 shows the mean CHSP thresholds obtained with the "normal listeners" for the different stimuli placed at 40° and 90°. The mean thresholds obtained with these two source locations are offset horizontally for visual clarity.

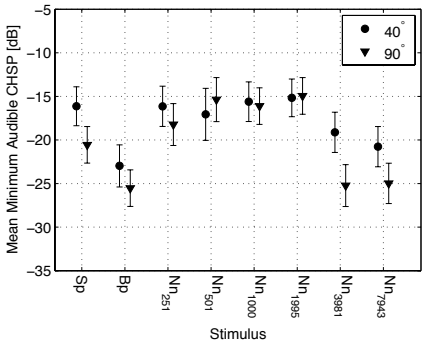


Fig. 7: Means of the minimum audible CHSP obtained with the "normal listeners" as a function of the different stimuli and source location. Thresholds for the stimuli placed at 40° (circle) and 90° (triangle) are offset horizontally for visual clarity. The error bars indicate the 95% confidence interval.

In contrast to the pattern observed in the experiment I, when carrying a pairwise comparison between stimuli, the

narrow-band noises with center frequencies below 4 kHz showed no significant differences. Yet, the band-pass noise was found to be significantly different to these stimuli and the speech signal ($p < 0.001$). The band-pass noise with crosstalk contains not only conflicting ILDs but also conflicting ITDs. Additionally, in this experiment the channel separation had different levels for each ear, which correspond to a listener displaced to the left or to the right from the nominal center position. These differences in channel separation increase also the changes in ILD. This might result in a large energy of conflicting cues (i.e. ILDs and ITD) for the band-pass noise, making discrimination of crosstalk easier for this stimuli than for the narrow-band stimuli at low frequencies and the speech signal.

Likewise the results obtained with experiment I, the mean thresholds for the narrow-band noises centered at 4 kHz and 8 kHz lie below 20 dB and are also significantly lower than the thresholds obtained with the narrow-band noises centered at lower frequencies.

There is also a clear difference between the mean CHSP thresholds obtained when the source was placed at 40° and when the source was placed at 90°. Similar to the trend observed with the CHSP thresholds obtained with the listener placed at the nominal center position, this difference is more pronounced with the high frequency narrow-band noises (4 kHz and 8 kHz). In this case, however, the location of the sound source shows also a significant difference with the speech signal.

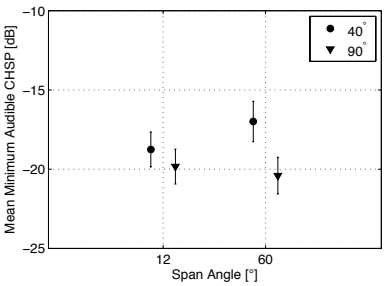


Fig. 8: Total average of the minimum audible CHSP obtained with the "normal listeners" as function of span angle and source position. The mean thresholds for each source position are offset horizontally for visual clarity. The error bars indicate the 95% confidence interval.

Figure 8 shows the average CHSP thresholds as a function of span angle and source position. Then again, the mean differences between source positions is more pronounced with the 60° span angle.

6. GENERAL REMARKS

During the experiments, subjects were asked to give their impressions about the stimuli. They were specifically asked what they thought the difference was when they could hear a difference. Most subjects described the difference as one sound being placed on one of their sides while the other sound was either inside the head or above. Some subjects described the differences as a change in distance.

In our simulated crosstalk, we were basically changing the ILD of the binaural signals. In both experiments the ITD was kept constant, thus being the changes in ILD the principal discrimination cue. Some experiments had shown that when changing the ILD or the ITD of binaural signals, a lateralization effect usually occurs, which causes the virtual image to move to the center of the head, especially when changes in ILD do not correlate with changes in ITD [3]. In the case of the "delay sensitive listener", when the ILD difference were not longer audible, this lateralization effect was still present due to the audible differences in ITD.

7. CONCLUSIONS

This paper presents a subjective evaluation of the minimum audible channel separation for binaural reproduction systems through loudspeakers. Using a simplified model, the crosstalk was simulated by adding to the direct signals the cross-terms attenuated by a gain factor and a delay. The sounds were then reproduced through headphones. Two listening experiments were carried out in which the minimum audible channel separation was evaluated for the listener placed at the nominal center position and the listener laterally displaced 5 cm to the left/right of the center position. Eight different stimuli were evaluated, among which there were speech, band-pass noise and 1-octave-band noises with center frequencies between 0.2 and 8 kHz. Each stimulus was convolved with 40° and 90° HRTFs. Two different span angles were simulated. This resulted in a total of sixty four different stimuli. The minimum audible channel separation was measured using a 3AFC discrimination procedure and the simple adaptive algorithm with weighted up-down method [9].

The results obtained from the first experiment showed a clear pattern in which the channel separation thresholds for the narrow-band noises centered at 1 kHz and 2 kHz are significantly larger than the thresholds obtained with the rest of the stimuli. Additionally, the minimum audible CHSP obtained with the 4 kHz narrow-band noise placed at 90° and 8 kHz narrow-band noise placed at 40° and 90° were significantly smaller than the rest of the stimuli. At frequencies above 1.6 kHz, the ILD is generally large and it is known to be the main binaural cue used by humans to localize sound sources. Thus, reductions in ILD caused by the crosstalk at high frequencies might lead to lateralization of the virtual image in a greater extend than changes in ILD at lower frequencies. This could explain why at high frequencies the differences in ILD are easier to discriminate than at low frequencies. We also believe that when a broad-band stimuli such as the band-pass noise is used, the differences in ILD at high frequencies are somehow masked by the binaural cues present at lower frequencies, making discrimination of crosstalk slightly more difficult with broad-band noises when the listener is placed at the nominal center position.

In the middle frequencies (1 kHz and 2 kHz), where the main localization mechanism changes from phase differences to level differences, our localization accuracy decreases [15]. This can explain why the crosstalk showed to be more difficult to discriminate in this region than at lower or higher frequencies.

In the second experiment we observed a particular trend: some subjects could consistently discriminate the stimuli with channel separation below -40 dB when the simulated span angle was 60°. This suggests us that the introduced differences in ITD make the discrimination still possible when the crosstalk should not be audible anymore. This is in agreement with results presented in [13] in which the sweet spot size as a function of temporal changes was evaluated for different loudspeaker configurations. In that study it was shown that the 60° span angle is not robust to lateral displacements when evaluating the ITD changes. Those results are based on a minimum audible ITD of $10\mu s$. However, a large variance of the minimum audible ITD has been observed when evaluating it with naive listeners [8]. We believe that is the reason why not all the subjects were able to discriminate the differences in ITD.

On the other hand, audibility of coloration due to comb-filter distortions can also be a possible explanation of the thresholds obtained with the "delay sensitive listeners". Even though none of the subjects reported to have perceived coloration or pitch differences during the test, we consider that this possibility should not be completely discharged.

Sound sources placed at 90° showed a general tendency of smaller thresholds than the sound sources placed at 40°. This differences are more pronounced at high frequencies and with the 60° span angle. The ILD at 90° is significantly larger than the ILD at 40°, due to the natural head shadowing effect. This is especially observed at high frequencies. Thus, it is hypothesized that the crosstalk in a sound source placed at 90° is easier to discriminate than when the source is at 40° due to the significant changes in ILD at high frequencies. Additionally, with the 60° span angle the delay differences between the channels is larger than with the 12° span angle. This could act as an additional cue used to discriminate the signal with crosstalk.

Disregarding the group of subjects that could discriminate the differences in ITD, the results obtained with the listener laterally displaced from the nominal center position follow similar trends to the results obtained with the listeners placed at the nominal center position. However, in this case we did not observe significant differences between the narrow-band noises centered at the middle frequencies (1 kHz and 2 kHz) and the narrow-band noises centered at lower frequencies. Yet, the band-pass noise showed lower thresholds in this scenario in comparison with the thresholds obtained at the nominal center position.

In summary, we could observe that the minimum audible channel separation lies below -15 dB for most of the evaluated stimuli. Furthermore, in the case of broad-band signals such as the band-pass noise employed in the experiment, the minimum audible channel separation lies around -20 dB when the listener is at the nominal center position and -25 dB when the head is laterally displaced. Previously, it has been suggested to set the maximum channel separation to -12 dB or -10 dB [2, 17]. However, the results obtained from this study suggest us that those limits could be rather relaxed. In most applications of binaural systems, the reproduced signals are either speech or broad-band signals in general. Therefore, according to the results presented in this study, the channel separation limits should be set below or around -20 dB instead.

Most of the subjects described the stimuli with crosstalk as being closer to, above or inside their heads. This is in agreement with lateralization experiments, in which changes in ILD or ITD causes the virtual image to fall inside the head. It is not expected that the virtual images will fall inside the head when reproduced through

loudspeakers, but rather that they will be placed at the location of the loudspeakers. Such an effect can have unfortunate consequences in binaural reproduction systems through loudspeakers. If for example a virtual environment is simulated and some of the images happen to be wrongly placed at the loudspeakers position due to insufficient channel separation at some key frequencies, the whole virtual experience can be degraded to a large extend. Hence, if a proper virtual reproduction is desired, care should be taken when designing the crosstalk cancellation filters and sufficient channel separation should be allowed in the target frequency band.

8. ACKNOWLEDGMENT

Part of this work was carried out at the Sound and Music Innovation Technology (SMIT) laboratories at the National Chiao-Tung University, Taiwan. The authors would like to thank the valuable input received from professor Mingsian Bai and the help given by his students with the experiments.

9. REFERENCES

[1] Mingsian R. Bai and Chih-Chung Lee. Development and Implementation of Cross-Talk Cancellation System in Spatial Audio Reproduction Based on Subband Filtering. *Journal of Sound and Vibration*, 290:1269–1289, August 2005.

[2] Mingsian R. Bai and Chih-Chung Lee. Objective and Subjective Analysis of Effects of Listening angle on Crosstalk Cancellation in Spatial Sound Reproduction. *Journal of the Acoustic Society of America*, 120(4):1976–1989, October 2006.

[3] Jens Blauert. *Spatial Hearing*. Hirzel-Verlag, 3rd edition, 2001.

[4] Bjarke P. Bovbjerg, Flemming Christensen, Pauli Minnaar, and Xiaoping Chen. Measuring the Head-Related Transfer Functions of an Artificial Head with a High Directional Resolution. In *109th Convention of The Audio Engineering Society*, page 5264, Los Angeles, CA, September 22-25 2000.

[5] George E. P. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for Experimenters: design, discovery, and innovation*. Wiley-Interscience, 2nd. edition, 2005.

[6] John Eng. Sample Size Estimation: How Many Individuals Should Be Studied? *Radiology*, 227(2):309–313, 2003.

[7] Helmut Haas. Influence of a Single Echo on Audibility of Speech. *Journal Audio Engineering Society*, 20(2):146 – 159, 1972.

[8] Pablo F. Hoffmann and Henrik Møller. Audibility of Differences in Adjacent Head-Related Transfer Functions. *Acta Acustica united with Acustica*, 94:945–954, 2008.

[9] Christian Kaernbach. Simple Adaptive Testing With the Weighted Up-Down Method. *Perception & Psychophysics*, 49:227–229, 1991.

[10] Christian Kaernbach. Adaptive Threshold Estimation with Unforced-choice Tasks. *Perception & Psychophysics*, 63(8):1377–1388, 2001.

[11] Ole Kirkeby and Philip A. Nelson. Digital Filter Design for Inversion Problems in Sound Reproduction. *Audio Engineering Society*, 47(7/8):583–595, July/August 1999.

[12] Yesenia Lacouture Parodi and Per Rubak. Objective Evaluation of the Sweet Spot Size in Spatial Sound Reproduction Using Elevated Loudspeakers. *Journal of the Acoustical Society of America*, 128(3), September 2010.

[13] Yesenia Lacouture Parodi and Per Rubak. Sweet Spot Size in Virtual Sound Reproduction: A Temporal Analysis. In *Principles and Applications of Spatial Hearing*. World Scientific, in press.

[14] H. Levitt. Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, 49(2 (Part 2)):467–477, 1971.

[15] A. W. Mills. On The Minimum Audible Angle. *Journal Acoust. Soc. Am.*, 30(4):237 – 246, 1958.

[16] Philip A. Nelson, Felipe Orduña-Bustamante, and Hareo Hamada. Inverse Filter Design and Equalization Zones in Multichannel Sound Reproduction. *IEEE Transactions on Speech and Audio Processing*, 3(3):185 – 192, May 1995.

[17] John Rose, Philip Nelson, Boaz Rafaely, and Takashi Takeuchi. Sweet Spot Size of Virtual Acoustic Imaging Systems at Asymmetric Listener Locations. *Journal of the Acoustic Society of America*, 112(5):1992–2002, November 2002.

[18] Darren B. Ward. Joint Least Squares Optimization for Robust Acoustic Crosstalk Cancellation. *IEEE Transactions on Speech and Audio Processing*, 8(2):211–215, February 2000.

6 MANUSCRIPT E

Lacouture Parodi, Y. and Rubak, P. (2010), **Binaural reproduction system using three stereo-dipoles**, *Journal of the Audio Engineering Society*, submitted.

Portions of this manuscript were presented in "Evaluation of a Binaural Reproduction System Using Multiple Stereo-Dipoles", Proceedings of the 128^{th} Convention of the Audio Engineering Society 2010, May 22-25, London, UK.

Binaural reproduction system using three stereo-dipoles*

Yesenia Lacouture Parodi[1] and Per Rubak[1]

[1]*Aalborg University, Aalborg, DK-9220, Denmark*

Correspondence should be addressed to Yesenia Lacouture Parodi (ylp@es.aau.dk)

ABSTRACT

The sweet spot size of different loudspeaker configurations was investigated in a previous study carried out by the authors. Closely spaced loudspeakers showed a wider control area than the standard stereo setup. The sweet spot with respect to head rotations showed to be especially large when the loudspeakers are placed at elevated positions. In this paper we describe and evaluate a system that attempts to make use of the robustness to head rotations of the loudspeakers placed above the listener combined with the wide sweet spot that closely spaced loudspeakers exhibit. Three pairs of closely spaced loudspeakers comprise the proposed system: one pair placed in front, one placed behind and one placed above the listener. The system is based on the idea of dividing the sound reproduction into regions to reduce front-back confusions and enhance the virtual experience without the aid of a head tracker. A set of subjective experiments with the intention of evaluating and comparing the performance of the proposed system are described. The results indicate that a reduction of front-back confusions is obtained when the three-loudspeaker pair system reproduces the virtual images. However, there are still some pending issues regarding coloration changes between the loudspeaker pairs and the errors introduced by the used of non-individual head related transfer functions and the regularization at high frequencies.

1. INTRODUCTION

One of the biggest limitations of binaural reproduction systems through loudspeakers is the rather narrow sweet spot. In order to counteract the crosstalk (i.e. the signals that are to be heard in one ear are heard in the other) it is necessary to introduce proper inverse filters into the reproduction chain. Those filters are usually designed for a fixed position. Thus, head movements can easily introduce significant errors in the reproduction, destroying in that way the virtual image.

The sweet-spot is usually defined as the area in which the amount of head movements is within the maximum allowed such that the introduced errors are negligible. In the past, different loudspeaker configurations had been proposed in order to widen the sweet-spot [2, 3, 7, 10, 11, 22]. However, the overall sweet-spot remains rather narrow and most of the proposed configurations show some limitations with respect to head rotations. This is, when the listener rotates his/her head, a sufficient amount of rotation can make the loudspeakers to be located in the same side of his/her head. As a result, the virtual images will be wrongly placed at the loudspeakers location.

The problem of narrow sweet-spot is often counteracted by introducing a head tracker and dynamic crosstalk cancellation filters into the system. However, such a system requires extra computational power, is usually expensive and unpractical in some cases.

Front-back confusions is also an issue that has not being solved in binaural reproduction systems through loudspeakers. Most researchers agree on that a significant reduction of the front-back confusions is obtained when the sound pressures at the ears changes due to head rotations [24, 9, 19]. But as mentioned before, when the head is rotated the virtual acoustic image can be easily destroyed in the absence of a head tracker.

In [7] a system including two pairs of loudspeakers, one in the front and one in the back of the listener is proposed. The way the system approaches the problem is by letting the frontal and rear loudspeakers reproduce only the images situated in their respective hemisphere and by adding

*Portions of this work were presented in "Evaluation of a Binaural Reproduction System Using Multiple Stereo-Dipoles," Proceedings of the 128th Convention of the Audio Engineering Society 2010, 22–27 May, London, UK

cross-fading to account for the transitions from the front to the back. Thus head rotations will naturally place the images in their respective hemisphere. However, this does not apply for sound images located above, below and at the sides of the listener.

The sweet spot size of several different loudspeaker configurations was analyzed in a previous study conducted by the authors [14, 15]. The sweet spot was evaluated with respect to changes in the magnitude ratios between crosstalk and the direct signals, as well as the changes of the temporal cues. These analysis were done for lateral displacements, frontal displacements and head rotations. The results obtained in those investigations showed a large agreement with previous studies in which closely spaced loudspeakers result in a wider sweet spot than the standard stereo configuration [11, 2]. It was additionally found that when placing the loudspeakers above the head a rather large sweet spot is obtained with respect to head rotations. With respect to lateral displacements and the changes of temporal cues, the largest control area was obtained when closely spaced loudspeakers were placed on the horizontal plane.

In this paper we evaluate a system that combines the properties of the loudspeakers placed above the head, on the horizontal plane and the idea of dividing the sound reproduction into regions. Here, we describe a binaural reproduction system using three pairs of closely spaced loudspeakers: one pair placed in front, one placed in the back and one placed above the listener [18]. The virtual reproduction is thus divided into three zones and each pair of loudspeakers will only reproduce the virtual images placed in their proximity. This is, the frontal loudspeakers will reproduce sources placed in front of the listener, the rear loudspeakers will reproduce sources placed in the back of the listener and the upper loudspeakers will reproduce sources located above, on the sides and below the listener. A linear crossfading between regions is also introduced in order to avoid artifacts due to sudden changes between zones [16].

This paper is organized as follows: Section 2 presents a description of the proposed system. In section 3 the implementation of the system is outlined. In section 4 two subjective experiments carried out to assess the performance of the proposed system are described and the results are discussed. A general discussion of the results observed and the limitations of this study are summarized in section 5.

2. THREE STEREO-DIPOLE SYSTEM

In the past decades, it has been broadly shown that when using two closely spaced loudspeakers - the so called stereo-dipole [11] - the control area of a binaural reproduction system is considerably widened [2, 14, 15, 21]. One of the reasons for this is that when the listener moves his/her head there is less variation of the path lengths between the loudspeakers and the ears with closely spaced loudspeakers than with loudspeakers placed farther apart. However, when the stereo-dipoles are placed on the horizontal plane, small head rotations can make the loudspeakers to be located in the same side of the listener's head. In this situation, even with sufficient crosstalk cancellation, the virtual images will be misplaced at the loudspeakers position as a consequence of the precedence effect or the law of the first wavefront [4].

In [14] and [15] we observed that when placing the loudspeakers above the head the system is especially robust to head rotations. This is due to the fact that the location of the loudspeakers relative to the head does not change significantly when the listener rotates his/her head. Thus, if we combine these characteristics and the idea of dividing the reproduction of the sound into zones we might be able to obtain a system that could be robust enough to head rotations and in that way we can enhance the virtual experience without the aid of a head tracker.

Therefore, we propose a binaural reproduction system that makes use of three stereo-dipoles (TSD) placed in the front, back and above the listener [18]. Figure 1 illustrates this idea.

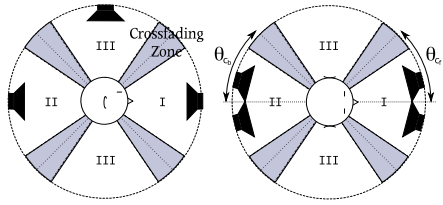


Fig. 1: Scheme of the different zones in the three stereo-dipoles (TSD) setup. $2\theta_{c_i}$ is the aperture of the cones that define the limits for the different zones.

The virtual reproduction is thus divided into three zones as indicated in figure 1 by the roman numbers I, II and III. Each loudspeaker pair will thus reproduce the sound

sources that are placed at their respective zone. This is, the frontal loudspeakers will reproduce sources placed in zone I, the rear loudspeakers will reproduce sources placed in zone II and the upper loudspeakers will reproduce sources placed in zone III.

A sudden change between zones migth introduce audible artifacts to the sound [16]. Thus, we need to define a crossfading region between each zone. These crossfading regions are defined by the intersections with a sphere of the right circular cones with aperture $\theta_{c_{f1}}$ and $\theta_{c_{f2}}$ on the frontal region and $\theta_{c_{b1}}$ and $\theta_{c_{b2}}$ on the rear region as depicted in figure 2. The head of the listener is placed right in the center of the sphere.

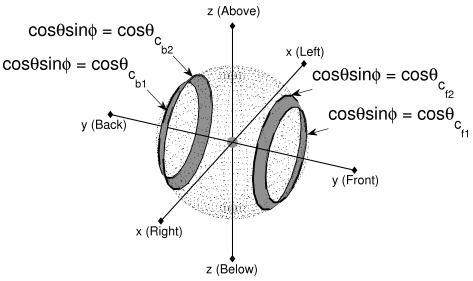


Fig. 2: Definition of the crossfading zones in a sphere around the listener's head. The smaller sphere in the center represents the listener's head. The gray areas indicate the crossfading zones.

The intersection between a right circular cone and a sphere is a circle. In our case, these circles are defined in spherical coordinates by

$$\cos\theta\sin\phi = \cos\theta_{c_{fi,bi}} \tag{1}$$

To do the crossfading we propose to apply a weighting function to the gain of the crosstalk cancellation filters, depending on the circles placed inside the crossfading region. This is, the circles defined by the inequality

$$\cos\theta_{c_{f1,b1}} \leq \cos\theta\cos\phi \leq \cos\theta_{c_{f2,b2}} \tag{2}$$

Hence, the gain weighting functions in the crossfading regions can be defined as a function of circles as follows:

$$W_{f,b}(\phi,\theta) = \begin{cases} 0 & \Psi(\phi,\theta) > \Omega_1 \\ \frac{\Psi(\phi,\theta)-\Omega_1}{\Omega_2-\Omega_1} & \Omega_1 \leq \Psi(\phi,\theta) \leq \Omega_2 \\ 1 & \Psi(\phi,\theta) < \Omega_2 \end{cases} \tag{3}$$

where $\Omega_i = \cos\theta_{c_{fi,bi}}$ define the circles that limit the crossfading regions, the function $\Psi(\phi,\theta) = \cos\theta\sin\phi$ corresponds to the circle in which the virtual image is located and the subscripts f and b stand for front and back respectively. This is equivalent to a linear crossfading between circles.

Figure 3 presents a simplified block diagram of the three dipoles system with the proposed linear crossfading. CTC stands for crosstalk cancellation filters and $x(\theta,\phi)$ is a vector with the desired binaural signals and an extra channel with information about the instant location of the source.

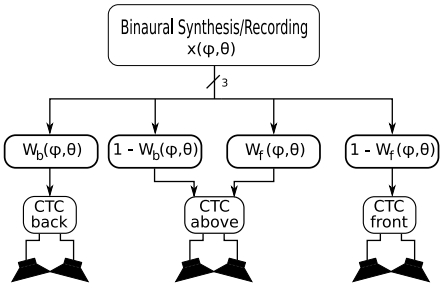


Fig. 3: Simplified block diagram of the proposed three stereo-dipole (TSD) system. CTC stands for crosstalk cancellation filters.

3. SYSTEM IMPLEMENTATION

As shown in figure 3, the TSD system has independent crosstalk cancellation filters for each stereo-dipole pair. These filters are calculated by using the transfer functions from each loudspeaker to the ears of an artificial head. For this purpose, we made use of Valdemar, the artificial head designed at the acoustic laboratories at Aalborg University [6].

There exists a number of methods to calculate crosstalk cancellation filters. In [14] and [12] three different crosstalk cancellation methods were evaluated. One of the methods - referred to as the generic crosstalk canceler - is based on the exact matrix inversion and calculates the inverse filters by employing the minimum-phase components of the impulse responses from the loudspeakers to the ears. The excess-phase components of the impulse responses are modeled as a frequency independent delay.

The other two methods are based on a least square approximation, one in the frequency domain and the other in time domain.

The performance of the methods has been first evaluated assuming the listener placed at symmetric positions with respect to the loudspeakers and disregarding the loudspeakers impulse responses [12]. In that study it was shown that in general the least square methods result in better channel separation and less errors than the generic crosstalk canceler.

In [14], the performance of the above mentioned methods was evaluated at asymmetric positions and taking into account the reproduction chain's impulse response. In that study, no significant differences between the methods with respect to sweet spot size were observed when using closely spaced loudspeakers.

Based on those results, we decided to use the least square method in the frequency domain to calculate the crosstalk cancellation filters. We chose this method given that its implementation is fairly simple and does not require as much memory and computation power to calculate the filters as the least square method in the time domain.

Now, the transfer functions from the loudspeakers to the ears present steep notches around and above 8 kHz. The inversion of such a system will thus contain large peaks around those frequencies. Additionally, there is usually a roll-off at low frequencies due to the loudspeakers drivers inherent limitations. Thus, we need to incorporate a frequency dependent regularization in order to control the gain of the inverse filters and avoid singularities. We modeled the regularization as a shape factor and a gain factor. The shape factor corresponds to the absolute magnitude response of an FIR stop-band filter, with stop-band frequencies at 500 and 6000 Hz. The gain factor is set such that the average power of the crosstalk cancellation filters is approximately the same for all the configurations. This constraint is set under the assumption that by ensuring similar power, the audibility of differences between the filters will be decreased. When choosing the proper regularization constant, it was also ensured that the gain of the filters did not exceed 12 dB, in order not to overdrive the loudspeakers. Previous experiments had showed that sufficient crosstalk cancellation is obtained with that limitation. A detailed description of the methods and their implementation can be found in [14] and [12].

Even though a real-time implementation of the TSD is feasible, to evaluate its performance we decided to do all the processing off-line for practical reasons, i.e. the crosstalk cancellation filters, the binaural reproduction and the crossfading algorithm. All the processing and reproduction was implemented in Matlab.

4. SUBJECTIVE EVALUATION OF THE TSD

In order to evaluate our hypothesis that an improvement in localization performance as well as a significant reduction of front-back confusions is achieved with the TSD system, we carried out a set of subjective experiments. First, we evaluated the localization performance with each loudspeaker pair independently and estimated the appropriate crossfading region. Then, we evaluated the overall performance of the TSD with dynamic sources.

4.1. Subjective Experiment I

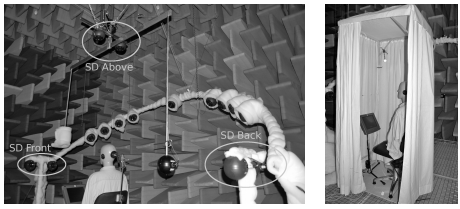
There exists a number of studies on sound localization with stereo-dipoles placed on the horizontal plane [21, 20, 23]. In general, it has been observed that virtual images outside the range of the loudspeakers are incorrectly placed due to the lack of the natural high frequency separation provided by the head shadowing [2]. Thus by intuition, it seems natural to define the crossfading region in accordance with those observations, i.e. $\pm 30^{\circ}$. However, a clear boundary where the stereo-dipoles are effective has not been properly defined - to the best of our knowledge. Additionally, sufficient information on sound localization when the stereo-dipoles are placed above the listener's head has not been found in the literature known by the authors.

Thus, the objective of this experiment is to evaluate the region in which each single stereo-dipole pair is effective reproducing binaural signals. This is with the aim to define the optimum crossfading region. Additionally, the influence of head rotations on the localization performance of virtual sources when reproduced through each of the three loudspeaker pair is also investigated.

4.1.1. Apparatus

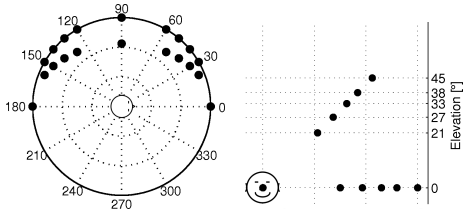
A systematic localization experiment with static sources was carried out. The localization performance of virtual sources reproduced by each of the three stereo-dipoles were evaluated and compared with the localization performance with real sources. For this purpose, binaural recordings were made with Valdemar placed inside the anechoic chamber where the experiment took place.

Figure 4 shows the setup of the experiment. A half circular arc of 170 cm radius is placed inside the anechoic



(a) (b)

Fig. 4: Experimental setup. a) Arc with static sources and the three stereo-dipoles (SD front, SD above and SD back) highlighted by the gray circles, b) Cabin surrounded by an acoustically transparent curtain where the subjects sat.



(a) Top view (b) Frontal view

Fig. 5: Diagram of the tested directions. a) Top view, b) Frontal view

chamber and is mounted on two supporting poles which make it possible to rotate the arc. In this way, different sound sources at different azimuth and elevations can be reproduced. As can be seen in the picture, most of the loudspeakers are concentrated in the hypothesized cross-fading area, i.e. between 30° and 60° azimuth if the arc is placed on the horizontal plane.

In order to evaluated different elevation angles, the arc was placed at 0° elevation (horizontal plane) and 45° elevation. The latter placement results in six different elevation angles. Sources on the median plane were evaluated in a pilot experiment. Subjects found this localization rather difficult - even with real sources - and most of them reported to feel frustrated with those specific sessions. Therefore, in order to reduce the length of the test and the frustration of the subjects, we decided not to include the median plane in the final test. Figure 5 summarizes the locations of the evaluated sources viewed from the top and from the front.

The frontal and rear stereo-dipoles are placed at the arc and the upper stereo dipoles are placed on a supporting pole located right above the center of the circumference (highlighted in figure 4(a) by the gray circles). The span angles of all three dipoles is 12° and they are placed at a distance of 150 cm from the center of the circumference.

Subjects were surrounded by an acoustically transparent curtain as shown in figure 4(b) and were blind folded before entering the chamber. Thus, they had no knowledge of the location of the sources. An adjustable chair with a head rest was placed inside the cabin to ensure that all subjects heads were positioned correctly.

The orientation of the cabin could be also changed, hence sources at both sides of the listener could be reproduced. However, in order to reduce the length of the test, the sources were randomly distributed to the left or to the right among subjects. This means that half of the subjects evaluated the sound sources placed at their left while the other half evaluated the sources placed at their right. When doing this distribution is was also ensured that each subject listened to sound sources at both sides to reduce bias and tiredness. In order to maintain a proper sample size for each stimulus, in the analysis of the localization results we assumed symmetry with respect to the median plane. Thus, the localization results obtained with the sources at the left of the listeners are combined with the respective symmetric location at the right of the listeners.

4.1.2. Stimulus

We used pink noise as a base signal for the stimuli. The signal was band limited between 200 Hz and 8 kHz. A raised cosine with 300 ms onset and offset was applied to each stimulus.

The stimuli consisted on a reference sound and the test sound. The reference sound had a duration of 1 second and was located at 0° azimuth and 0° elevation, i.e. right in front of the listener. There was a pause of 1 second between the reference and the test signal. The test signal had a duration of 2 seconds. Subjects were informed about the location of the reference sound prior to the test. The reference sound was introduced not only to familiarize the subject with the stimulus spectrum, but also because in pilot experiments we observed that when using the reference sound, subjects found the location task easier. All the stimuli were reproduced at nominal level of 65 dBA at the ear position.

4.1.3. Experimental design and Localization method

This experiment was divided into four scenarios: 1) real source - head fixed, 2) virtual source - head fixed, 3) real source - head rotations, 4) virtual source - head rotations. In order to control the head movements, three markers were placed inside the cabin. One marker was placed at 0° azimuth 0° elevation (right in front of the listener) and the two other markers were placed at $\pm 20^\circ$ azimuth also on the horizontal plane. In the head fixed scenario, subjects were asked to point their noses towards the marker right in front of them and to keep their head still while each stimulus was reproduced. In the head rotation scenario, subjects were asked to point their nose towards the marker in the front and keep their head still while the reference sound was reproduced. Once the reference sound had ended, they were asked to turn their head towards the markers placed at $\pm 20^\circ$ while the test sound was reproduced.

The experiment was divided into a training session and three main sessions of two hours each, distributed over different days. Each session consisted on six blocks of approximately ten minutes each. Subjects had breaks between blocks and during the breaks they were asked to fill out a questionary with comments about the experiment. Each block consisted of one of the four aforementioned scenarios and sources placed either on the horizontal plane or at different elevations. In the virtual sources scenarios, each loudspeaker pair was evaluated in separate blocks. The presentation order of the blocks was randomized using a Greco-Latin square design. The order of the stimulus presentation in each block was also randomized.

Inside the cabin, subjects had access to a touch-screen where they could interact with the user interface. Before each stimulus was reproduced, they had a reminding of the correct placement of the head, i.e. either to keep their head still or to rotate it while the test sound was reproduced. They could push play whenever they felt ready for the next stimulus and could repeat each stimulus as many times as they wanted in case they were not sure about the location of the source. After each stimulus was reproduced they were asked to select first the perceived azimuth angle on a two-dimensional diagram with markers every 10° displayed on the screen. Then they had to select the perceived elevation angle on a diagram of similar characteristics. Subjects were trained for at least two hours to get familiarized with the localization paradigm and the stimuli spectrum.

4.1.4. Subjects

Seven paid subjects (3 females and 4 males) between 23 and 33 years old participated in the test. The subjects' hearing thresholds were checked using a standard pure-tone audiometry in the frequency range from 250 Hz to 8 kHz and it was ensure that their thresholds were not larger than 20 dB HL.

Some of the subjects had previous experience with localization experiment and all of them participated in the pilot experiment. The pilot experiment was also used as a screening procedure and only subjects who where able to localize the real sound sources within the expected error could continue with the rest of the experiment. Thus, we consider that the subjects who participated have enough experience with the localization paradigm and are rather good localizers. Subjects received written and oral instructions before starting the test.

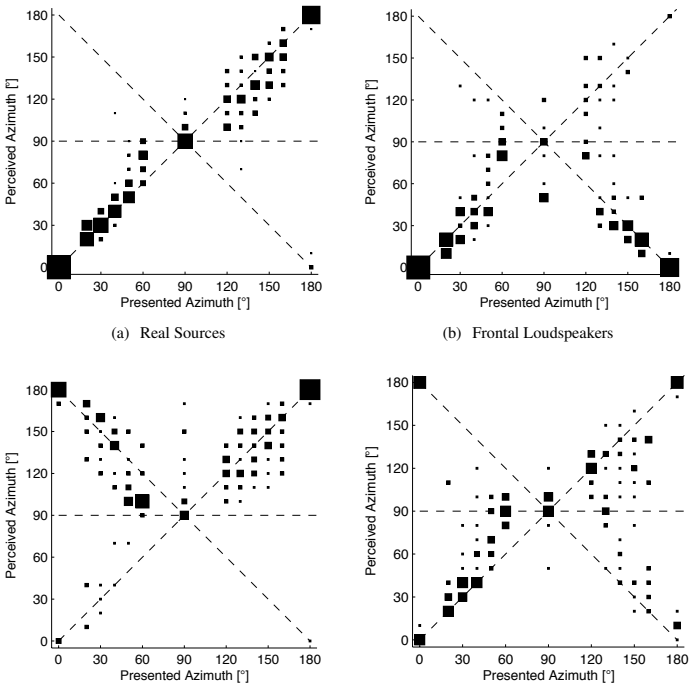
Given that the resolution of the test is 10°, we would like to ensure that the 95% confidence intervals are not larger than $\pm 10^\circ$. Thus, according to equation 3 in [8] the minimum sample size should be $N = 21$. This is assuming that the average standard deviation of the localization experiment is 23°, observed in the pilot experiment. Therefore, the test was designed so each subject evaluated each stimulus three times, resulting in 21 samples per stimulus.

4.1.5. Results

The results of the localization experiment when the listeners were instructed to keep their head still are presented in figures 6 and 7 in terms of presented azimuth/elevation versus perceived azimuth/elevation. The area of the squares is proportional to the amount of subjects that perceived the sound coming from the same location.

Figure 6(a) shows the azimuth localization performance with real sources. Fewer front-back confusions than expected are observed. For azimuth angles larger than 50° we can observe a slight bias towards the frontal plane. As expected, the localization performance of rear sources is poorer than the frontal sources. Nevertheless, we could say that the error is still within the expected range since subjects were not allowed to look towards the source while it was played.

The localization accuracy of the human auditory system decreases as the sources approaches the frontal plane [17]. This could explain the situations in which subjects confused sources placed at 60° or 120° with sources at 90°, i.e. right at their side. Additionally, we observed



(a) Real Sources (b) Frontal Loudspeakers

(c) Rear Loudspeakers (d) Upper Loudspeakers

Fig. 6: Presented azimuth versus perceived azimuth with real and virtual sources. The head of the listener is at the optimal position and is kept still while the stimuli were reproduced. The area of the squares is proportional to the amount of subjects that perceived the same angle. The dashed lines correspond to the perfect localization (45° line), front-back confusions (−45° line) and frontal plane (constant 90° line). a) Real sources, b) Virtual sources reproduced by the frontal stereo-dipole, c) Virtual sources reproduced by the rear stereo-dipole, d) Virtual sources reproduced by the stereo-dipole above the head.

this bias more markedly when the sources were at different elevations, where our localization accuracy tends to decrease even more.

Looking at the azimuth localization performance with the virtual sources we can observe a large amount of back-front and front-back confusions when the virtual sources are reproduced by the frontal and rear loudspeakers respectively (figures 6(b) and 6(c)). The localization error of the virtual sources with azimuth angles bellow 40° reproduced by the frontal loudspeakers is comparable with the error observed with real sources. Likewise the localization error of virtual sources with azimuth angles above 120° reproduced by the rear loudspeakers is within the expected value. There is again a bias towards the frontal plane for azimuth angles larger than 50° which is still comparable to the accuracy observed with the real sources.

When the virtual sources are reproduced by the upper loudspeakers, fewer front-back confusions are observed (figure 6(d)). However, we can see a larger bias towards the frontal plane for virtual sources with azimuth angles larger than 50° in comparison to the frontal and rear loudspeakers.

Figure 7 shows a large variability in the elevation localization performance not only with the virtual sources but also with the real sources. There is a clear bias towards the horizontal plane for the real sources and the virtual sources reproduced with the frontal loudspeakers (figures 7(a) and 7(b)). The rear and upper loudspeakers seem to produce a better perception of elevation (figures 7(c) and 7(d)). Yet, there is still a large variability among the responses.

It is well known that our accuracy to estimate the elevation of a sound source is rather poor, especially when there are no visual cues. In this experiment, subjects could neither see the location of the sources nor had markers as elevation reference. Subjects reported to find it difficult to estimate the elevation angle, even though they perceived a source coming from an elevated position.

Additionally, it has been shown that the main cues for judgement of elevation lie at frequencies above and around 5 kHz and that the power within 5 to 10 kHz relative to the power of frequencies below and above that band give additional elevation cues [1]. The stimuli used in this experiment were band limited to 8 kHz and thus the elevation cues above that frequency were omitted. This could result in insufficient elevation cues when the sounds were reproduced by the real sources.

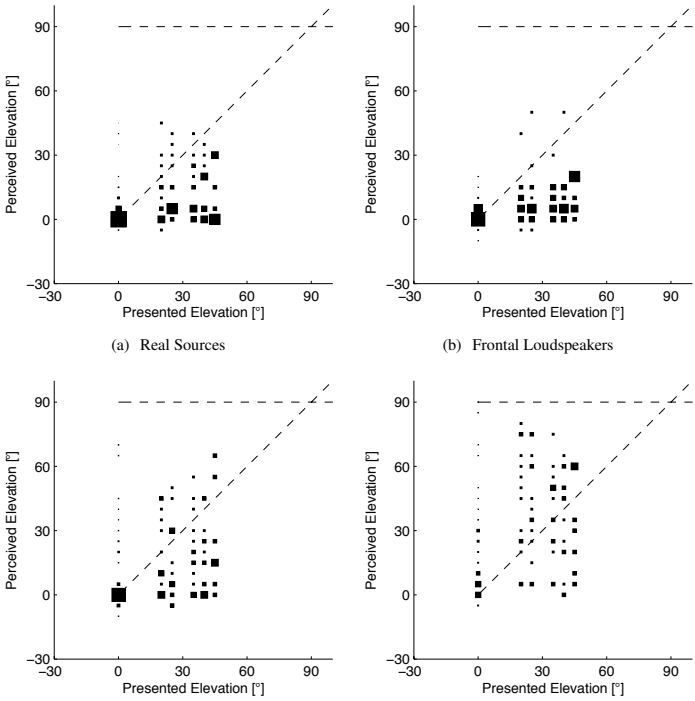
In the case of the virtual sources, the crosstalk cancellation filters used in this experiment are regularized above and around 6 kHz[1]. This implies that the channel separation above that frequency is rather poor and therefore ambiguous elevation cues are introduced. When using the upper loudspeakers, natural elevation cues are introduced in the sounds due to the poor crosstalk at frequencies above 6 kHz. This could explain why a better elevation perception is obtained with this setup as well as the large bias towards the frontal plane observed with the perceived azimuth. It also explains why sources placed on the horizontal plane were more often perceived at elevated positions if compared to the other setups.

The azimuth and elevation localization performances when the subjects were instructed to rotate their heads are presented in figures 8 and 9. We can see that front-back confusions are completely resolved with the real sources when the subject is allowed to move the head (figure 8(a)). This is in agreement with previous studies in which head movement have shown to help resolving front-back confusions[24, 19, 9].

As expected, when the listener rotates the head and virtual sources are reproduced by the frontal and rear loudspeakers, all the rear/frontal images are localized in the vicinity of the loudspeakers that are reproducing them (figures 8(b) and 8(c)). Nevertheless, virtual images below 40° reproduced by the frontal loudspeakers and virtual images above 140° are localized within a reasonable error.

In the case of the virtual images reproduced by the upper loudspeakers, there is a considerable increase in the bias towards the frontal plane (figure 8(d)). In [12] and [14] it was shown that head rotations do not affect considerably the performance of the crosstalk cancellation when the loudspeakers are placed above the listener. This also implies that the dynamic binaural cues expected by head rotations are not present in the sound, i.e. updates of the interaural time delays (ITD) and interaural level differences (ILD) that correlate to the location of the source. We believe that this lack of updates with head rotations do not give the additional cues that the subject's auditory system is expecting. Adding the misleading elevation cues due to the poor cancelation above 6 kHz, this ambiguity is then resolved by locating the virtual image somewhere near the frontal plane.

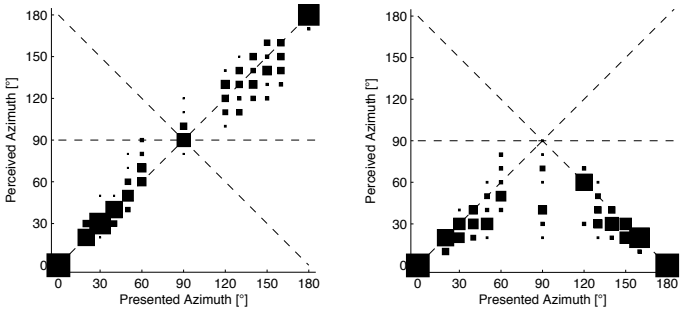
[1]The shape factor used to calculate the filters is a band-stop filter with stop band between 500 Hz and 6 kHz. Since this filter shape does not change abruptly with frequency, some frequencies below 6 kHz are also regularized. Details on the shape factor design can be found in [12].



(a) Real Sources (b) Frontal Loudspeakers

(c) Rear Loudspeakers (d) Upper Loudspeakers

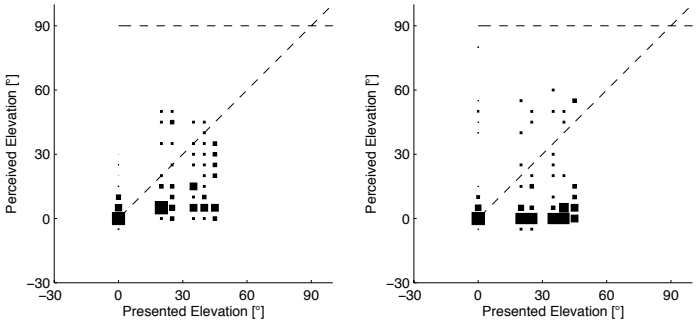
Fig. 7: Presented elevation versus perceived elevation angles with real and virtual sources. The head of the listener is at the optimal position and is kept still while the stimuli were reproduced. The area of the squares is proportional to the amount of subjects that perceived the same angle. The dashed lines correspond to the perfect localization (45° line) and localization above the head (constant 90° line). a) Real sources, b) Virtual sources reproduced by the frontal stereo-dipole, c) Virtual sources reproduced by the rear stereo-dipole, d) Virtual sources reproduced by the stereo-dipole above the head.



(a) Real Sources (b) Frontal Loudspeakers

(c) Rear Loudspeakers (d) Upper Loudspeakers

Fig. 8: Presented azimuth versus perceived azimuth with real and virtual sources. The listener rotate his/her head while the test stimuli were reproduced. The area of the squares is proportional to the amount of subjects that perceived the same angle. The dashed lines correspond to the perfect localization (45° line), front-back confusions ($-45°$ line) and frontal plane (constant 0° line). a) Real sources, b) Virtual sources reproduced by the frontal stereo-dipole, c) Virtual sources reproduced by the rear stereo-dipole, d) Virtual sources reproduced by the stereo-dipole above the head.



(a) Real Sources (b) Frontal Loudspeakers

(c) Rear Loudspeakers (d) Upper Loudspeakers

Fig. 9: Presented elevation versus perceived elevation angles with real and virtual sources. The listener rotate his/her head while the test stimuli were reproduced. The area of the squares is proportional to the amount of subjects that perceived the same angle. The dashed lines correspond to the perfect localization (45° line) and localization above the head (constant 90° line). a) Real sources, b) Virtual sources reproduced by the frontal stereo-dipole, c) Virtual sources reproduced by the rear stereo-dipole, d) Virtual sources reproduced by the stereo-dipole above the head.

A similar bias towards the horizontal plane in the perceived elevation angles is observed when the listeners were instructed to rotate his/her head (figure 9). However, we can see that when the virtual images are reproduced by the upper loudspeaker most of the elevation angles are over estimated (figure 9(d)), which is also a consequence of the poor cancellation for frequencies above 6 kHz.

Table 1 shows the percentage of responses that could be categorized as being within a $\pm 10^\circ$ error from the presented azimuth, front-back confusions or biased towards the frontal plane. Supporting the previous observations we can see that while the percentage of responses within a $\pm 10^\circ$ error increases with head rotations for the real sources, it decreases for the virtual sources. This is especially noticeable with the upper loudspeakers. The amount of front-back confusions increases considerably with head rotation for virtual sources reproduced with the frontal and rear loudspeakers. The percentage of responses biased towards the frontal plane increases dramatically with head rotations when reproduced by the upper loudspeakers.

4.1.6. Summary and general remarks

After each block subjects were asked to assess the difficulty of the block and to write additional comments to their experience. Some of the subjects found the sessions with the upper loudspeakers to be the most difficult ones, especially when they were asked to rotate their heads. They reported that they did not feel confident determining elevation. One of the subjects mentioned that in those particular sessions some of the sources did not have a clear location and that instead she perceived them as a long loudspeaker.

Even though it has been shown that the sweet spot size increases considerably when the loudspeakers are placed above the head, it seems that a large channel separation is not sufficient to render a proper virtual image, especially when head rotations are allowed. We believe that since the ITD changes introduced by head rotations when the loudspeakers are above the head are negligible, this can also result in ambiguities that make the localization task difficult. This could explain the large variability observed in the results obtained with the head rotations and the bias towards the frontal plane. On the other hand, we could observe that elevation cues are better rendered by the upper loudspeakers even though sources on the horizontal plane are also perceived as being elevated.

The virtual sources used in this experiment were binaural recording done with Valdemar. What's more, the crosstalk cancellation filters were calculated using impulse responses measured with Valdemar as well. It is known that for frequencies around and above 6 kHz, individual differences of the head related transfer functions (HRTFs) become significantly large. The fact that non-individual HRTFs were used in this experiment, increases the error of the results and could have potentially degraded the performance of the measured binaural systems.

In spite of these observed limitations, we decided to continue with the implementation of the TSD system and its evaluation. From the results observed in this experiment, we can say that the frontal and rear loudspeakers have an acceptable performance up to $40^\circ - 50^\circ$ with head fixed and head rotations. This lead us to conclude that the optimal crossfading angle should be somewhere between that range. Thus, we decided to set the crossfading angle to 45°.

4.2. Subjective Experiment II

Now that we have estimated the optimal placement of the crossfading region, we need to evaluate the overall performance of the TSD system. For this purpose, a localization experiment with dynamic sources was carried out. In this experiment a comparison between real sources, the three stereo-dipoles working independently and the TSD system were carried out. For the sake of completeness, the influence of head rotations in the perception of the sources was also evaluated.

Before running this experiment it was necessary to do a pilot test in order to define the optimum width of the crossfading region. Binaural recordings of a rotating loudspeaker moving between zones were used and the width of the crossfading region was systematically varied from 10° to 30°. It was found that a width of 20° resulted in the smoothest transition between regions.

The same seven subjects that participated in the first experiment took part in this experiment. Therefore, they were already familiar with the environment and the procedures.

4.2.1. Apparatus

In order to reproduce a dynamic source, a rotating loudspeaker was placed inside the anechoic chamber where the first experiment took place. The loudspeaker is mounted onto a transversal rotator connected to a rotating base. The rotating base is placed outside the inner room of the anechoic chamber and hence the noise coming from the motor is not audible inside. The loudspeaker is located

	Real Source		SD front		SD back		SD above	
	Head Fixed	**Head Rotations**	**Head Fixed**	**Head Rotations**	**Head Fixed**	**Head Rotations**	**Head Fixed**	**Head Rotations**
Localization within $\pm 10^\circ$ error	70%	79%	33%	31%	33%	27%	45%	27%
Front-back confusions	3%	0%	39%	45%	32%	43%	24%	27%
Bias towards frontal plane	10%	4%	11%	1%	9%	1%	19%	33%

Table 1: Percentage of azimuth responses within $\pm 10^\circ$ error of the presented angle, front-back confusions and responses biased towards the frontal plane.

at 0° elevation with respect to the listener's head and can rotate 360° continuously around the listener. The distance between the listener's head and the loudspeaker is 1 m. The angular speed of the rotating base is fixed to $4.5^\circ/s$ and it can only make clock wise rotations. We considered this speed to be sufficiently moderated for the subjects to perceive the movement since it is above the threshold for perceptibility of directional changes [4]. The position of the loudspeaker can be controlled via the parallel port of the PC used in the experiment. Figure 10 shows a close-up to the rotating loudspeaker.

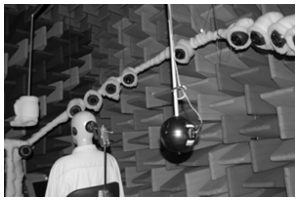


Fig. 10: Rotating loudspeaker.

In this experiment subjects were again blind folded before entering the chamber and sat inside the cabin with the acoustically transparent curtain used in experiment I.

The virtual sources were binaural recordings of the rotating loudspeaker done with Valdemar. The recordings have a duration of 15 s and are moving across the crossfading zones. We defined the evaluated displacements as right-front (RF), right-back (RB), left-back (LB) and left-front (LF). Figure 11 depicts the four different evaluated trajectories.

4.2.2. Stimulus

We used pink noise and an extract of an anechoic recording of a drum as base signals for this experiment. Both

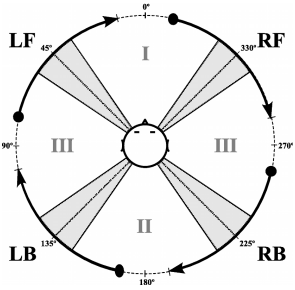


Fig. 11: Diagram of the trajectories of the dynamic sources denoted as right front (RF), right back (RB), left back (LB) and left front (LF). The roman numbers correspond to the different zones and the gray areas to the crossfading region.

signals were band limited between 200 Hz and 8 kHz. The stimulus consisted of either one of the signals reproduced by the rotating loudspeakers or a binaural recording of the rotating loudspeaker reproduced by one of the stereo-dipoles or the TSD system.

4.2.3. Experimental design and Localization method

As in experiment I, this experiment consists on four scenarios: 1) real source - head fixed, 2) virtual source - head fixed, 3) real source - head rotation, and 4) virtual source - head rotation. The movement of the head was controlled in the same fashion as in experiment I.

The experiment was divided into a training session and four main sessions of approximately 20 minutes each. Each session consisted on one of the two stimuli (pink noise or drums) and either the head fixed or head rotations

scenarios. The order of scenarios and stimuli presentation was also randomized using a Greco-Latin square design [5].

Subjects had also access to a touch-screen and the user interface inside the cabin. Before each sound was reproduced, subjects were reminded to place their head correctly. After each stimulus was reproduced they were asked to asses the perceived starting point and end point of the trajectory in a two-dimensional diagram with markers in a circle distributed with a 10° resolution, which represented the locations around them. The starting point was then marked as A in the screen and the ending point was marked as B in the screen. If they perceived that the source moved to one place and came back, they were instructed to mark three points as starting (A), turning (B) and ending point (C) of the trajectory.

Once they had selected the trajectory of the source, they were asked whether they perceived changes in the elevation of the source. They were instructed to roughly adjust a set of slide bars distributed from point A to point B to represent the perceived elevation changes in the trajectory.

After determining the elevation changes, subjects were asked to assess the following affirmations/questions:

a) The trajectory of the source was clearly localizable all the time.

They were instructed to click "no" on this affirmation if they perceived that at some point between A and B the source was not localizable.

b) The trajectory is properly represented by the user interface.

They were instructed to click "no" on this affirmation if the perceived trajectory was not going around them but through them, above them or in a straight line.

c) Where there coloration changes in the sound?

d) Did the source jumped from A to B?

Subjects were instructed to click "yes" if they perceived that source jumped suddenly from A to B.

e) Where there any artifacts or distortion?

Additional to those questions, they were encouraged to comment verbally after each stimulus if they thought the user interface was not sufficient to represent their perception. Before starting the test, subjects were given written and spoken instructions about the localization paradigm.

4.2.4. Results

Even though some subjects mentioned that they found the localization task more difficult with the drums than with the pink noise, we found no significant differences between the responses obtained with these stimuli. Additionally, we did not find significant differences between the responses obtained with each of the reproduced trajectories, i.e. RF, RB, LB and LF. Thus, we decided to combine the results obtained with both stimuli and the four trajectories in order to get a better overview of the responses.

When analyzing the data obtained in this experiment, we found that the responses could be clearly classified into eight categories as follows:

1. **Correct trajectory:** The perceived trajectory is within a 30° error from the original trajectory and neither jumps nor straight movements were reported. We decided to relax the localization error to 30° given that subjects could only select the perceived trajectory once the stimuli had ended and there were no repetitions. Thus, we estimate that the localization error increases due to the fact that the stimuli were rather long and thus the task requires some memory.

2. **Front-back confusions:** The perceived trajectory is within a 30° error from the original trajectory on the opposite hemisphere.

3. **No Movement:** No movement was perceived.

4. **Back and forth:** The beginning of the trajectory is within a 30° error from the original trajectory but at some point it comes back to the starting point.

5. **Front-back confusions with jumps:** The perceived trajectory falls into the front-back confusion definition but sudden jumps were perceived.

6. **Straight move:** The trajectory is not circular but a straight line from A to B. This category includes responses in which the subjects mentioned that the source was going towards them or above their heads.

7. **Blur:** The source was not always localizable.

8. **Color change:** There were coloration changes of the source.

Figure 12 shows the observed percentage distribution of the responses in each of the mentioned categories when the subjects were instructed to keep their heads still. The results obtained with the real sources, virtual sources reproduced with the frontal, rear and upper loudspeakers independently and virtual sources reproduced with the TSD system are compared. The bars indicate the 95% confidence intervals calculated independently for each source type.

The percentage of correctly perceived trajectories is above 60% for the real sources and above 40% for the virtual sources reproduced by the frontal and upper loudspeakers. In contrast, only 28% of the trajectories of the virtual sources reproduced by the TSD system were perceived correctly while the rear loudspeakers show a rather low percentage of trajectories perceived within the defined error (below 10%). There are significant differences between the percentage of correct responses observed with the TSD, the rear loudspeakers and the real sources.

The frontal and upper loudspeakers show the largest amount of front-back confusions (larger than 20%) followed by the TSD system with a 16% of front-back confusion responses which is larger though comparable to the 14% front-back confusions obtained with the real sources.

When the listeners were instructed to keep their head still, static sources were reported only with the virtual sources reproduced by the rear loudspeakers. We can also observe that the rear stereo-dipole shows a larger percentage of trajectories perceived as going back and forth in comparison to the other loudspeaker systems and real sources. That difference is found to be statistically significant at the 0.05 level.

Even though there are very few front-back confusions with jumps, we decided to keep the data in the results. When subjects reported these jumps, they also mentioned that they either perceived the source trajectory starting at their back and suddenly jump to the front or vice-versa. This means that the starting point of the trajectory was correctly perceived, but it suddenly jumped to the opposite hemisphere. This is observed with the real sources and the virtual sources reproduced by the rear and upper loudspeakers.

Around 20% of the responses observed with the rear loudspeakers and the TSD system were perceived as going into a straight line or towards the subjects. Interestingly and despite of being a small percentage, around 8% of the responses observed with the real sources were also perceived as a straight line. This suggests that there are some ambiguities with respect to loudness changes when the source approaches the frontal plane. This change in loudness could have been perceived as a change in distance and that could explain why some subjects perceived even real sources as coming towards them. We do not observe significant differences of these percentages between the real sources, the frontal loudspeakers and upper loudspeakers.

The TSD and the rear loudspeakers show also the largest amount of sources rated as not always localizable. However, that percentage is still below 15% and no significant differences are observed between the different sources types. In this category, some subjects commented that they could clearly localize the starting point of the source but not the ending point and vice versa.

We can not observe any significant differences between the perceived coloration changes between the real sources and the virtual sources. This indicates that subjects might have misunderstood the concept of coloration and thus no conclusions in that regard can be drawn from these results.

Figure 13 shows the percentage distribution of the different responses categories when the subjects were instructed to rotate their heads. The bars indicate the 95% confidence intervals calculated for each source type independently. While the percentage of correctly perceived trajectories increased dramatically for the real sources, it decreased for the virtual sources reproduced by the frontal and upper loudspeaker pairs and the TSD system. The latter shows a degradation of approximately 20% in comparison to the results observed with the head fixed. However, confirming one of our initial hypothesis, the TSD also shows a significant reduction of front-back confusions when head rotations are allowed. As expected, no front-back confusions are observed with the real sources.

Fewer sources are perceived as static when head rotations were allowed than when the head is kept still. Yet, the amount of trajectories perceived as going back and forth increases for all the virtual sources, especially when reproduced by the rear loudspeakers. The percentage of sources perceived as going towards the listeners increased about 10% for the TSD when head rotations are allowed.

We can also see in the results, that when head rotations are allowed the percentage of sources perceived as not always localizable increases. This is especially observe with sources reproduced by the rear loudspeakers.

With regard to the perceived elevation changes, a large variability was observed among and within subjects. Most

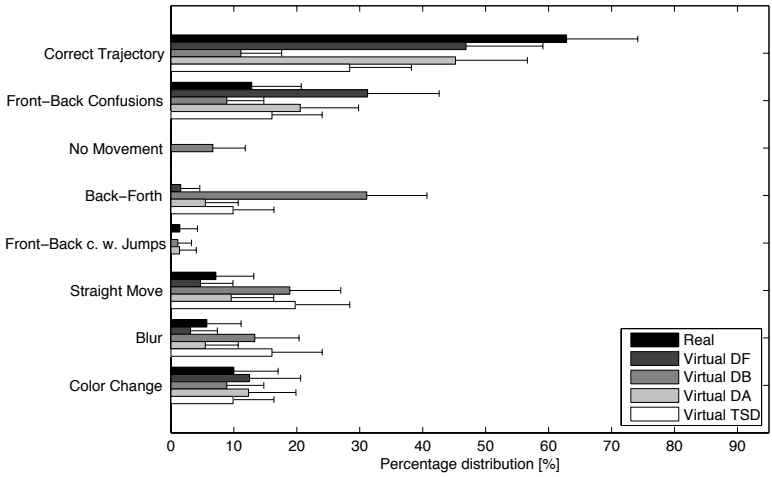


Fig. 12: Percentage distribution of the different answers categories observed in the location experiment with dynamic sources. The listener kept his/her head still while the stimuli were reproduced. The bars correspond to the 95% confidence intervals.

subjects perceived that the sources elevation changed along the trajectory. This was observed not only with the virtual sources but also with the real ones. Some subjects perceived the starting points of the trajectory from an elevated position and slowly going down, whereas other subjects reported sources moving up and down during the trajectory. This infers that the lack of high frequency cues and the natural loudness changes due to a source moving towards the frontal plane could have been perceived as changes in elevation. Yet, we could not observe any meaningful trends between sources trajectories and loudspeaker setups and therefore those results are not presented in this paper.

5. DISCUSSION

The results obtained in the first experiment are in good agreement with studies in which head rotations are shown to improve the localization performance of real sources [24, 19, 9]. Additionally, our initial assumption that head rotations will place the virtual images closer to the loudspeaker that reproduces them was confirmed by the results obtained in the localization experiments with the frontal and rear loudspeakers.

The localization performance with static virtual sources rendered by the upper loudspeakers showed to be poorer than expected, especially when head rotations were allowed. We observed that there is a strong bias towards the frontal plane and when head rotations are allowed, this bias increases significantly. We believe that this observation is a consequence of the negligible changes of interaural time delays (ITD) and interaural level differences (ILD) when the subject rotates his/her head. With head rotations the auditory system is expecting temporal and spectral changes that correlate to the location of the source. Since the changes perceived are negligible when the virtual sources are reproduced by the upper, then the sources will most likely be localized somewhere near the frontal plane and at elevated position. Besides, the use of non-individual HRTFs and the poor channel separation at frequencies above 6 kHz contributed to introduce unavoidable bias in the localization results.

The results of the second experiment confirmed the hypothesis that when using the three-stereo-dipoles (TSD) system and dividing the reproduction into zones, front-back confusions are reduced. We observed that the TSD

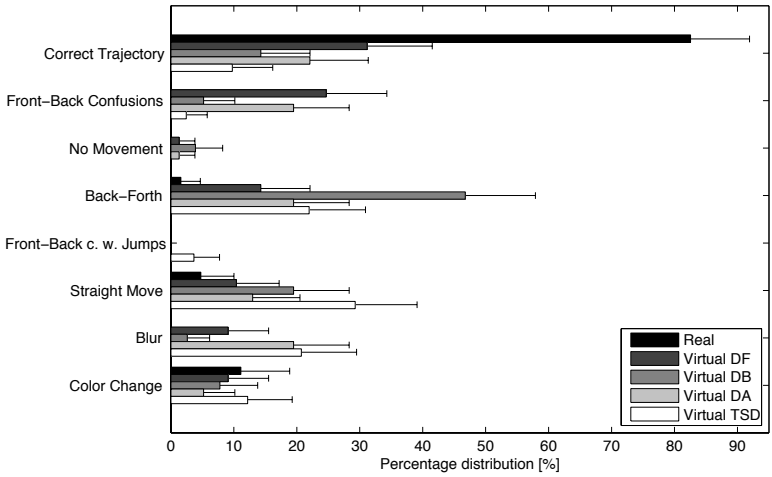


Fig. 13: Percentage distribution of the different answers categories observed in the location experiment with dynamic sources. The listener rotated his/her head from left to right while the stimuli were reproduced. The bars correspond to the 95% confidence intervals.

system yielded fewer front-back confusions if compared to the stereo-dipoles located on the front and above. Contrary to what it was expected, the frontal loudspeaker performed better with respect to correct perception of the trajectory of the virtual images. What's more, the TSD system's overall performance decreased substantially with head rotations, resulting in more blur images, back and forth movements and sources perceived as coming towards the listener.

The observed performance of the TSD system could be a consequence of the poor localization performance observed in the first experiment with the upper loudspeakers when head rotations were allowed. Some subjects mentioned that when listening to the virtual images reproduced by the upper loudspeakers, some of the sources were not clearly localizable. Thus, the limitations observed with that setup in the first experiment are certainly a source of ambiguities and introduce bias when the three stereo-dipoles work together.

We observed an improvement in localization accuracy and elimination of front-back confusions with the dynamic real sources when the subjects were instructed to rotate their heads. In principle, the movement of the source should contain enough dynamic cues to resolve the front-back confusions. Nevertheless, front-back confusions were still observed with the dynamic real sources when the head was kept still. In [24] it was shown that movements of the source have no positive effect on the front-back components of the perceived position. Thus, even when the sound source is moving, head rotations play an important role in the perception of the source position. In the case of the virtual sources, these head movements lead to conflicting ITD and ILD updates. This results in a larger amount of non-localizable sources and front-back confusions when the stereo-dipoles work independently. When the virtual sources are reproduced by the three stereo-dipoles working together, it can be argued that there is a combination of conflicting cues perceived from the frontal and rear loudspeaker plus a lack of ITD and ILDs updates perceived from the upper loudspeakers. This degrades the performance of the TSD system more than expected when head rotations are allowed.

During the pilot and main tests we observed a slight change in coloration between the loudspeaker pairs. These coloration changes could have also influenced the perceived trajectory and elevation of the source. How-

ever, subjects reported almost the same amount of coloration changes with the real and virtual sources. We thus believe that the concept of coloration was not properly understood and some subjects might had described a coloration change when there was actually a loudness change in the case of the real sources for example.

Even though the filters used in the experiment were designed to have the same power, the placement of the loudspeakers results in unavoidable cancellation differences due to the different characteristics of the impulse responses used to calculate the filters. Likewise, the regularization applied to the filters to frequencies above 6 kHz results in audible crosstalk leakages at those frequencies. It has been shown that at high frequencies the crosstalk is more audible than at lower frequencies and the location of the source can be significantly affected [13]. In addition to that, the use of non-individual HRTFs is also a potential source of error in these experiments. These issues not only added extra coloration to the virtual sources but also could have provided conflicting cues, especially at high frequencies where the dips and peaks of the HRTFs are known to vary considerable among subjects and were important elevation cues are embedded.

During the second experiment it was also observed a large variability of the perceived elevation changes. We did not observe any trend with respect to real sources, virtual sources or loudspeaker setup and the perceived elevation changes varied considerable among subjects. This could be a consequence of the lack of cues above 8 kHz for the real sources which makes the elevation discrimination a difficult task. The crosstalk leakages above 6 kHz of the virtual images certainly influenced their perceived elevation.

In summary, front-back confusions in virtual images are considerably reduced with the proposed TSD system. However, contrary to what we expected there was no improvement in overall localization performance when using the three stereo-dipoles working together.

At this stage, we can not really reject our initial hypothesis and this study can be considered as an initial step towards the design of a system robust to head movements. First, we need to take a closer look into the reproduction of virtual sources and localization performance with the loudspeakers above the head. The fact that we are regularizing the crosstalk cancellation filters above 6 kHz is already degrading the overall localization performance. There are also some pending issues such as the use of individual HRTFs and the changes in coloration between loudspeaker pairs, which require further research.

6. REFERENCES

[1] Futoshi Asano, Yoiti Suzuki, and Toshio Sone. Role of Spectral Cues in Median Plane Localization. *Journal of the Acoustic Society of America*, 88(1):159–168, July 1990.

[2] Mingsian R. Bai and Chih-Chung Lee. Objective and Subjective Analysis of Effects of Listening angle on Crosstalk Cancellation in Spatial Sound Reproduction. *Journal of the Acoustic Society of America*, 120(4):1976–1989, October 2006.

[3] Jerry Bauck. A Simple Loudspeaker Array and Associated Crosstalk Canceler for Improved 3D Audio. *Journal Audio Engineering Society*, 49(1 - 2):3 – 13, 2001.

[4] Jens Blauert. *Spatial Hearing*. Hirzel-Verlag, 3rd edition, 2001.

[5] George E. P. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for Experimenters: design, discovery, and innovation*. Wiley-Interscience, 2nd. edition, 2005.

[6] Flemming Christensen, Clemen Boje Jensen, and Henrik Møller. The Design of VALDEMAR - An Artificial Head for Binaural Recordings Purposes. In *109th Convention of The Audio Engineering Society*, page 5253, Los Angeles, CA, September 22-25 2000.

[7] Richard David Clemow. Method of Synthesizing a Three Dimensional Sound-Field. U. S. Patent 6,577,736, June 2003.

[8] John Eng. Sample Size Estimation: How Many Individuals Should Be Studied? *Radiology*, 227(2):309–313, 2003.

[9] P. A. Hill, P. A. Nelson, O. Kirkeby, and H. Hamada. Resolution of Front-Back Confusion in Virtual Acoustic Imaging Systems. *Journal of the Acoustic Society of America*, 108(6):2901–2909, December 2000.

[10] Yuvi Kahana, Philip A. Nelson, Ole Kirkeby, and Hareo Hamada. A Multiple Microphone Recording Technique for the Generation of Virtual Acoustic Images. *Journal of the Acoustic Society of America*, 105(3):1503–1516, March 1998.

[11] Ole Kirkeby and Philip A. Nelson. The "Stereo Dipole" – A Virtual Source Imaging System Using Two Closely Spaced Loudspeakers. *Audio Engineering Society*, 46(5):387–395, May 1998.

[12] Yesenia Lacouture Parodi. Analysis of Design Parameters for Crosstalk Cancellation Filter Applied to Different Loudspeaker Configurations. In *125th Convention of The Audio Engineering Society*, San Francisco, C.A, October 02-05 2008.

[13] Yesenia Lacouture Parodi and Per Rubak. A Subjective Evaluation of the Minimum Audible Channel Separation in Binaural Reproduction Systems through Loudspeakers. In *128th Convention of the Audio Engineering Society*, London, UK, May 22 - 25 2010.

[14] Yesenia Lacouture Parodi and Per Rubak. Objective Evaluation of the Sweet Spot Size in Spatial Sound Reproduction Using Elevated Loudspeakers. *Journal of the Acoustical Society of America*, 128(3), September 2010.

[15] Yesenia Lacouture Parodi and Per Rubak. Sweet Spot Size in Virtual Sound Reproduction: A Temporal Analysis. In *Principles and Applications of Spatial Hearing*. World Scientific, in press.

[16] Tobias Lentz. Dynamic Crosstalk Cancellation for Binaural Synthesis in Virtual Reality Environment. *Audio Engineering Society*, 54(4):283–294, April 2006.

[17] A. W. Mills. On The Minimum Audible Angle. *Journal Acoust. Soc. Am.*, 30(4):237 – 246, 1958.

[18] Henrik Møller. Spatial Sound Reproduction System with Loudspeakers. International Patent WO2008135049 (A1), 13th November 2008.

[19] Stephen Perrett and William Noble. The Effect of Head Rotations on Vertical Plane Sound Localization. *Journal of the Acoustic Society of America*, 102(4):2325–2332, October 1997.

[20] John Rose, Philip Nelson, Boaz Rafaely, and Takashi Takeuchi. Sweet Spot Size of Virtual Acoustic Imaging Systems at Asymmetric Listener Locations. *Journal of the Acoustic Society of America*, 112(5):1992–2002, November 2002.

[21] Takashi Takeuchi and Philip A. Nelson. Robustness to Head Misalignment of Virtual Sound Imaging Systems. *Journal of the Acoustic Society of America*, 109(3):958–971, March 2001.

[22] Takashi Takeuchi and Philip A. Nelson. Optimal Source Distribution for Binaural Synthesis Over Loudspeakers. *Journal of the Acoustic Society of America*, 112(6):2786–2797, December 2002.

[23] Takashi Takeuchi and Philip A. Nelson. Subjective and Objective Evaluation of the Optimal Source Distribution for Virtual Acoustic Imaging. *Journal of the Audio Engineering Society*, 55(11):981–997, November 2007.

[24] Frederic L. Wightman and Doris J. Kistler. Resolution of Front-Back Ambiguity in Spatial Hearing by Listener and Source Movement. *Journal of the Acoustic Society of America*, 105(5):2841–2853, May 1999.

A APPENDIX

A.1 Binaural reproduction and beamforming

This report was written as part of the visit to the National Chiao Tung University in Hsin-chu, Taiwan during autumn 2009. It contains a summary of a short study carried out on the possibility of using loudspeaker arrays and beamforming to improve the sweet spot size of binaural reproduction systems through loudspeakers. Its contents are not included in the dissertation.

INTERNAL REPORT: Binaural Reproduction Through Loudspeakers and Beamforming

Yesenia Lacouture Parodi

Department of Electronic Systems, Section of Acoustics, Aalborg University, Aalborg, DK-9220, Denmark

1. INTRODUCTION

The problem of narrow sweet spot in binaural reproduction systems through loudspeaker have been usually addressed by proposing different span angles or loudspeaker configurations [1, 3, 7, 9, 10, 14]. The sweet-spot is usually defined as the area in which the amount of head movements is within the maximum allowed such that the introduced errors are negligible. However, the overall sweet-spot remains rather narrow and most of the proposed configurations show still limitations with regards to head movements.

In [12] and [13] a study of the sweet spot size of different configurations was presented. In that study only two- and four-channel configurations were considered.

One of the main issues when calculating inverse filters for crosstalk cancellation networks is the redundancies that exists between the direct path and the cross-paths. This results in singularities in the inverse matrix and requires the use of more regularization. This leads to the idea that if that redundancy is reduced then a better crosstalk cancellation might be achieved. This redundancy can be reduced by for example narrowing the directivity pattern of the loudspeakers.

This report explores the possible advantages of using beamforming to improve the robustness to head misalignments of binaural reproduction systems through loudspeakers. It is not intended as a thorough scientific investigation, but as a preliminary step towards the understanding of beamforming techniques and their possible applicabilities in binaural reproduction.

The report contains a set of simulations carried out in Matlab and short descriptions of the beamforming techniques applied in each simulation. The results obtained are thus discussed and possible development of the ideas are considered.

This work was done under the supervision of Professor Mingsian Bai during the author's visit to the National Chiao Tung University in Taiwan, in autumn 2009.

2. SIMULATIONS

In order to get a better understanding of beamforming techniques, different methods were investigated and simulated. The sweet spot was then analyzed employing the different techniques and compared with the sweet spot observed with a single two-channel configuration. Two span angles were simulated: 16° and 60°. The span angle of the equivalent arrays are calculated from their physical center.

2.1. Crosstalk Cancellation Method

The crosstalk cancellation filters were calculated using the fast deconvolution method [11]:

$$\mathbf{C}_o(k) = [\mathbf{H}^H(k)\mathbf{H}(k) + \beta\mathbf{B}^H(k)\mathbf{B}(k)]^{-1}\mathbf{H}^H(k)e^{\frac{-j2\pi(k-1)m}{N_c}} \quad (1)$$

for $k = 1, \ldots, N_c$, where N_c is the filter length and m is the modeling delay in samples. $\mathbf{H}$ is the plant matrix. The matrix $\mathbf{B}(k)$ is defined as the shape factor $B(k)$ multiplied by the identity matrix $\mathbf{I}$. The shape factor is a digital filter used to limit the gain of the filters at selected frequencies. In these simulation a stop-band filter with stop bands at 500 and 6000Hz was used (see [12] for more details). The filter lengths were set to 512 taps.

The sweet spot is simulated using the HRTFs form Valdemar's database [5] and the loudspeakers are assumed to be point sources. The HRTFs are then convolved with the frequency response of the array corresponding to the angle between the array and the ears for each lateral displacement.

Three beamforming techniques were analyzed: delay and sum, super directive and near field beamformers. The following summarizes the simulations carried out and presents the obtained results.

2.2. Delay-and-sum beamforming:

The simplest way to do a beamformer is by using the so-called delay-and-sum (DSB) beamforming. As its name suggest it, it delays the input to each transducer so that the summation of the output results in the desired beampattern. The response of a DSB with L transducers in a uniform linear array is given by,

$$B(\omega,\theta) = \sum_{n=-M}^{M} e^{-jn\omega(\sin\theta_o - \sin\theta)d/c} \qquad (2)$$

where L=2*M+1 is the number of loudspeakers, θ_o is the steering direction and d is the spacing between loudspeakers.

In the simulations done with the DSB, the following settings were used:

- M = 5
- d = 4.2 (Maximum working frequency is thus $f_h = \frac{c}{2d} = 4000Hz$)
- Steering angle (θ_o) = 25°

The steering angle was selected considering the width of the beam and the separation of the ears. Figure 1 shows the beam pattern of the DSB implemented.

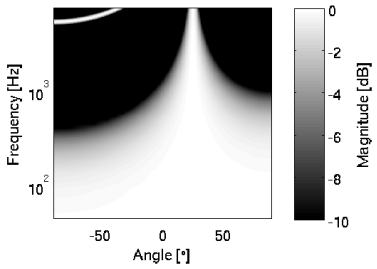


Fig. 1: Beampattern of the delay-and-sum beamformer with steering angle $\theta_o = 25°$.

Figure 2 shows the inverse filters calculated for the 16° span angles. Filters are calculated for a two-channel configuration (solid lines) and loudspeaker arrays with the DSB (dotted lines). There are no siginificant differences between the filters calculated with or without array. Only at high frequencies, there are some differences between the filters due to the side-lobes of the beamformer.

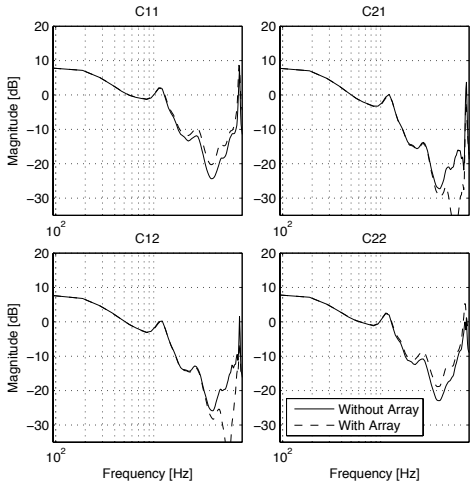


Fig. 2: Crosstalk cancellation filters for a span angle of 16°. Solid line: single loudspeakers, dotted line: loudspeaker array

Figures 3 and 4 show the simulated channel separation at the left ear as a function of frequency and lateral displacement for the 16° and 60° span angle configurations. The right panels show the channel separation obtained with the single loudspeaker pair and the left panels show the equivalent array configuration with the DSB. The upper panels correspond to the channel separation at the right ear and the lower panels to the channel separation at the left ear.

In both configurations, no real improvement of the sweet spot is observed. Besides, there is a slight degradation of the sweet spot at high frequencies and there is an additional ringing frequency present at around 4 kHz. This must be due to the side lobes of the beamformer, which start appearing above 4 kHz (see figure 1).

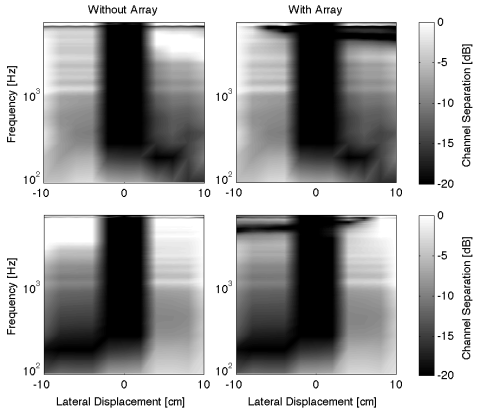


Fig. 3: Channel separation as function of frequency and lateral displacement of a single loudspeaker pair (left) and a loudspeaker array (right). Span angle 16°

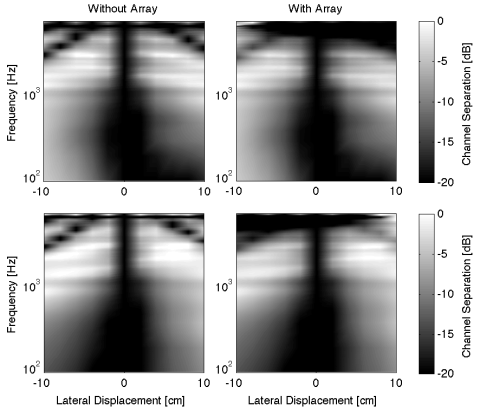


Fig. 4: Channel separation as function of frequency and lateral displacement single loudspeaker pair (left) and a loudspeaker array (right). Span angle 60°

2.3. Super-directive beamformer

The principle of a super-directive beamformer (SDB) is that instead of maximizing the output of the array in target direction, it maximizes the ratio between one sensor and the array output [2, 4]. The larger the array gain, the higher the ability of the array as a spatial filter to suppress the noise. The weights of the loudspeakers are defined as [2]:

$$w_{\beta} = \frac{(\Gamma_{nn} + \beta_{BF}\mathbf{I})^{-1}\mathbf{d}}{\mathbf{d}^{H}(\Gamma_{nn} + \beta_{BF}\mathbf{I})^{-1}\mathbf{d}} \tag{3}$$

where $\mathbf{d}$ is the steering vector , β_{FB} is a regularization factor used to ensure the stability of the array and Γ_{nn} is the coherence matrix and is defined as (assuming spherically isotropic noise) :

$$\Gamma_{nm} = sinc(k(n-m)d) \tag{4}$$

where n and m denote the indexes of the loudspeaker elements, k is the wave number and d is the distance between loudspeakers.

The settings applied to the super-directive beamforming in the simulations are the following:

- M = 5
- d = 4.2 (Maximum working frequency is thus $f_h = \frac{c}{2d} = 4000Hz$)
- Steering angle (θ_o) = 5°
- $\beta_{BF} = 10^{-13}$

Since the super-directive beamformer is expected to have a narrower beamwidth than the DSB, a steering angle of 5° is found to be sufficient to direct the beam towards the ears. Figure 5 shows the directivity pattern of the super directive beamformer simulated.

Figure 6 shows the inverse filters calculated for the two-channel setup with span angle of 16° and the array equivalent with the SDB weights applied to it. As with the DSB, there are no significant differences between the inverse filters without and with beamformer.

The channel separation as a function of frequency and lateral displacements for the two-channel case and the array case with SDB with span angles of 16° and 60° are presented in figures 7 and 9 respectively. The upper

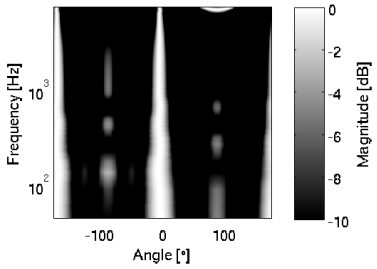


Fig. 5: Directivity pattern of the super directive beamformer with steering angle $\theta_o = 5°$.

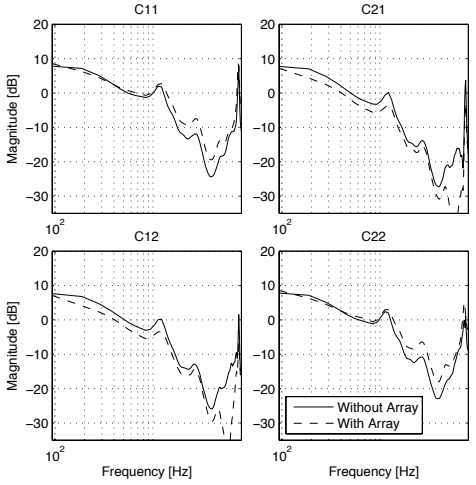


Fig. 6: Crosstalk cancellation filters for a span angle of 16°. Solid line: single loudspeakers, dotted line: loudspeaker array with super directive beamformer

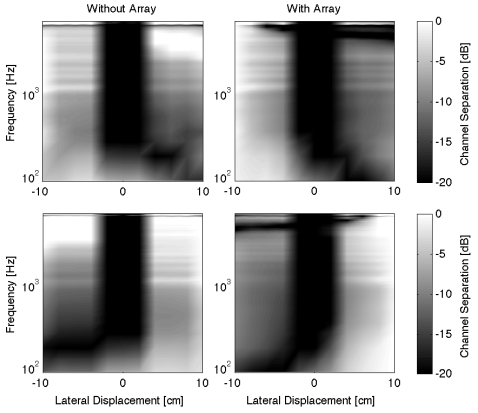


Fig. 7: Channel separation as function of frequency and lateral displacement single loudspeaker pair (left) and a loudspeaker array with super directive beamformer (right). Span angle 16°

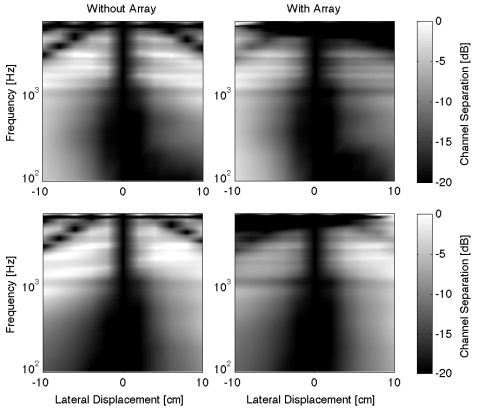


Fig. 8: Channel separation as function of frequency and lateral displacement single loudspeaker pair (left) and a loudspeaker array with super directive beamformer (right). Span angle 60°

panels correspond to the channel separation at the right ear and the lower panels to the channel separation at the left ear.

Similar to the results observed with the DSB, there is no significant improvement of the sweet spot with the SDB.

Looking at the condition number with and without loudspeaker array, it can be seen that the contidion of the matrix actually improves with the use of the array. Figure 9 shows the condition number as a function of frequency without array and with the SDB. This must be due to the fact that the redundancies between the cross-terms and the direct signals are reduced, reducing at the same time the singularities of the matrix.

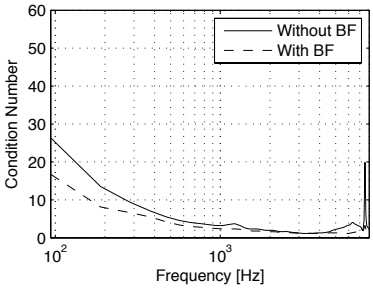


Fig. 9: Condition number of the plant matrix. Solid line: without array, Dashed line: with array and super directive beamformer. Span angle 16°

Thus, the remaining question is why there is no improvement on the sweet spot even though not only the condition of the matrix improves with the array, but also the directivity pattern. Figure 10 shows the beamwidth of the delay and sum beamformer (DSB) and the superdirective beamformer (SDB). As expected, it can be seeing that whereas the DSB is practically omnidirectional at frequencies below 400Hz, the SDB presents a much narrower beamwidth.

To get a more clear view of what it is happening physically, a scale drawing of the hypothesized beam for two arrays spanning 16° at two particular frequencies is presented in figure 11. On the upper figures a beamwidth of 70° is illustrated. That beamwidth would correspond to the beamwidth at 400Hz for the SDB and 1.5Hz for the DSB approximately (see figure 9). It can be observed that there is still a considerable amount of overlapping

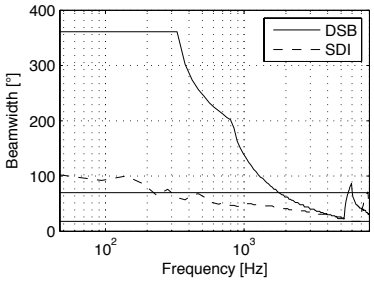


Fig. 10: Beamwidth of the Delay and Sum Beamformer (solid line) and the Super Directive Beamformer (dashed). The two constant lines (solid) correspond to beam widths of 70° and 26° respectively.

of both beams. In order to reduce the overlapping, the beams should be steered 27° at least as shown in the upper-left panel in figure 11.

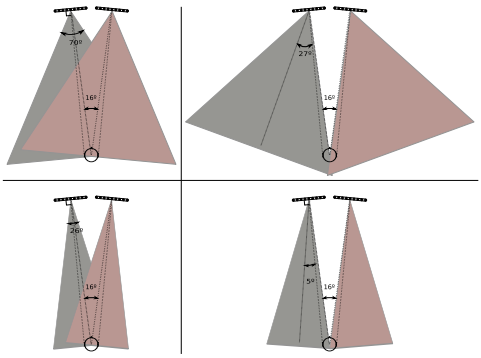


Fig. 11: Sketch of the beampattern for two arrays spanning 16° from each other

The lower panels in figure 11 show a beamwidth of 26°, which would correspond to the beam width at 3 kHz for the SDB and the DSB approximately. It can be observed that there is still a considerable overlapping between beams and in order to reduce the overlapping, the beams should be steered about 5°. Thus, if we steer the beams sufficiently to avoid overlapping at low frequen-

cies, the head will be outside the main lobes at high frequencies. That can explain the degradation observed at high frequencies in figures 3 and 4. On the other hand, if we steer them enough such that the high frequencies are not overlapping and the ears are still in the region of the main lobes, there will be too much overlapping at low frequencies. This could explains the little variations observed in the previous simulations.

2.4. Nearfield beamformer

The beampattern of the delay-and-sum beamformer in the near field is given by (using Green's function for omnidirectional point sources)

$$B(\omega,\theta) = \sum_{n=-M}^{M} W_n(\omega) h_n(\omega,\theta) \tag{5}$$

where W_n are the complex weights of the loudspeakers and

$$h_n(\omega,\theta) = \frac{1}{4\pi r_s(\theta)} e^{-jkr_s(\theta)} \tag{6}$$

$$r_s(\theta) = \sqrt{(r_f \sin(\theta) - nd)^2 + (r_f \cos(\theta))^2} \tag{7}$$

is the distance from each loudspeaker to the evaluation point and r_f is the distance from the center of the array to the evaluation point. d is the distance between loudspeakers. The weights that compensates the delays of the Green's functions are given by [8]:

$$W_n(\omega) = \frac{h_n^*(\omega,\theta_o)}{\sum_{n=-M}^{M} |h_n(\omega,\theta_o)|} \tag{8}$$

where θ_o is the steering angle. Figure 12 shows the beampattern of a near field beamformer for $\theta_o = 0^\circ$ and $r_f = 0.5m$

The settings of the simulated near field beamformer are the follwing:

- M = 5
- d = 4.2 (Maximum working frequency is thus $f_h = \frac{c}{2d} = 4000Hz$)
- Steering angle (θ_o) = 0°

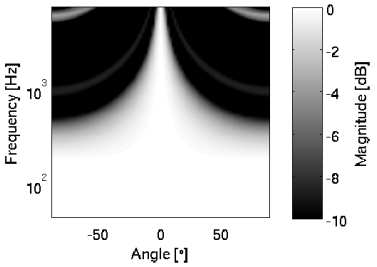


Fig. 12: Beampattern of the Near field beamformer. $\theta_o = 0^\circ$

The characteristics of the beamforming are similar to the ones observed with the DSB in the far field (Fig. 1). Especially at low frequencies the array shows omnidirectional characteristics. Thus, it is expected that similar characteristics to the ones observed with the DSB and the SDB will be obtained with this approach. However, in the near field region, the HRTFs follow different pattern and it is thus necessary to carry a separate analysis of the crosstalk cancellation system and the sweet spot when the loudspeakers are close to the head.

2.4.1. HRTFs in the near field

Brungart et. al concluded in [6] that as the sound sources move closer to the head, there is a substantial increase in the interaural level differences. That is especially observed at low frequencies. Additionally, there is a clear attenuation of the high frequencies i.e. the HRTFs show a low-passed behavior. This can be observed in figure 13 taken from Brungart's results.

From the figure it can be observed that the closer the source gets to the head, the more dramatic are the changes of the HRTFs. This is, small head movements will result in large changes of the HRTFs. Thus, it can be argued that when implementing a binaural reproduction system through loudspeakers in the near field, the sweet spot will be even narrower than the sweet spot of systems in the far field, especially at low frequencies.

3. DISCUSSION

This report explores the use of beamforming as a tool to improve the robustness to head misalignments of binaural reproduction systems through loudspeakers. A set of simulations of three different beamforming techniques

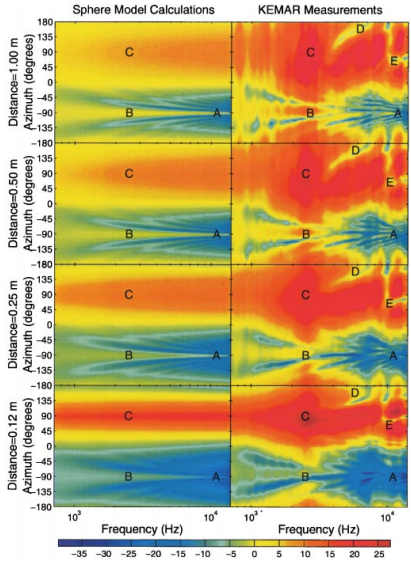


Fig. 13: Contour plots of the HRTFs predicted by the sphere model (left column) and measured with KEMAR (right column). Azimuth is shown at 3° intervals, and frequency is shown at 1/12-oct intervals from 500 Hz to 14 kHz. Figure taken from [6]

are presented and the sweet spot of the simulated arrays are analyzed.

From the preliminary results observed in this report it can be estimated that the use of beamforming with crosstalk cancellation systems might not result in a significant improvement of the sweet spot. However, the use of beamforming might still be useful when the crosstalk cancellation system is placed in a reverberant environment - which is the normal application. Thus, the idea should not be discharged completely, but is the authors opinion that another focus should be adopted. This requires further analysis in the available beamforming techniques and a more thorough analysis of the crosstalk cancellation networks applied to loudspeaker arrays.

4. REFERENCES

[1] Mingsian R. Bai and Chih-Chung Lee. Objective and Subjective Analysis of Effects of Listening angle on Crosstalk Cancellation in Spatial Sound Reproduction. *Journal of the Acoustic Society of America*, 120(4):1976–1989, October 2006.

[2] Mingsian R. Bai and Chengpang Lin. Microphone Array Signal Processing With Application in Three-Dimensional Spatial Hearing. *Journal Acoust. Soc. Am.*, 117(4):2112–2121, April 2005.

[3] Jerry Bauck. A Simple Loudspeaker Array and Associated Crosstalk Canceler for Improved 3D Audio. *Journal Audio Engineering Society*, 49(1 - 2):3 – 13, 2001.

[4] Marinus M. Boone, Wan-Ho Cho, and Jeong-Guon Ih. Design of a Highly Directional Endfire Loudspeaker Array. *Journal of the Audio Engineering Society*, 57(5):309 – 325, May 2009.

[5] Bjarke P. Bovbjerg, Flemming Christensen, Pauli Minnaar, and Xiaoping Chen. Measuring the Head-Related Transfer Functions of an Artificial Head with a High Directional Resolution. In *109th Convention of The Audio Engineering Society*, page 5264, Los Angeles, CA, September 22-25 2000.

[6] Douglas S. Brungart and William M. Rabinowitz. Auditory localization of nearby sources. head-related transfer functions. *Acoustical Society of America*, page 1465 1479, 1999.

[7] Richard David Clemow. Method of Synthesizing a Three Dimensional Sound-Field. U. S. Patent 6,577,736, June 2003.

[8] Markus Guldenschuh and Alois Sontacchi. Transaural Stereo in a Beamformin Approach. In *Proceedings of the* 12^{th} *International Conference on Digital Audio Effects (DAFx-09)*, Italy, September 1-4 2009.

[9] Yuvi Kahana, Philip A. Nelson, Ole Kirkeby, and Hareo Hamada. A Multiple Microphone Recording Technique for the Generation of Virtual Acoustic Images. *Journal of the Acoustic Society of America*, 105(3):1503–1516, March 1998.

[10] Ole Kirkeby and Philip A. Nelson. The "Stereo Dipole" – A Virtual Source Imaging System Using Two Closely Spaced Loudspeakers. *Audio Engineering Society*, 46(5):387–395, May 1998.

[11] Ole Kirkeby, Philip A. Nelson, Hareo Hamada, and Felipe Orduna-Bustamante. Fast Deconvolution of Multi-Channel Systems Using Regularization. *IEEE Transactions on Speech and Audio Processing*, 6(2):189–195, 1998.

[12] Yesenia Lacouture Parodi and Per Rubak. Objective Evaluation of the Sweet Spot Size in Spatial Sound Reproduction Using Elevated Loudspeakers. *Journal of the Acoustical Society of America*, 128(3), September 2010.

[13] Yesenia Lacouture Parodi and Per Rubak. Sweet Spot Size in Virtual Sound Reproduction: A Temporal Analysis. In *Principles and Applications of Spatial Hearing*. World Scientific, in press.

[14] Takashi Takeuchi and Philip A. Nelson. Optimal Source Distribution for Binaural Synthesis Over Loudspeakers. *Journal of the Acoustic Society of America*, 112(6):2786–2797, December 2002.