Technical University of Denmark

Juan David Gil Corrales

Reproduction of realistic background noise for testing telecommunications devices

Master thesis, August 31, 2014

DTU Electrical Engineering
Department of Electrical Engineering

JUAN DAVID GIL CORRALES

Reproduction of realistic background noise for testing telecommunications devices

Master thesis, August 31, 2014

Supervisors:

Wookeun Song, Research Engineer, Brüel & Kjær Sound and Vibration Measurement A/S

Ewen MacDonald, Associate Professor, Department of Electrical Engineering, DTU

DTU - Technical University of Denmark , Kgs. Lyngby

B&K - Brüel & Kjær Sound and Vibration Measurement A/S, Nærum

2014

Reproduction of realistic background noise for testing telecommunications devices
Reproduktion af realistisk baggrundsstøj til test af telekommunikationsudstyr

This report was prepared by:
Juan David Gil Corrales
s121312@student.dtu.dk

Supervisors:
Wookeun Song, Research Engineer, Brüel & Kjær Sound and Vibration Measurement A/S
Ewen MacDonald, Associate Professor, Department of Electrical Engineering, DTU

Technical University of Denmark
DTU Electrical Engineering
Elektrovej, Building 352
2800 Kgs. Lyngby
Denmark
Tel: +45 4525 3954

Brüel & Kjær Sound and Vibration Measurement A/S
Innovation Department
Skodsborgvej 307
2850 Nærum
Denmark
Tel: +45 4580 0500

Project period: March 03 2014 - August 31 2014

ECTS: 35

Education: MSc

Field: Engineering Acoustics

Remarks: This report is submitted as partial fulfilment of the requirements for graduation in the above education at the Technical University of Denmark in collaboration with Brüel & Kjær Sound and Vibration Measurement A/S.

Abstract

A deconvolution method for reproduction of realistic background noise is described in this project. The method calculates a set of inverse filters, that compensate for the response of the loudspeakers and the room in which the reproduction is done. A designed array of 8 microphones, and a set of 8 loudspeakers were used for the optimization of the sound field. Different background noises were simulated in the laboratory to create a reference, and the reproduction of sound was optimized under different conditions.

Five different values for the regularization parameter, three configurations of loudspeakers, four background noise programs, and two techniques to improve the performance of the reproduction, were used to investigate the deconvolution method applied to reproduction of sound. The quality of the reproduction was analysed objectively by means of the magnitude spectrum difference, coherence, and phase difference between the reproduced- and reference sounds. This analysis was complemented with a listening test to evaluate the similarity between the two sounds.

It was found that the highest regularization parameter provided the best results, there was no difference in the performance of setups using 4 or 8 loudspeakers, and for different background noises the error was the same, no matter the program material. The first technique to improve the performance of the reproduction at the target positions was not optimal due to the impact at other positions; and the second technique improved the perception of audible artefacts, reduced the errors in the reproduction, but only in a small amount.

Table of Contents

List of Terms

System Audio workstation (a.k.a. a computer with the latest version of `MATLAB`) connected to the sound card and to the loudspeakers, the room, the Head And Torso Simulator (HATS) on which the designed microphone array is mounted, possibly a telecommunications device, and the microphones themselves. This system from the audio workstation to each of the microphones is characterized and inverted.

Inverse filter Filter that has the inverse response of the system, and that is calculated using regularization.

Performance of the processing (-method, or -reproduction) Matrix convolution of the transfer function of the system with the transfer function of the inverse filter. Ideally the performance is an identity matrix.

Designed microphone array Microphone array designed to cover an area around the head, where telecommunications devices are usually placed.

Target microphone positions Actual microphone positions of the designed microphone array for which the reproduction of sound is optimized.

Validation microphone positions Microphone positions defined in between some of the target positions to inspect the reproduction of sound at places outside of the designed microphone array.

Background noise program Program material that is selected to test the device. In this project four background noise programs were used: Pink noise, Pop music, Classical music, and Speech.

Reference sound sources Sources that are placed in the room to playback the background noise programs. They act as the actual noise sources that are wanted to be reproduced.

Reference sound Sound recorded at the microphone positions when the reference sound sources playback the background noise programs.

Loudspeaker signals Signals calculated by filtering the recorded reference sound with the inverse filter.

Loudspeaker array Actual set of loudspeakers used to reproduce the sound that will have the same properties as the reference.

Reproduced sound Sound recorded at the microphone positions when the loudspeaker signals are played back by the loudspeaker array.

Introduction

Background noise is part of everyday listening experiences. Manufacturers of telecommunications devices, such as mobile phones, headsets, hearing aids, etc., have the need to develop better technologies in the devices to handle the background noise. Algorithms of noise suppression, echo cancellation, or speech enhancement, need accurate acoustic conditions for testing. This means the ability to, in the lab, reproduce the noise source as experienced on location. However, the test facilities and equipment contribute to the recorded sound when played back, making the sound in the lab differs from what it would be experienced at the real source of the noise. Therefore, in order to recreate the "true" noise, to be used for testing the efficiency of processing algorithms in telecommunications devices, a method is needed to remove the effect of the room and equipment, and reproduce the properties of sound as realistically as possible.

The standard method of reproduction of sound for testing telecommunications devices, ETSI EG 202 396-1 [1], is based on the reproduction of binaurally recorded signals using four loudspeakers in a square shape. The reproduction procedure consists of feeding the signal recorded with the left ear of a Head and Torso Simulator (HATS), to the front-left and rear-left loudspeakers, and a similar procedure with the right channels. All the loudspeakers should be equalized to the ears of the HATS, with Infinite or Finite Impulse Response (IIR/FIR) filters, in such a way that the reproduction is as transparent as possible. Several procedures to equalize the loudspeakers involve pairing two loudspeakers on each side, or front and back, in order to obtain a flat response. The equalization procedure does not take into account the crosstalk cancellation, meaning that the left ear will be equalized for the left loudspeakers only, and there will be some error due to the leaking from the loudspeakers on the right-hand side; the same will happen with the right ear.

Another technique widely used to reconstruct a desired sound field in laboratories; makes use of a spherical microphone array to measure the spherical harmonics of the sound field of interest. This technique, known as Ambisonics, uses a loudspeaker array, for which the loudspeaker signals are calculated based on the spherical harmonics measured. It has the advantage that the three-dimensional sound field can be reproduced; however, the quality of reproduction depends on the resolution of the microphone and loudspeaker arrays, which usually require a lot of channels to get good performance in a usable frequency range. Different improvements in the microphone array [2] have been proposed to optimize the performance of Ambisonics in the horizontal plane, by having the same amount of transducers more densely located in the equator of the sphere.

The method studied in this project is based on the active control of sound as proposed by Kirkeby et al. [3]. It is known as the deconvolution or fast deconvolution method, inverse filtering of sound, crosstalk cancellation, regularization method, and several other names referring to the same idea. The fundamentals of the method is the calculation and optimization of inverse filters, in the least-squares sense, that will cancel the effect of the room and the loudspeakers. The main difference between this method and the previous

ones, is that the optimization of the reproduction of sound is made only at particular positions defined by a microphone array. Ideally, the error around those positions is zero across the whole audible frequency range. By using large number of microphones, reproduction in a defined area can be optimized up to certain frequency. One advantage is that the equalization of the loudspeakers, as well as the crosstalk cancellation, are taken into account during the calculation of the inversion filters, which makes this a method of fast implementation. Unfortunately, the optimal solution at different points, than those in the array, might not be perfect in the frequency range of interest. The implementation, advantages and drawbacks of this method will be studied in this project.

Recent studies, have proposed the deconvolution method for different applications. Minnaar et al. [4] implemented a system composed of 29 loudspeakers, and a microphone array with 32 capsules, to study signal processing algorithms in hearing aids. Due to the large amount of loudspeakers and microphones employed, the reproduction of the magnitude and the phase of the sound was guaranteed in an area the size of a human head up to 3kHz, but for higher frequencies only the magnitude was correct. Another application of this method is the equalization of sound systems in cars. Due to the complexity of materials, the disposition of the loudspeakers, and the small space inside a car, the sound reproduction of good quality becomes a difficult task, that can be solved by using digital filters and deconvolution, as studied in [5].

In the last months, a new standard for the reproduction of sound using deconvolution, for testing telecommunications devices using HATS has been under development [6]. The proposal is rather challenging, because it makes use of a small setup with 8 microphones and 8 loudspeakers. This project is the implementation of that proposal, with some suggestions and different approaches to the evaluation of the quality in the reproduction.

The method of reproduction of sound using deconvolution consists of a microphone array for recording the desired sound, and a loudspeaker array for reproducing the optimized sound. By measuring transfer functions from the loudspeakers to the microphones and calculating their inverses, ideal reproduction of sound can be achieved in a wide frequency range. However, the calculation of the inverse matrix is not a trivial task. Due to the (possible) proximity of the microphones in space, the matrix of transfer functions at low frequencies can be ill-conditioned, meaning that a direct inversion is not possible due to singularities [7]. The workaround is to use regularization. By doing this, the effort that the inverse filter makes to correct the transfer function measured is limited, and even though this makes the inversion possible to realize, it makes the inverse filter non-causal which introduces audible artefacts [8] that are not easy to measure or to compensate for.

The subjective performance of the reproduction of sound using the deconvolution method has been addressed with different purposes. Minnaar et al. [4] asked through listening tests, how good the algorithms implemented in a hearing aid were. Norcross. et al. [8], [9], [10], have carried out several investigations on the method itself and the effects of regularization on the audible quality of the reproduced sound. In those works the quality of the reproduction was subjectively evaluated using the MUSHRA test (MUlti Stimulus test with Hidden Reference and Anchor) [11]. This test was implemented and used in this project.

As stated above, in this project, the reproduction of sound using the deconvolution method is performed for the application of testing telecommunications devices. Even though measurement with actual terminals were not done, a methodology to evaluate the performance of the method in positions where microphones of telecommunications devices could be mounted, is presented with a set of analyses of the parameters involved in the method. By means of objective and subjective evaluation of the error of the reproduction, results on the specific case of implementations are provided as a guide for future implementations. Analysis on the regularization, the number of loudspeakers used, the signals used for reproduction, and two proposed techniques to improve the perceived quality, are presented.

CHAPTER 1

Theory

Reproduction of sound using the deconvolution method is based on measurements of the transfer functions from each loudspeaker of a reproduction system, to each microphone of a recording system. A cost function is minimized in the least-squares sense, and the matrix of transfer functions is inverted using regularization. The formulation of the problem is based only on electric signals, which means that no assumptions about the physical system are needed, except that it is linear [12].

The problem can be illustrated in the next block diagram, after Kirkeby et al. [13]:

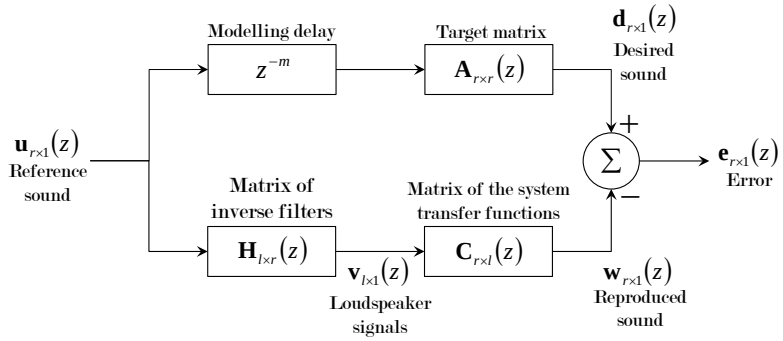


Figure 1.1: Block diagram form of the multichannel sound reproduction problem using the deconvolution method.

The reproduced sound $\mathbf{w}$, recorded at r microphone positions, is the sound produced by l loudspeakers playing back the loudspeaker signals $\mathbf{v}$, which are filtered by the electro-acoustic path $\mathbf{C}$ from each loudspeaker to each microphone. This can be written as a linear system in equation 1.1.

$$\begin{pmatrix} w_{1,1}(z) \\ w_{2,1}(z) \\ \vdots \\ w_{r,1}(z) \end{pmatrix} = \begin{pmatrix} C_{1,1}(z) & C_{1,2}(z) & \cdots & C_{1,l}(z) \\ C_{2,1}(z) & C_{2,2}(z) & \cdots & C_{2,l}(z) \\ \vdots & \vdots & \ddots & \vdots \\ C_{r,1}(z) & C_{r,2}(z) & \cdots & C_{r,l}(z) \end{pmatrix} \begin{pmatrix} v_{1,1}(z) \\ v_{2,1}(z) \\ \vdots \\ v_{l,1}(z) \end{pmatrix} \tag{1.1}$$

The role of the target matrix $\mathbf{A}$, in the block diagram, is to define the desired sound $\mathbf{d}$ in terms of the reference sound $\mathbf{u}$. In the case of this project, the target matrix is an identity matrix, which means that the same recording system is used to record the reference sound and to evaluate the error of the reproduction in the lab. This setting of the target matrix is also known in literature as the crosstalk cancellation problem [14]. The modelling delay z^{-m} is a delay purposely introduced in the processing, to ensure that is possible to implement the inverse filters with the condition that they are causal, and the value of m is half the length of the inverse filter.

The goal of the deconvolution method is to find the matrix of inverse filters $\mathbf{H}$; which, given a set of recordings of a reference sound, allows one to solve for the loudspeaker signals:

$$\begin{pmatrix} v_{1,1}(z) \\ v_{2,1}(z) \\ \vdots \\ v_{l,1}(z) \end{pmatrix} = \begin{pmatrix} H_{1,1}(z) & H_{1,2}(z) & \cdots & H_{1,r}(z) \\ H_{2,1}(z) & H_{2,2}(z) & \cdots & H_{2,r}(z) \\ \vdots & \vdots & \ddots & \vdots \\ H_{l,1}(z) & H_{m,2}(z) & \cdots & H_{l,r}(z) \end{pmatrix} \begin{pmatrix} u_{1,1}(z) \\ u_{2,1}(z) \\ \vdots \\ u_{r,1}(z) \end{pmatrix} \tag{1.2}$$

By setting the condition that the impulse responses of the inverse filters must be stable, $|z| = |e^{j\omega\Delta}| = 1$, where ω is the angular frequency, and Δ is the sampling interval; a cost function J can be defined as the sum of two terms, the error and the effort [3]; each of which represent how well the reference sound is reproduced; and, the total energy of the loudspeaker signals, respectively.

$$J = \mathbf{e}(e^{j\omega\Delta})^H \mathbf{e}(e^{j\omega\Delta}) + \beta \mathbf{v}(e^{j\omega\Delta})^H \mathbf{v}(e^{j\omega\Delta}) \tag{1.3}$$

where $\mathbf{e} = \mathbf{d} - \mathbf{w}$, is the performance error; H is the Hermitian, or conjugate transpose; and β is called regularization parameter.

By changing the regularization parameter from zero to infinity, the solution will change from minimizing only the error, to minimizing only the effort. This means that, for low regularization, the reproduction will be more precise, but the energy of the loudspeaker signals will be high; whereas for higher values, the input energy to the loudspeakers will decrease, in order to protect them from saturation, but there will be more error in the reproduction. Another effect of the regularization is to make the filter non-causal, which generates audible artefacts in the result [15], [16]. So, one must be careful setting this value in order to reduce the audibility of those artefacts.

For regularization parameters higher than zero, the cost function is minimized in the least-squares sense by a vector

$$\mathbf{v}(e^{j\omega\Delta}) = \frac{\mathbf{C}(e^{j\omega\Delta})^H \mathbf{A}(e^{j\omega\Delta}) \mathbf{u}(e^{j\omega\Delta})}{\mathbf{C}^H(e^{j\omega\Delta})\mathbf{C}(e^{j\omega\Delta}) + \beta\mathbf{I}} \tag{1.4}$$

and, by comparing equation 1.4 to equation 1.2, it can be seen that the wanted matrix of inverse filters, is described by equation 1.5:

$$\mathbf{H}(e^{j\omega\Delta}) = \frac{\mathbf{C}(e^{j\omega\Delta})^H \mathbf{A}(e^{j\omega\Delta})}{\mathbf{C}^H(e^{j\omega\Delta})\mathbf{C}(e^{j\omega\Delta}) + \beta\mathbf{I}} \tag{1.5}$$

CHAPTER 2

Methodology

Reproduction of sound using the deconvolution method makes use of a *loudspeaker array* and a HATS, on which a *designed microphone array* is mounted. The procedure consists of using the recording system, composed by the HATS and the designed microphone array, on location (e.g. a train station, a pub, a lecture room, etc.). The sound is recorded and brought to the lab, and by following the next steps, the reproduction of sound at the microphone positions of the microphone array will be ideally identical to the sound on the original location[1].

1. System identification: Measurement of the system transfer function from every loudspeaker position to every *target microphone position*. The *system* is the whole chain from the audio workstation, the loudspeakers, the room, and the HATS (or other elements around like mobile phones, headsets, etc), to the microphone array.

2. Calculation of inverse filters: Filters to compensate for the response of the previously identified system are calculated by inverting the matrix of transfer functions of the system. The direct solution might introduce problems, because at low (and possibly at very high) frequencies the matrix becomes ill-conditioned, resulting in high gains that can saturate the reproduction system. To avoid this, the solution using regularization is implemented instead. An optional correction of the pre-ringing created by the regularization can be applied in this step.

3. Recording of the reference sound: In order to adjust and control the performance of the inversion, the original *reference sound sources*, from which different *background noise programs* are played back, are simulated in the lab with additional loudspeakers. This makes it possible to know which are the original sound sources that produced the *reference sound*, and to repeat measurements that can be compared.

4. Filtering of the recorded reference sound: This is the process of obtaining the *loudspeakers signals*, which will interfere in space in such a way that, at the target microphone positions, the *reproduced sound* is ideally the same as the reference sound.

5. Recording of the reproduced sound: Actual reproduction and recording of the optimized sound that recreates the reference sound as realistically as possible. Optionally, the designed microphone array can be changed for the telecommunications devices to be tested (this was not done in this project).

6. Calculation of the error: The performance can be objectively evaluated by means of magnitude spectrum difference (called simply 'error' in the coming pages), coherence, and phase difference, between the reproduced- and reference sounds.

[1] Please refer to the List of Terms; for definitions of the terms written in italic font.

7. Correction of the response: Based on the response at *validation microphone positions*, the overall response of the inverse filters can be adjusted to compensate for the amplitude errors at positions different to the target microphone positions.

8. Subjective evaluation: Listening test designed to evaluate the processing involved in the method. By asking how similar the reproduced sound is to the reference, information about the audible artefacts can be related to the objective measures of the reproduction error.

2.0 Description of the room, the electro-acoustic system and the calibration procedure

2.0.1 Room

The room in which the method was implemented and tested is located at the headquarters of Brüel & Kjær Sound and Vibration Measurement A/S. It is a small room used for various purposes that was adapted for the installation of the setup. A sketch with the dimensions of the room is shown in figure 2.1.

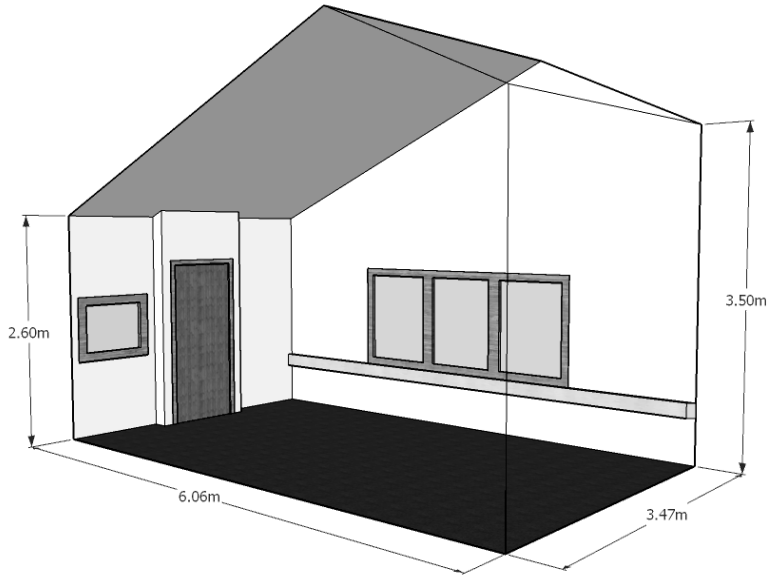


Figure 2.1: Sketch of the test room.

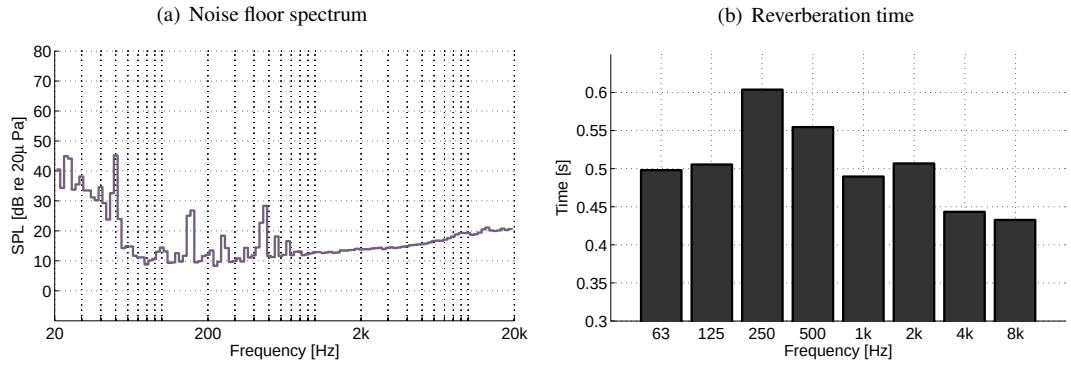


Figure 2.2: Noise floor spectrum and reverberation time of the room.

Background noise is a term commonly used in acoustics for the noise measured when no acoustic events of interest are taking place. In this project, it will be called *noise floor* in order to avoid confusion with the actual background noise program that is meant to be reproduced when testing telecommunications devices.

The spectrum of the noise floor in 1/12-octave bands, measured in the middle of the room, is presented in figure 2.2; together with the estimated reverberation time [17] in 1/1-octave bands, measured in the same position with a loudspeaker placed far from any hard surface in the room. The measurement was not rigorous according to standards, but it was made to give an idea of how the reverberant the room is.

2.0.2 Electro-acoustic system

In figure 2.3, a photograph of the room shows the reference sound sources, the loudspeaker array, and the HATS wearing the designed microphone array. The reference sound sources, B&K type 4224, labelled with numbers 9 and 10 are sources used to reproduce the reference background noise programs. By having these sources, the reference background noise source was always the same, which enables comparison across experiments. The loudspeaker array, labelled with numbers 1 to 8, is an array of 8 loudspeakers placed in a squared shape around the HATS. The distance from the center of the array to the nearest loudspeaker is approx. 1.41m, and 2.00m to the farthest. The loudspeakers used in this array are TANNOY VXP8; which are concentric active loudspeakers in which the low-mid frequency driver is in the same axis as the high frequency driver, presenting a more symmetric radiation with respect to the center of the loudspeaker.

Figure 2.3: Photograph of the loudspeaker array (1 to 8), reference sound sources (9 and 10), and HATS wearing the designed microphone array.

The microphone array that was designed for this project is shown in figures 2.4 and 2.5. After extensive investigations [6], the positions of the target microphones (labelled with numbers 1 to 8) were selected to cover more densely the region around the head, where telecommunications devices are usually located [2]. The microphones used in the microphone array are the B&K type 4958. Four additional microphones B&K type 4935, labelled V1 to V4, were added to the microphone array in between of some of the target microphones in order to validate the reproduction.

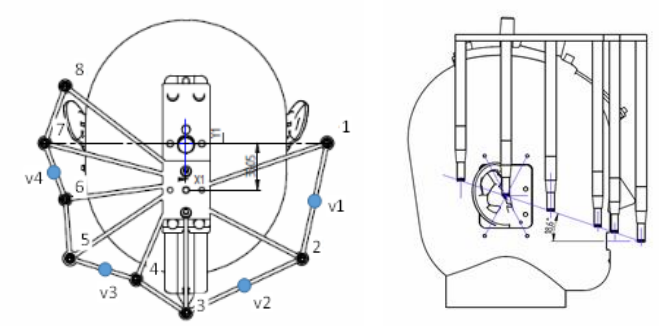


Figure 2.4: Sketch of the designed microphone array.

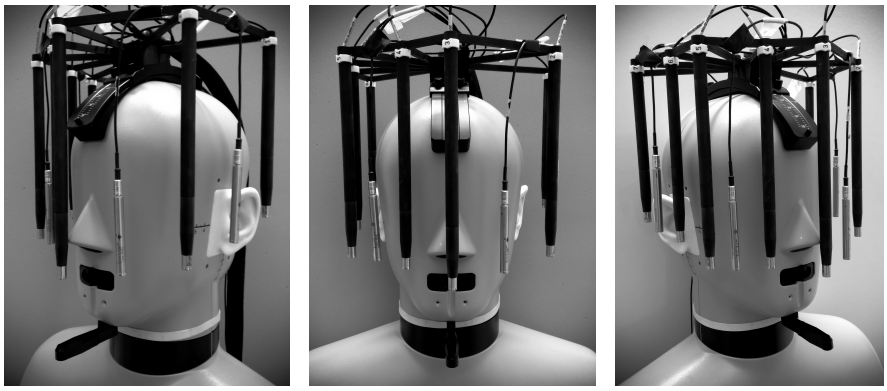


Figure 2.5: Designed microphone array and validation microphones.

2.0.3 Calibration

The system was calibrated by adjusting the input channels sensitivity and the output gain to the loudspeakers. For the input channel sensitivity, a calibrator B&K type 4231 was used to generate 94dB SPL @ 1kHz at the microphones; which were connected to the conditioning amplifier Nexus B&K type 2693. A calibration coefficient was calculated to read the correct value of sound pressure level in MATLAB. The output gain of the loudspeakers was adjusted using an omnidirectional microphone B&K type 2670 placed in the middle of the loudspeaker array while pink noise was played back through each of the loudspeakers. The gain of the amplifiers of each loudspeaker was adjusted to a nominal value, and the corresponding output gain of the sound card, RME FIREFACE UFX, was increased/reduced until obtaining 90dB SPL at the omnidirectional microphone. All recordings and signal processing were carried out in MATLAB at a sampling rate of 48kHz; and a bit depth of 24bits.

[2] In the original sketches, the microphone number 3 was specified to be at the level of the Mouth Reference Point (MRP), however an error in the construction of the array made all the positions to be raised 5mm up relative to this point. This difference is not considered important for the analysis presented in this project, but is a noteworthy information.

The calibration levels of the external sound sources were adjusted to 80dB SPL instead of 90dB SPL. In this way, the reproduction system had more dynamic range than the reference sound sources, so saturation of the reproduction system was avoided in the calculation of the loudspeaker signals.

2.1 System identification

Measurements of transfer functions using sweeps show advantages against distortion and time variance of the system; that are not obtained with methods like MLS-measurements [18]. A system can be characterized by its complex transfer function as defined by equation 2.1:

$$H(f) = \frac{O(f)}{I(f)} \tag{2.1}$$

where $O(f)$ and $I(f)$ are the spectra of the output and the input, respectively.

For the measurement of the transfer function of the present system, this method was implemented by playing back one sweep of approximately 3 seconds and truncating the measured impulse response to 1.5 seconds.

For this implementation, the set of functions used to control the playback and recording of the sweep had an intrinsic delay between input and output, so the characterization of the system was made from the excitation signal in MATLAB ($I(f)$); to every input channel from the designed microphone array ($O(f)$). In this way, the inverse filter designed would compensate for this delay when the recorded reference sound was processed and played back.

The measured transfer functions of the system are presented in figure 2.6. The energy of the impulse response is plotted in dB so the low values are more visible. In these matrices of impulse- and frequency responses, the rows represent the target microphones and the columns are the loudspeakers. for each frequency response, two curves are plotted; the opaque line is the fine resolution frequency response measured for all the available frequency bins, the thick line is the average energy per 1/12-octave band, shown just for graphical purposes. The diagonal elements are highlighted to emphasize the non-crosstalk channels.

2.2 Calculation of inverse filters

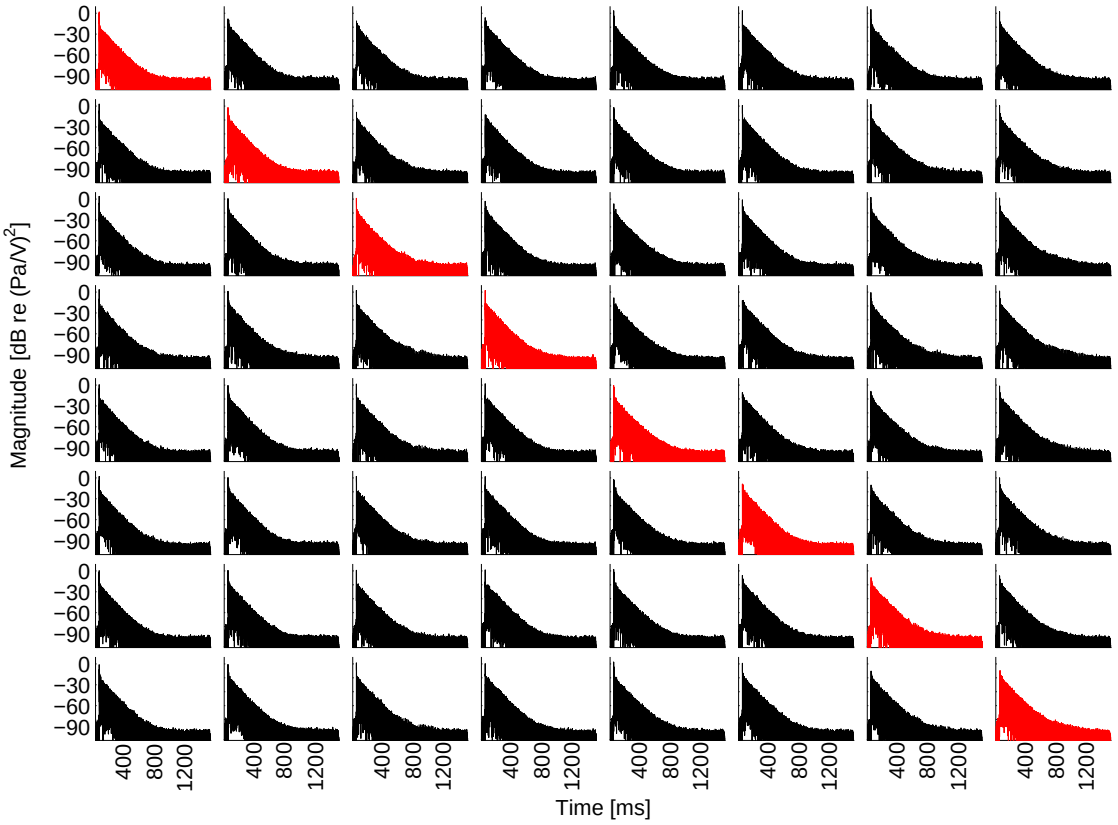
Inverse filters were calculated using equation 1.5. The filters were derived for five different values of the regularization parameter[3], $\beta = -20$dB, 0dB, 10dB, 20dB, and 40dB. In figures 2.7 and 2.8, examples of the transfer function measurements and inverse filter of a 2-by-2 system are shown for two different regularization parameters, $\beta = 20$dB and $\beta = -20$dB, respectively.

2.2.1 Post-processing

The effect of regularization on the inversion of the system can be seen by comparing the transfer functions of the inverse filters shown in figures 2.7 and 2.8. When low regularization is applied ($\beta = 20$dB), the frequency response of the inverse filter (figure 2.7(e)) looks more like a mirrored version of the frequency response of the system (figure 2.7(d)), than when high regularization is applied ($\beta = -20$dB; figures 2.8(e) and 2.8(d)). Comparing these plots, a first conclusion about the performance of the inversion is that $\beta = 20$dB produces a more perfect inversion than $\beta = -20$dB [13].

[3]The values of regularization in linear scale correspond to: $\beta = 10$, 1, 0.3162, 0.1 and 0.01, respectively. So that is why $\beta = -20$dB is referred to as high regularization, and $\beta = 40$dB as low regularization.

(a) System impulse response.



(b) System frequency response.

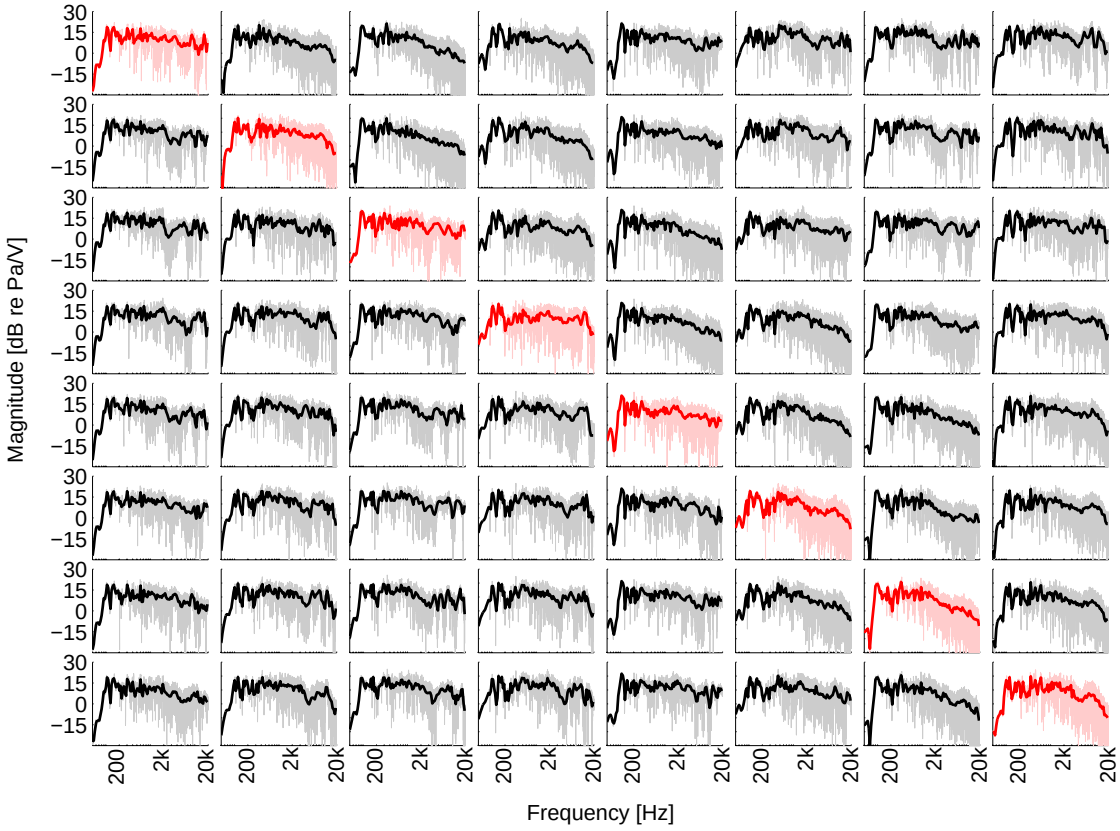


Figure 2.6: 8-by-8 system transfer functions measurements. Rows represent the target microphones and columns are the loudspeakers.

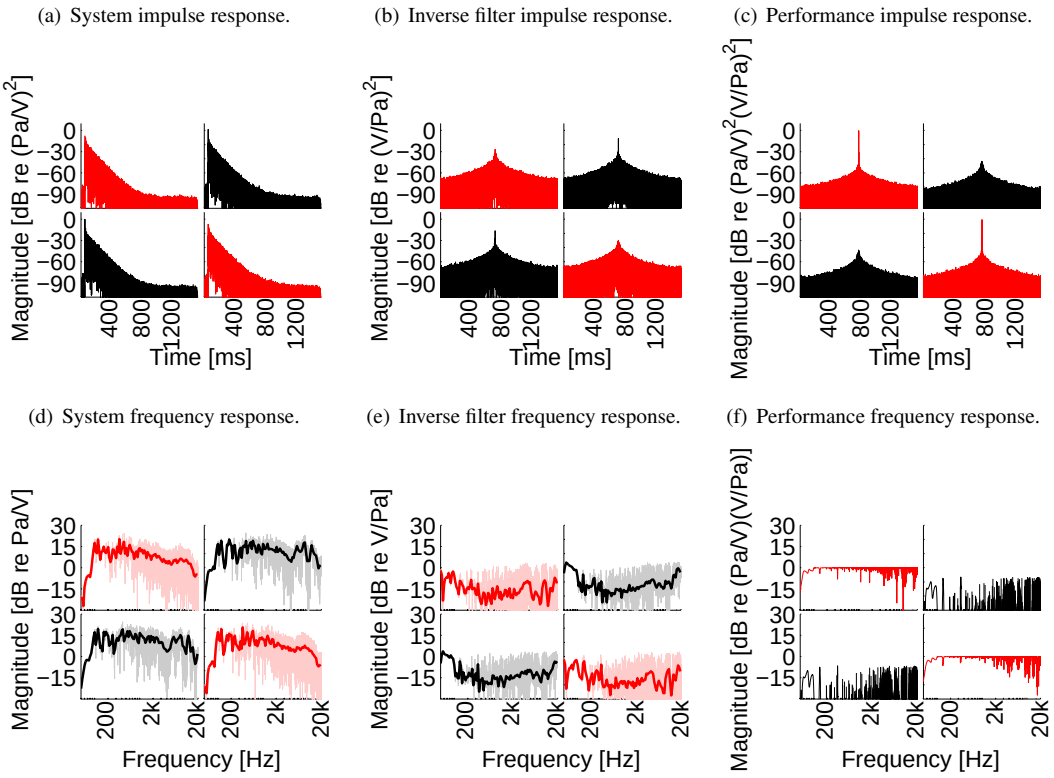


Figure 2.7: 2-by-2 system transfer functions measurements and performance, using regularization of $\beta = 20$dB.

The performance of the inversion is defined as the convolution of the transfer function of the system, $\mathbf{C}$, with the transfer function of the inverse filter, $\mathbf{H}$. In the previous example (when $\beta = 20$dB is used), it can be seen how the diagonal elements of the performance tend more to 0dB, while the off-diagonal elements or crosstalk channels tend to negative values. See figure 2.7(f).

The impulse responses of the inverse filters also show differences in the time before and after the impulse occurs. Regularization has the property of making a causal inverse filter, non-causal. Non-causal filters have a characteristic pre-response, that generates strong audible artefacts mostly in the form of pre-echoes [16]. The delay of the peak of the inverse filter, with respect to the peak of the system impulse response, is a consequence of a modelling delay implemented to avoid the non-causalities that would otherwise make this filter impossible to realize [15]. However, the audible effects of the non-causal part will still be present in the result of the processing. The level of these audible effects is proportional to the level of the raising- and releasing part of the impulse response of the inverse filter [8], [9]. As it can be seen from both filters, in figures 2.7(b) and 2.8(b), the ratio between the peak and the lower amplitudes of the impulse responses are about 40dB for $\beta = 20$dB and 80dB for $\beta = -20$dB. this might indicate that $\beta = -20$dB would have less audible effects than $\beta = 20$dB.

This first look at the inversion problem applied to reproduction of sound shows the ambiguity in choosing the regularization parameter. If the regularization parameter is low ($\beta = 40$dB), the inverse filter is a more precise inversion of the system, but audible artefacts could be introduced in the signal, making the reproduced sound more different from the reference. On the other hand, high regularization ($\beta = -20$dB) gives rise to less audible effects, but the inversion is not so accurate.

In an attempt to attenuate the audible artefacts created by the inversion, a post-processing of the inverse filter is proposed in this project, in a way similar to what Mukai et al. [19] did to eliminate pre-echo noise of

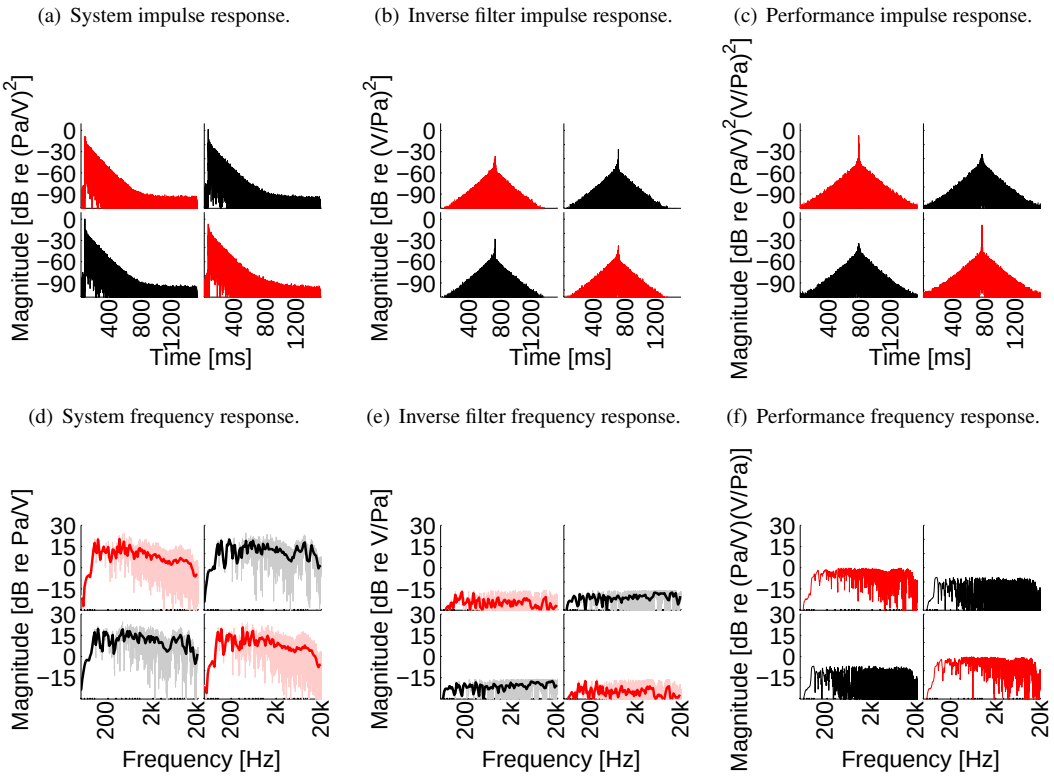


Figure 2.8: 2-by-2 system transfer functions measurements and performance, using regularization of $\beta = -20$dB.

separating filters used for blind source separation. The post-processing consists of applying a window to the impulse response of the inverse filter; this window will fade out the pre- and post-ringing of the response, and improvements in the quality of the reproduction of sound are expected. After many trials and comparison with other windows and preliminary evaluations of the quality, a Gaussian window was chosen. The effect on the impulse response of the inverse filters is shown in figure 2.9.

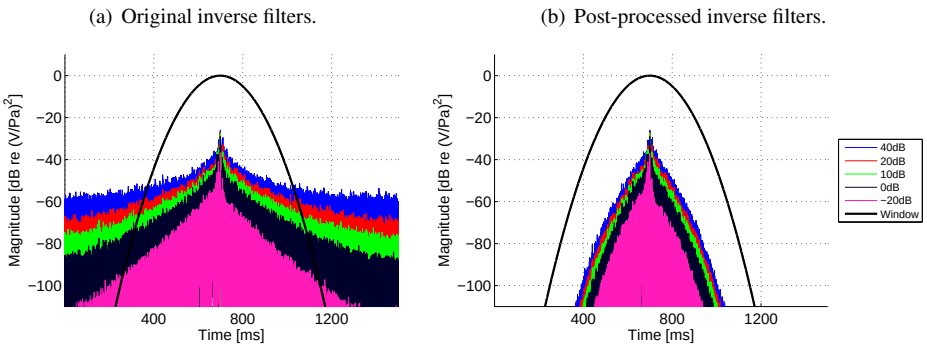


Figure 2.9: Post-processing Gaussian window applied to the channel (1,1) of the inverse filter impulse response of the 2-by-2 system shown in figure 2.7.

In figure 2.10, the calculation of the inverse filter was made using $\beta = 20$dB as in figure 2.7, but in this case the Gaussian window was applied to its impulse response; the effect on the impulse response can be seen in figure 2.10(b). The performance of the inversion tends very roughly, to keep the diagonal elements around zero dB, and tries to push down the crosstalk channels to negative values, as shown in figure 2.10(f).

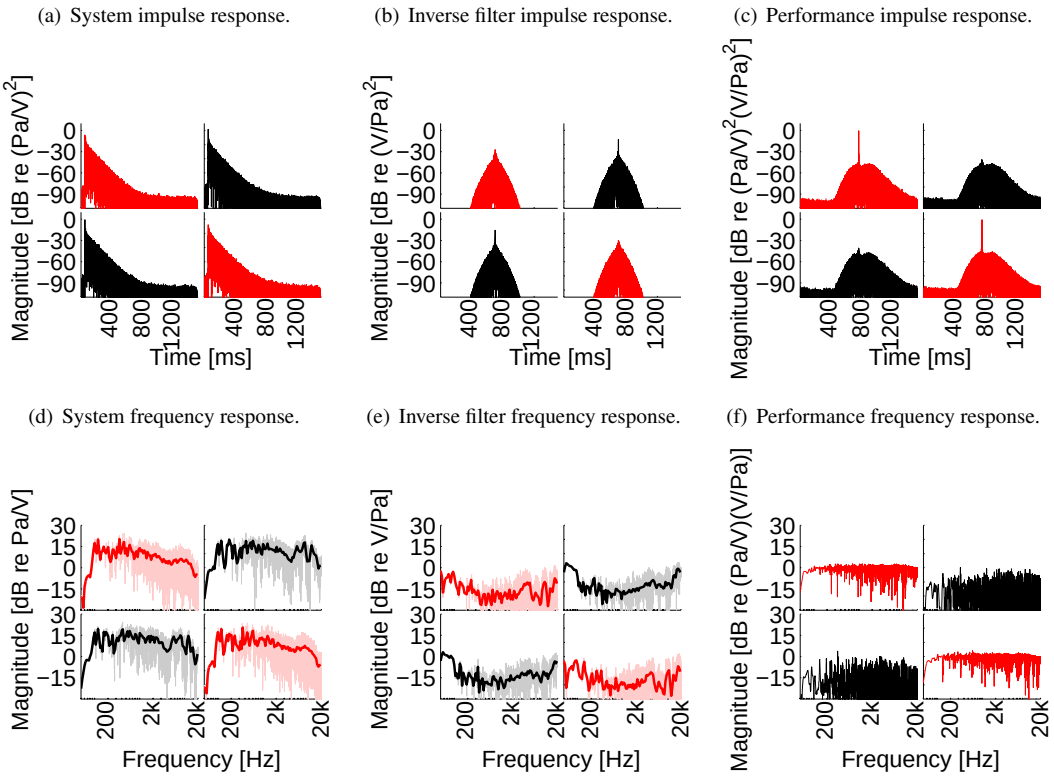


Figure 2.10: 2-by-2 system transfer functions measurements and performance, using regularization of $\beta = 20$dB, and applying post-processing of the inverse filter.

2.3 Recording of the reference sound

The reference background noise programs selected to implement the method are:

- Pink noise: Two channels of incoherent pink noise, generated in `MATLAB`.
- Pop music: Killer queen, by Queen.
- Classical music: String Quartet In E Flat, Op. 33/2, "The Joke" - 1. Allegro Moderato, by The London Haydn Quartet.
- Speech: Female and male mono samples arranged in a stereo sound file, by IEEE 269-2010 [20].

These samples were selected to cover different styles of music: Pop music being more dynamic and having percussive sounds; Classical music being softer and continuous. Speech, because the final application of the method is meant to be the testing of communications devices; and Pink noise, to use as a control signal for comparison and analysis.

The wave files were played back through the reference sound sources (loudspeakers 9 and 10 in figure 2.3), and recorded at the target and validation microphone positions. The frequency content of the recorded sound at microphone 7 is shown in figure 2.11 as an example of the spectra that were treated as reference.

2.4 Filtering of the recorded reference sound

The loudspeaker signals were obtained by filtering the recorded sound with the inverse filters. The way to do this is shown in equation 1.2. The loudspeaker signals are signals that are fed to every channel of the

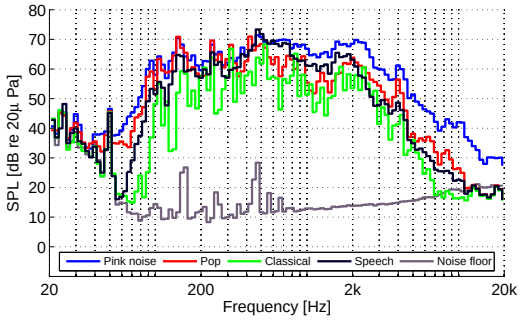


Figure 2.11: Spectrum of the reference sound of different background noise programs, recorded at microphone 7 of the designed microphone array.

loudspeaker array, and that would interact with the system originally identified. This interaction occurs in such a way that, at the target microphone positions, the sound field is optimized to recreate the reference sound as accurately as possible.

2.5 Recording of the reproduced sound

The reproduced sound generated by the loudspeaker array playing back the loudspeaker signals, is recorded at the target and validation microphone positions. Three different configurations of loudspeakers (and microphones) were selected: 8, 4, and 2 channels. When 8 channels were chosen, all the loudspeakers, labelled from 1 to 8 in figure 2.3, and all the microphones of the designed microphone array were used for the optimization. When 4 channels were chosen, the loudspeakers 2, 4, 6, and 8 and microphones 1, 2, 5 and 7; were used. And for 2 channels, the loudspeakers 2 and 8 and microphones 1 and 7 were used. An example of the frequency content of the recorded reproduced sound at microphone 7 is shown in figure 2.12.

2.6 Calculation of the error

The objective evaluation of the method is mainly based on the calculation of the error, the coherence, and the phase difference between the reproduced- and reference sounds in each of the microphone positions. An example of the error of the reproduction of Pink noise, using 8 loudspeakers and applying regularization $\beta = 20$dB, is shown in figure 2.13.

It can be seen that the reproduction of sound at the target positions has errors of around 0dB for frequencies higher than 100Hz and, all the way up to 20kHz. The coherence is higher than 90% for frequencies up to 10kHz, and the phase difference is constant for all frequencies.

At the validation positions, the error follows the same shape as the error at the target positions at low-mid frequencies, but it increases for frequencies higher than 1.5kHz. The same tendency is seen for the coherence and the phase difference.

One criterion to judge the quality of the inversion is that the error should remain within the $\pm$3dB range to be acceptable. For the coherence and phase difference there is not a clear convention for this, only to achieve as much coherence and little phase difference as possible; coherence being a measure of the similarity of the signals, and phase difference showing the time variance of one signal with respect to the other, in the frequency domain.

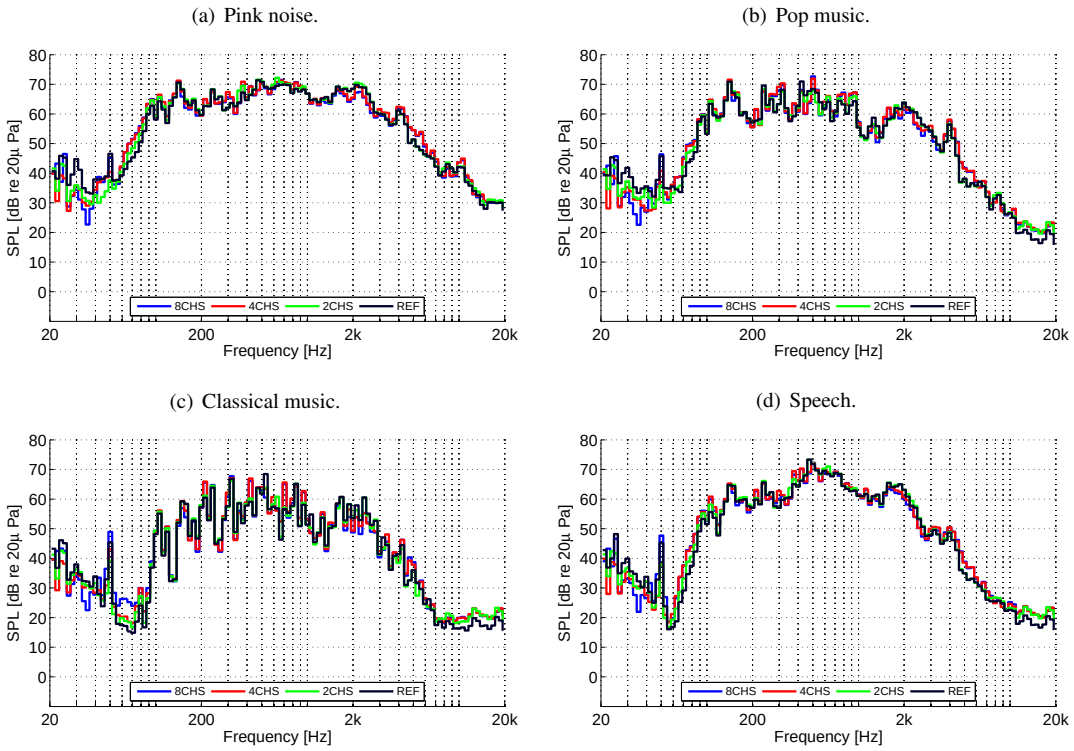


Figure 2.12: Spectrum of the reproduced sound with different numbers of loudspeakers (8, 4, and 2), compared to the spectrum of the reference sound (REF), optimized with $\beta = 20$dB. Recorded at microphone 7 of the designed microphone array, for different background noise programs.

2.7 Magnitude correction of the inverse filters

As shown in section 2.6, the reproduction error is higher at the validation positions than at the target positions at high frequencies. It is essential to minimize the error of the inversion in other positions outside of the designed microphone array, because the microphones of the telecommunications device might not coincide with positions in the microphone array. The reproduced sound was optimized only to those positions, so the tests on the device could be based on sound fields that are not realistic.

A method proposed by HEAD Acoustics GmbH [6], to adjust the magnitude of the error at the validation positions, consists of calculating an equalization filter for the error at the validation positions; and applying that filter to each channel of the inverse filters. The error of the inversion, when this correction is applied, is shown in figure 2.14. At the target microphone positions, it is evident that the side effect of this correction is to push the error out of the $\pm$3dB range. At the validation positions the same effect is seen. However, because the error in this case was very positive before, the correction now adjusts the error to fit inside a range around the $\pm$3dB region.

2.8 Subjective evaluation

A listening test was designed to evaluate the quality of the signal processing involved in this method. The main goal was to learn how much the reproduced sound resembles the reference sound, for different settings and configurations of the inverse filters and the reproduction system.

The MUSHRA test consists in the presentation of a set of test signals that must be compared to a reference. Among the test signals, a copy of the reference is hidden, the objective of which is to set the

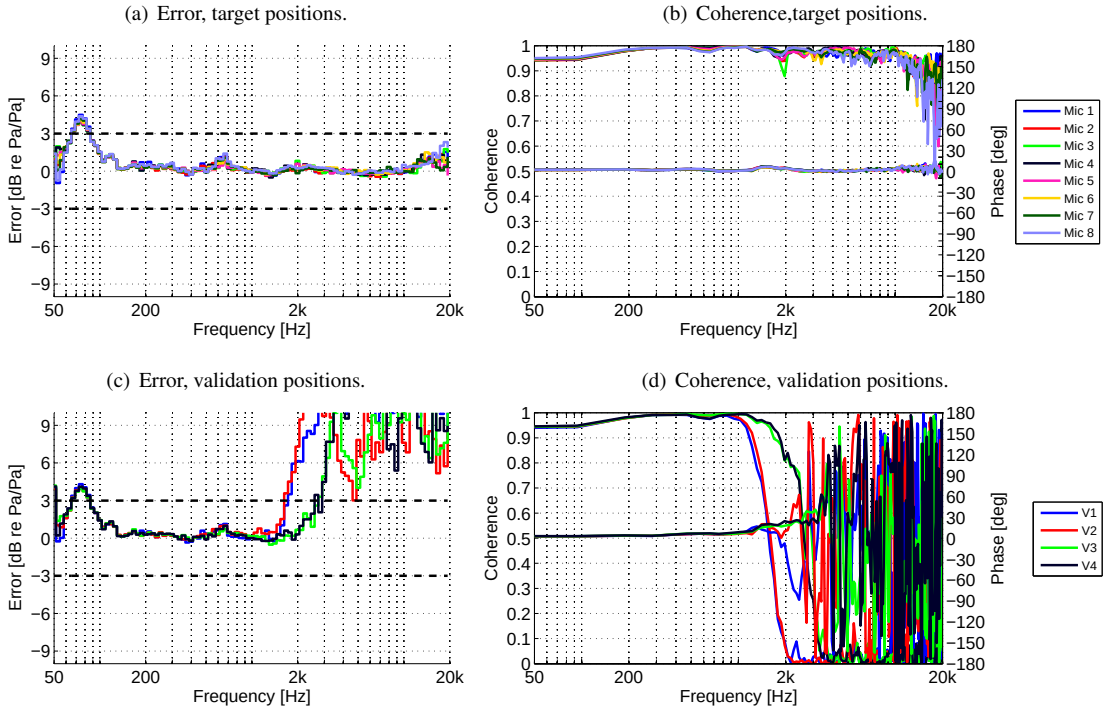


Figure 2.13: Error, coherence and phase difference between the reproduced- and reference sounds, at the target and the validation microphone positions. Regularization parameter: $\beta = 20$dB. Reproduction using 8 loudspeakers of the program material Pink noise.

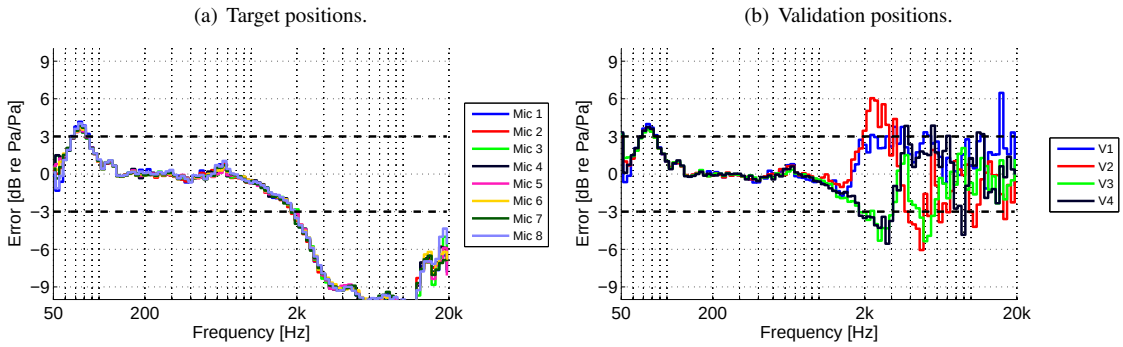


Figure 2.14: Error between the reproduced- and reference sounds, at the target and validation positions. Regularization parameters: $\beta = 20$dB. Reproduction using 8 loudspeakers of the program material Pink noise. Correction applied to the inverse filter.

highest point on the rating scale. The complete test includes an Anchor, which is a signal that is the most different from, or that has the worst quality compared to the rest, setting the lowest point on the rating scale.

A MUSHRA-like test was implemented here to ask the subjects how similar the reproduced sounds were to the reference sound, on a scale from 0 to 100 divided in the categories: "Very different", "Different", "Similar", and "Very similar". In the formal definition, this test was not a MUSHRA because an anchor was not used, just a named reference and hidden reference among the other sounds.

Three experiments were carried out in this test, with the scope of answering the following questions:

1. Which regularization parameter makes the reproduced sound at the validation positions more similar to the reference? The validation positions were selected for the test because it is more likely to have

the telecommunications device placed at positions with similar characteristics to them, than to the target ones.

2. Is there any change in this perception, if the target rather than the validation positions are evaluated?

3. Due to the phase difference between the two microphones used to record the stimuli, is there any change in the rating of similarity, if the signal is presented diotically?

For the first question, three different program materials were presented to the subjects; Pop music, Classical music, and Speech. For the second and third questions, only Pop music was used. The reproduced- and reference sounds were recorded at the validation positions with microphones V1 and V4, and at the target positions with microphones 1 and 7. These microphone signals were presented binaurally through headphones, and in the diotic case the signal of microphone V4 was presented in both ears. The response of the headphones was calibrated in frequency to the HATS B&K type 4100; and in level to the same as was recorded in the room (around 80dB SPL).

In each combination of program material and number of loudspeakers, the optimization of the reproduced sound with different regularization parameters was made: $\beta = -20$dB, 0dB, 10dB, 20dB, 40dB; and 10dB, 20dB, 40dB, applying the post-processing described in section 2.2.1. While playing back approx. 18 seconds of recorded sound, the subjects had the option to smoothly switch between the 8 processed audio samples and the hidden reference (randomly ordered), and to compare to a known reference. See appendix B for details on the execution of the test and the test platform.

The listening test was performed in various meeting rooms of Brüel & Kjær Sound and Vibration Measurement A/S. Five subjects, aged between 25 to 37 years, did the experiment twice, and the average of both results was taken to make a comparison across subjects. All subjects reported to have musical training or experience judging audio quality. Even though it was not possible to perform an audiogram, none of the subjects reported any kind of hearing loss.

CHAPTER 3

Results

The regularization parameter is one of the most important parameters in the reproduction of sound using the deconvolution method. This parameter controls how precise the inversion of the system transfer function is, and depending on its value, some audible artefacts could be controlled (and perceived). The setup of the system used to reproduce the sound, has influence on the result as well; the hypothesis is that if more loudspeakers and microphones are used, the result will improve, but perceptually this might be different. In this section, the results for different numbers of channels will be presented and compared with the perceptual evaluation from subjects.

The signal to be reproduced also has an influence on the quality of its reproduction. If the level of reproduction is too different for different frequencies, there will be more errors in the frequencies where the signal has low amplitude. Some techniques to improve the performance of the reproduction were introduced in sections 2.2.1 and 2.7. The result of applying these techniques is presented here in more depth. Last, but not least, some results of the listening test, especially regarding the microphone positions selected and the listening mode, are presented as well.

In section 2.6, the error, coherence and phase difference between the reproduced- and reference sounds at the target and validation microphone positions, were introduced with two examples; see figure 2.13. As a means of comparison, the maximum error, and the average coherence and average phase difference across microphones will be presented here, so that only one curve, representing all the microphones of the measurement, is used.

3.1 Regularization parameter

The regularization parameter β was the only variable used to adjust the inverse filter in the calculation procedure. In this project, five values were chosen after preliminary tests: $\beta = -20$dB, 0dB, 10dB, 20dB and 40dB. This was done to show how much the error in the reproduction of sound changes for different values, but especially to highlight that a procedure to choose this number, in the specific problem of multichannel reproduction of sound, can not be based on a performance measure only.

In figure 3.1 the maximum error, average coherence and average phase difference are shown for a system composed of 8 loudspeakers and 8 microphones, in which the reference background noise is Pink noise, optimized with the different values of regularization. In appendix A, figures A.1 and A.2 show the same results for Pop music, Classical music, and Speech.

Comparing the error at target and validation positions, it is clear that $\beta = -20$dB is the only value of regularization that, in general, ensures low errors in both microphone positions, although for some frequencies the error exceeds the $\pm$3dB range. For other values of regularization the errors at the target

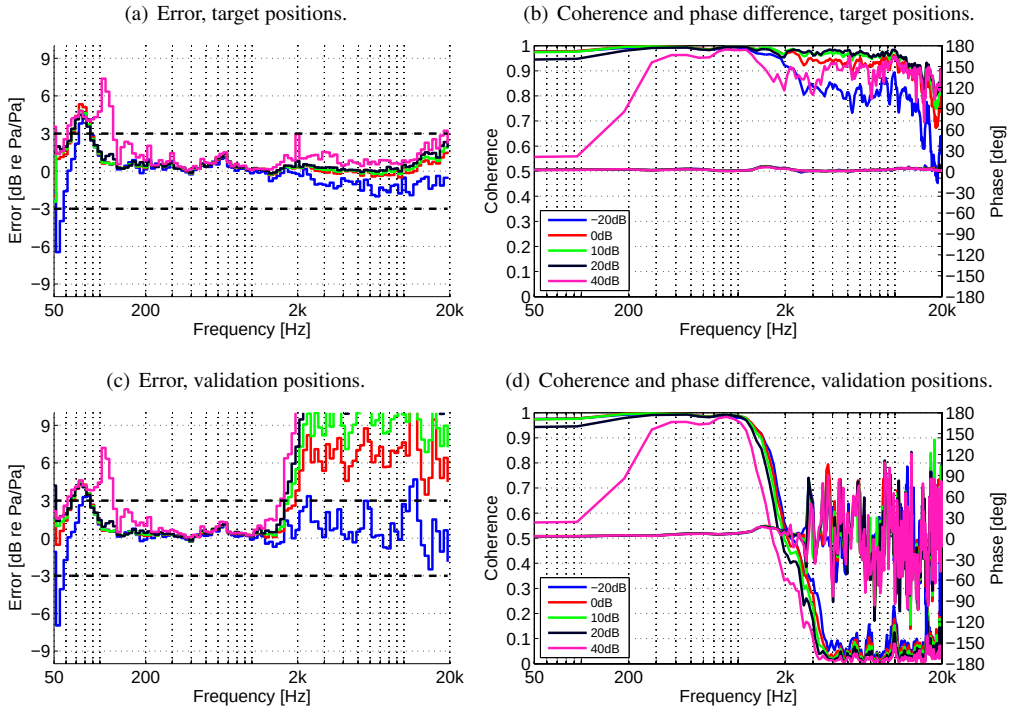


Figure 3.1: Maximum error, average coherence and average phase difference between the reproduced- and reference sounds, across the target and the validation microphones positions. Regularization parameters: $\beta = -20$dB, 0dB, 10dB, 20dB and 40dB. Reproduction using 8 loudspeakers of the program material Pink noise. No correction or post-processing applied to the inverse filter.

positions look better. Nevertheless, $\beta = -20$dB is still the best choice for this system because the error at validation positions does not increase to more than 6dB for frequencies higher than 1.5kHz, as it does for lower values of regularizations.

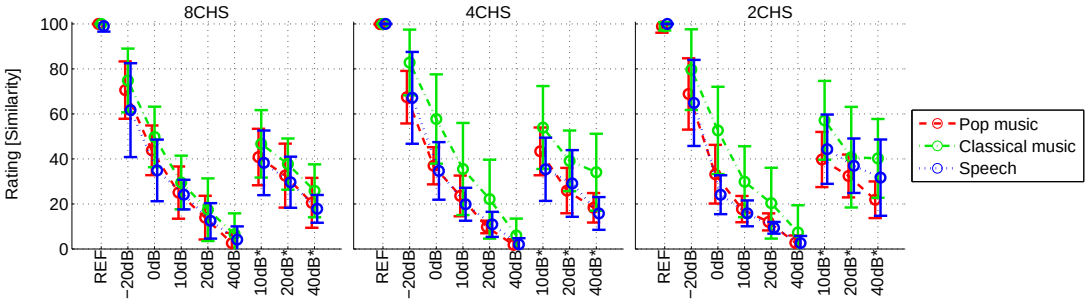
As was said, at validation positions the error at high frequencies increases because the optimization of the sound field is local to the designed microphone array and not to other positions. For $\beta = 40$dB, low values for the coherence are also present at low frequencies for both, target and validation positions. Between $\beta = -20$dB, 0dB, 10dB and 20dB the coherence and phase difference do not differ much. The error gradually increases as the regularization parameter becomes smaller.

The first result of the listening test is shown in figure 3.2. Three background noise programs were evaluated: Pop music, Classical music, and Speech; reproduced through systems of 8, 4, and 2 loudspeakers. Each of the plots shows three regions. One corresponding to the rating of the hidden reference, the second corresponding to the reproduced sound obtained with regularization parameters $\beta = -20$dB, 0dB, 10dB, 20dB and 40dB without post-processing. And the third region, showing the ratings for $\beta = 10$dB, 20dB and 40dB with post-processing of the inverse filter. The circles are the average ratings across subjects and the error bars represent the 95% confidence interval.

On average, the subjects were able to easily identify the hidden reference under all the conditions. It was found that $\beta = -20$dB was the regularization parameter with the highest rating, and the ratings decreased as the regularization parameter decreased. This was observed for all the program materials and number of loudspeakers. In most of the cases, the samples in which the post-processing of the inverse filter were applied got higher ratings than the corresponding regularization value without post-processing; however, the improvement was not dramatic.

When comparing the ratings given by the test subjects, figure 3.2(a), to the maximum error, average

(a) Number of channels vs. Reference background noise programs.



(b) Reference background noise program vs. Number of channels.

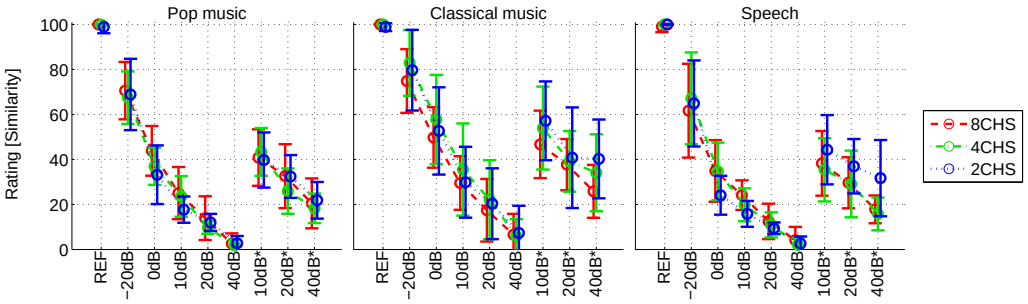


Figure 3.2: Average ratings across subjects, the error bars represent 95% confidence interval.

coherence and average phase difference measured, figure A.2, it can be seen that there is agreement between the measured error and the perceived similarity between the reproduced- and reference sounds. For $\beta = -20$dB, the highest rating, and the lowest errors were measured. For $\beta = -40$dB, the highest error and phase difference occur, and the lowest ratings were obtained.

3.2 Number of channels

The number of channels in the system refers to the number of microphones used to record the reference sound, and the number of loudspeakers to reproduce it. In all the cases the number of loudspeakers was the same as the number of microphones: 8, 4, or 2.

The objective evaluation parameters: maximum error, average coherence and average phase difference, are shown in figure 3.3; for different numbers of loudspeakers, all reproducing Pop music, and optimized with $\beta = 0$dB. For other values of the regularization parameter the same conclusions on the results apply, and the reproduction using different number of loudspeakers did not yield different conclusions when the number of channels was changed. With exception of $\beta = -20$dB, the results for which are shown in figure A.3.

It can be seen that at the target positions there are not differences on the optimization with different number of channels. At the validation positions there is a clear improvement in the reproduction from 2 to 4 channels, but the response is almost the same with 4 and 8 channels. This is due to the longer separation of the microphone positions, in the case of 2 channels.

When $\beta = -20$dB was used to calculate the inverse filters, the difference between the reproduced- and the reference sound was more sensitive to changes in the number of channels. Even though none of the error curves in figure A.3 fit inside the range of ± 3dB, there is less variation from target to validation microphone positions.

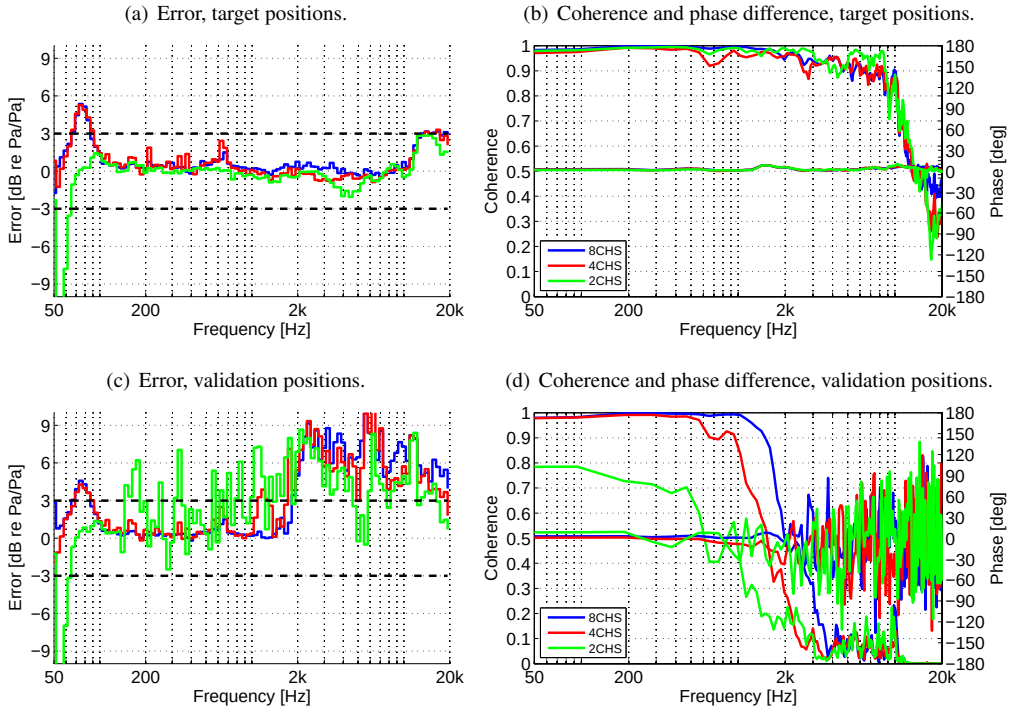


Figure 3.3: Maximum error, average coherence and average phase difference between the reproduced- and reference sounds, across the target and the validation microphones positions. Regularization parameter: $\beta = 0$dB. Reproduction using 8, 4 and 2 loudspeakers of the program material Pop music. No correction or post-processing applied to the inverse filter.

The result of the listening test presented in figure 3.2(b), shows the same result. For the average ratings across subjects, reproductions with different numbers of loudspeakers were all evaluated within the same range of similarity, with no clear preference of any of the setups.

3.3 Reference background noise program

In general, the maximum error between the reproduced- and reference sounds was not found different for different background noise programs. For both, target and validation microphone positions, the reproduction error for the four different program materials follow the same trend according to the regularization parameter used. Pink noise was not so sensitive to reduction of the regularization parameter, but the other three programs registered errors approx. 2dB higher for frequencies higher than 10kHz, see figure 3.4.

At the target microphone positions the coherence and the phase difference of the program materials Pop music, Classical music, and Speech, decreased for frequencies higher than 1.5kHz in higher amount compared to Pink noise, see figure A.4. This will be discussed in section 4.3.

It was checked for different number of loudspeakers (plots not shown in this report), and the behaviour described at the target microphone positions was the same. The coherence decreased for high frequencies and the phase difference between the reproduced- and reference sounds got the bent shape.

Regarding the listening test, figure 3.2, there seems to be a small tendency in the ratings of Classical music to be higher than other program materials; however, it is also the one with wider confidence intervals and the improvement is not enough to state a conclusion.

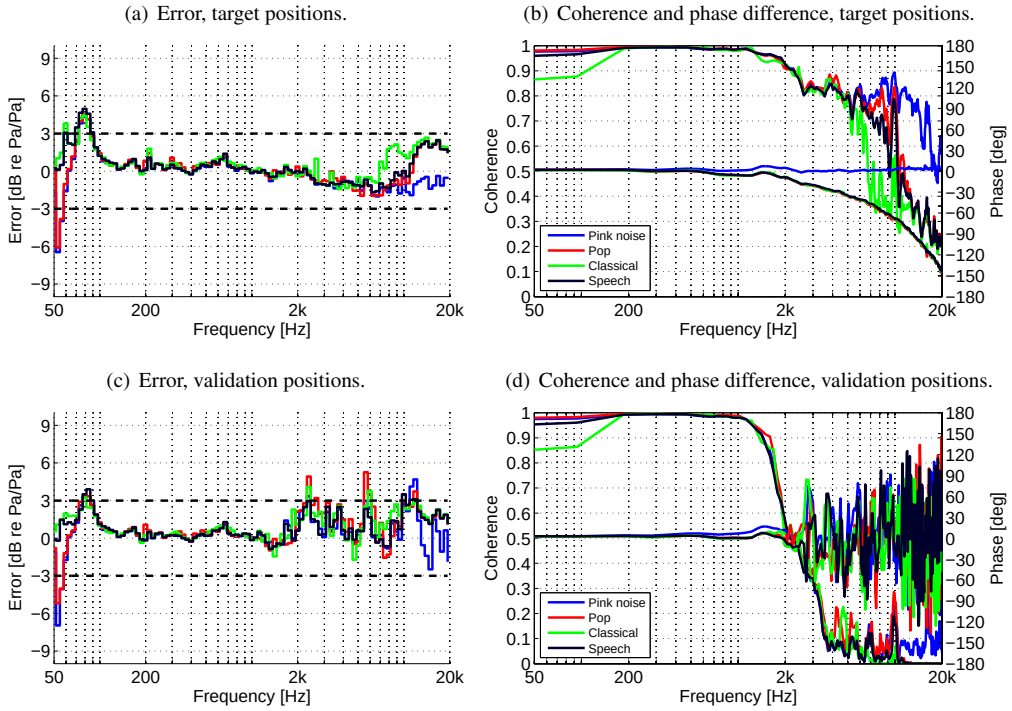


Figure 3.4: Maximum error, average coherence and average phase difference between the reproduced- and reference sounds, across the target and the validation microphones positions. Regularization parameter: $\beta = -20$dB. Reproduction using 8 loudspeakers of the program materials Pink noise, Pop music, Classical music and Speech. No correction or post-processing applied to the inverse filter.

3.4 Correction and post-processing of the inverse filter

Two techniques were proposed to improve the reproduction of sound. One technique is an adjustment of the inverse filter based on the magnitude error between the reproduced- and reference sounds at the validation positions. The other one is an adjustment in the impulse response of the inverse filter, by means of a window that fades out the pre- and post-ringing, whose levels have a direct relation with the audible artefacts produced by the filtering.

In figure 3.5 and 3.6, the maximum error, average coherence and average phase difference between the reproduced- and reference sounds are shown for a system optimized with regularization parameter $\beta = 10$dB, reproducing Pink noise with 8 loudspeakers. Both figures show the target and validation positions, when the correction after the magnitude error was measured at validation microphone positions, and when the post-processing is applied, respectively. Similar conclusions can be drawn for other values of regularization.

The correction applied, after the magnitude error at validation positions was measured, improves the error at the validation positions as expected. However, the effect of the equalization filter calculated for such correction has a damaging effect at the target positions, and possibly at other positions that are not accounted. For this kind of correction the coherence and the phase difference are not affected, but the side effect on the error is a sufficient argument to not recommend its application.

In the error plot, the effect of the post-processing means a very small improvement at high frequencies. The phase difference is not affected by the post-processing, but the coherence is lowered by almost 10% at frequencies higher than 1kHz. The same result for two other values of the regularization parameter are presented in the appendix A, figures A.5 and A.6. For $\beta = 20$dB, the conclusion is the same as before, but for $\beta = 40$dB the post-processing window improves the coherence between the reproduced- and reference

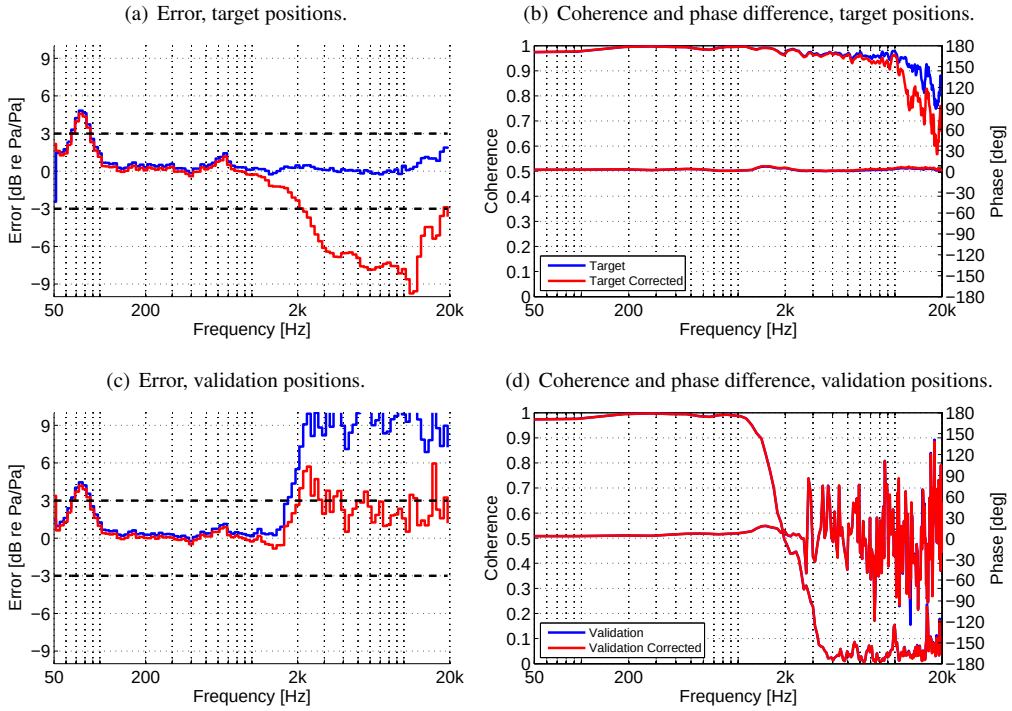


Figure 3.5: Maximum error, average coherence and average phase difference between the reproduced- and reference sounds, across the target and the validation microphones positions. Regularization parameter: $\beta = 10$dB. Reproduction using 8 loudspeakers of the program material Pink noise. No post-processing but correction applied to the inverse filter.

sounds at low frequencies.

The reproduction of Pop music was optimized with three regularization parameters, $\beta = 10$dB, 20dB, and 40dB, whose inverse filters were post-processed. These three samples were presented in the listening test, together with the samples that did not have post-processing in their inverse filters. The result of the rating given by the subjects is shown in figure 3.2. As it was said before, in all cases the ratings given for the post-processed samples were higher than the corresponding sample without post-processing.

3.5 Position of recording

It has been said that the method for reproducing sound using the deconvolution method is also called local reproduction of sound (when small number of loudspeakers are used) [3], because the optimization of the reproduced sound is only at the microphone positions of the designed microphone array. The procedure described in this project is based on the idea of optimizing the reproduced sound at positions around the head, near which the telecommunications devices could be mounted.

Ideally, the reproduced sound using the deconvolution method should be identical to the reference sound at the target microphone positions. In analogy to the Nyquist sampling theory [3], if a point is defined in the middle of two target positions, the reproduction at that point should be ideally perfect too up to the frequency $f = \frac{c}{2d}$, where c is the speed of sound, and d is the distance between the two adjacent target positions. So, validation microphone positions were defined in order to check the reproduction of sound outside of the microphone array.

Theoretically, the frequency limits for of the designed microphone array are 1786.5Hz for microphones V1 and V2, 3460Hz for microphone V3, and 3897.7Hz for microphone V4. The results (examples in figure

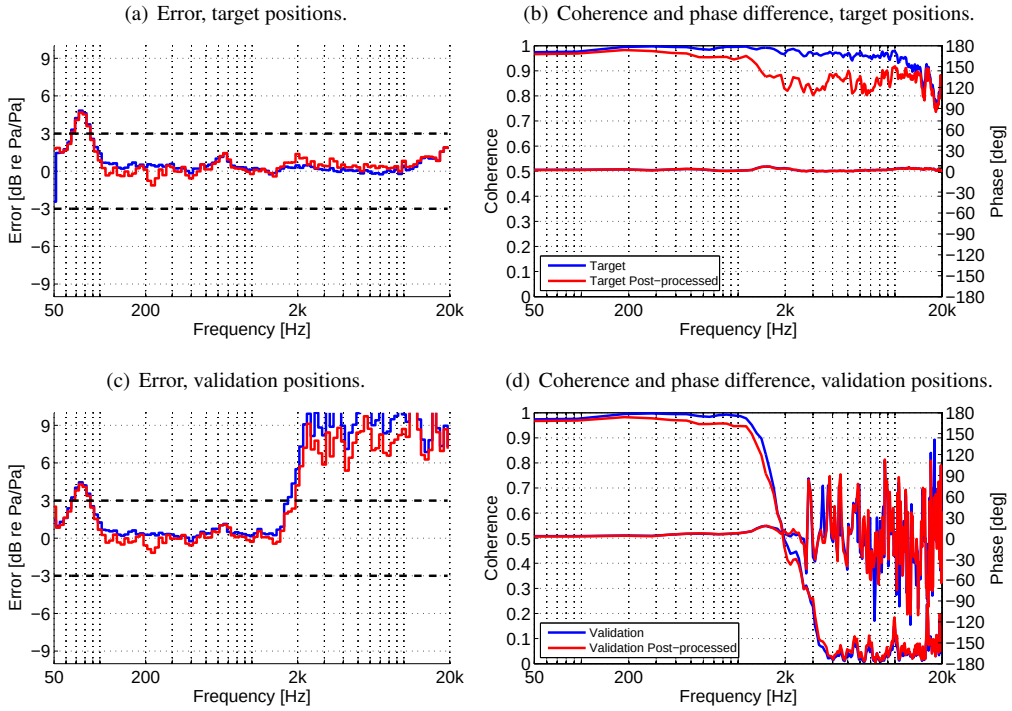


Figure 3.6: Maximum error, average coherence and average phase difference between the reproduced- and reference sounds, across the target and the validation microphones positions. Regularization parameter: $\beta = 10$dB. Reproduction using 8 loudspeakers of the program material Pink noise. No correction but post-processing applied to the inverse filter..

2.13) did not exhibit agreement with these values. The reason of this disagreement is that the frequency limit is the extreme value at which the method does not work at all; however, from lower frequencies, the error has started to increase because the error at validation positions does not depend only on the optimization of the reproduction at the adjacent target microphones, but the solution of the inverse problem takes into account the transfer functions of the whole set of loudspeakers to all the microphones.

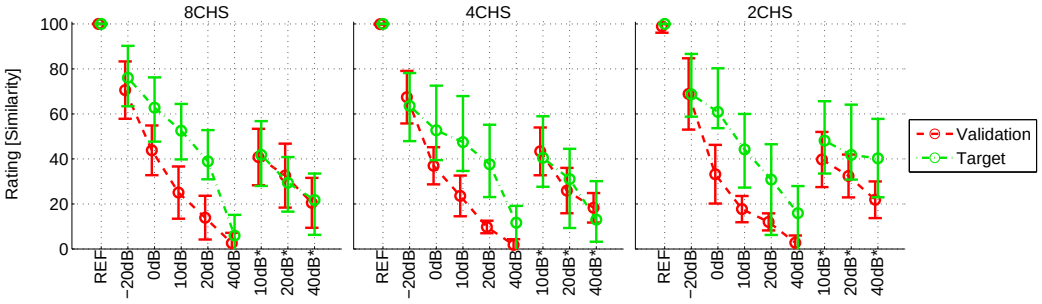
The main difference in the reproduction of sound at the target and validation microphone positions is that the error, coherence and phase difference between the reproduced- and reference sounds, are guaranteed in the approx. frequency range from 100Hz to 10kHz for the target positions, and from 100Hz to 1.5kHz, at the validation positions.

In reality, the telecommunications device that will be exposed to the reproduced sound, will not be exactly neither at the target microphone positions, nor at the validation microphone positions. This problem can be partially solved by changing the design of the microphone array, optimizing it to the location of the microphones of the device. The drawback is that depending on the size of the device the microphone array might change the sound field around it, which is not desired; or the inversion problem can become more challenging due to the proximity of the target microphones.

Within the limits of the setup of this project, the perception of the similarity between the reproduced- and reference sounds, at the validation compared to the target positions, was investigated in the listening test.

The program material Pop music was presented in two separate experiments, in which the reproduced sound presented was recorded at the validation and target positions, respectively. The conditions of comparison were the same in both experiments, five samples optimized with regularization parameters $\beta = -20$dB, 0dB, 10dB, 20dB, and 40dB without post-processing, three samples optimized with regularization parameters $\beta = 10$dB, 20dB, and 40dB, applying post-processing, and a hidden reference.

(a) Number of channels vs. Position of recording.



(b) Position of recording vs. Number of channels.

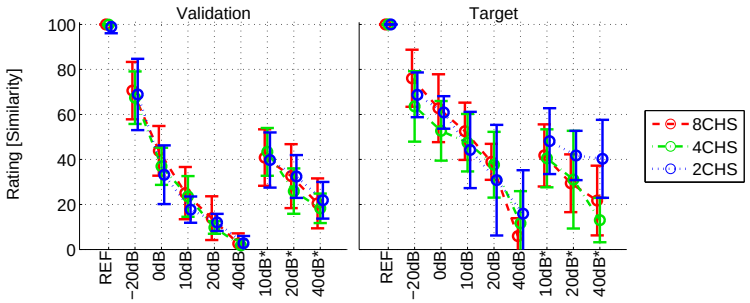


Figure 3.7: Average rating across subjects, the error bars represent 95% confidence interval.

The ratings given for the samples optimized with $\beta = 0$dB, 10dB, and 20dB were significantly higher for the target positions with respect to the validation positions, see figure 3.7. For the limits $\beta = -20$dB, and 40dB, the similarity at validation position, was almost the same as at the target position; and the post-processing did not have any effect in the microphone positions compared.

3.6 Presentation of the stimuli

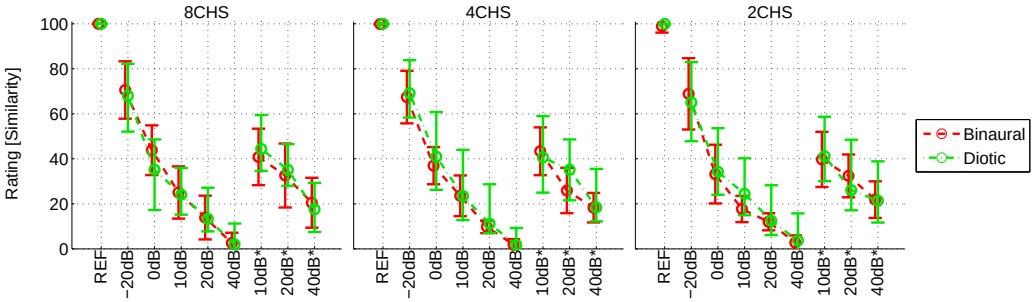
The stimuli presented in the main experiment of the listening test, were the signals recorded using the microphones V1 and V4, of the reproduced- and reference sounds optimized with different regularization parameters. These microphones were selected because they were the ones closer to the ears.

Due to the distance from the validation microphones to the ears, the signals presented in the listening test are not precisely adequate for reproduction using headphones. However, since the optimization of the sound around the HATS did not involved its microphones, the best approximation was to use these microphone. The goal of the listening test was to evaluate the quality of the signal processing involved, and the performance of the method itself for reproduction of sound with respect to a reference, not to study attributes of the stimuli.

Because the signals recorded, were not binaural stimuli per se., phase differences between both microphones are inherent due to recordings made at different points of space. The question about its presentation through headphones arose. Another experiment in the same listening test was made to investigate if there was any advantage in presenting the stimuli binaurally over diotically, or vice versa.

The reason to doubt about this is that it was unknown if the phase difference between the signals presented to left and right ear, would help to identify audible artefacts or not. The results in figure 3.8 show that the subjects did not get any improvement is the perception, for the binaural or the diotic listening, for any of the configurations of loudspeakers used.

(a) Number of channels vs. Presentation of the stimuli.



(b) Presentation of the stimuli vs. Number of channels.

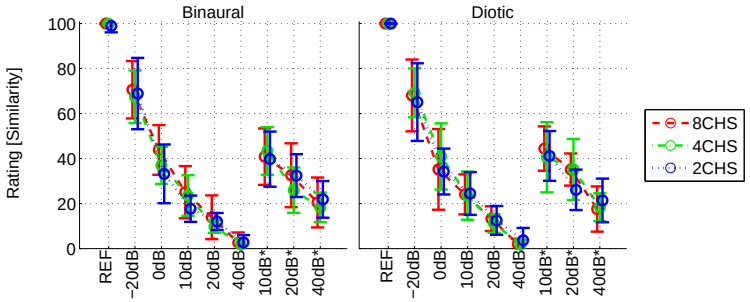


Figure 3.8: Average rating across subjects, the error bars represent 95% confidence interval.

CHAPTER 4

Discussion

In this section, some of the results presented in chapter 3 are re-described, analysed, compared or related to other results. The main goal of this project is not to stablish a procedure for the determination of the best setup of the system to reproduce sound using the deconvolution method, but to show how each of the decisions taken affect the result in practice.

4.1 Regularization parameter

According to literature [6], [14], one procedure to select the regularization parameter should start by choosing a number, and then lowering it by checking that the amplitude of the loudspeakers signals obtained do not saturate the reproduction system, mainly the loudspeakers. This procedure was shown to be not optimum because even for low values of regularization the error at target positions was higher than the error for $\beta = 20$ or higher. The error was unacceptable for validation positions at high frequencies. In all cases, the amplitude of the loudspeakers signals were in a safe range to not saturate the loudspeakers, so distortion was not one source of error.

In the listening test, the reduction in rating of similarity, as the regularization decreased, is a direct relation with the audible artefacts present in the sound. The audible artefacts can not be seen in the error; however, the low coherence measured for $\beta = 40$dB (figure A.2) suggests a relation with the audible artefacts and the coherence.

In opposition to what the theory suggests, better performance, objectively and subjectively, was achieved for the highest regularization parameter of the values chosen. The regularization parameter $\beta = -20$dB was the best choice for this system of 8 loudspeakers. Even though it is not shown in this report, similar results were found for the systems of 4 and 2 loudspeakers.

4.2 Number of channels

It is natural to think that having more channels will ensure better performance in the reproduction due to the higher density of microphones around the telecommunications device, and the energy that is radiated by more loudspeakers. This idea was not entirely true in the experiments carried out in this project.

Ideally, at the target positions there should not be any difference, because the local reproduction of sound guarantees the reproduction at those positions, no matter what the number of channels is. However, having more loudspeakers gives the possibility to reproduce sound at more points around the telecommunications device. At validation positions it was expected an improvement for higher number of channels that was not measured objectively or subjectively.

In figure 4.1, the z-score of the ratings presented in figure 3.2 are shown. The z-score, is defined as the number of standard deviations that a given observation is above the mean [21]. It is calculated using the formula 4.1:

$$z = \frac{x - \mu}{\sigma} \tag{4.1}$$

where x is the ratings of the experiment, μ is the mean, and σ is the standard deviation.

The z-scores calculated, show the optimal range that the subjects used to rate the audio samples, normalized across subjects. They show that the differences of ratings of the reproduction using different number of loudspeakers are negligible at the validation positions.

This confirms then, that $\beta = -20$dB was the highest rated regularization parameter, in accordance with the objective result presented in section 4.1.

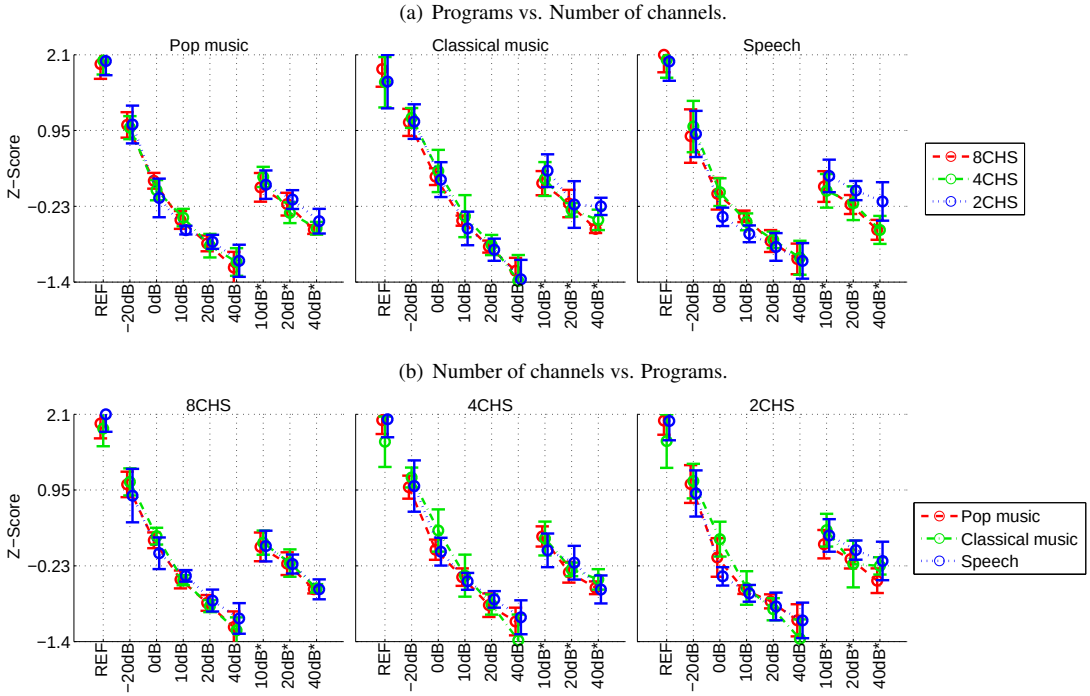


Figure 4.1: Z-score calculated across subjects, the error bars represent 95% confidence interval.

4.3 Reference background noise program

In figure 2.11, the spectra of the reference sounds of different background noise programs were presented for the recording made at one of the microphones of the designed microphone array. It is obvious that, even though the program materials might have different spectral content, the response of the reference sound sources (figure 4.2), in conjunction with the room and the position of the microphone, have shaped the spectra of the four signals recorded. At low frequencies ($f < 100$Hz) the spectral content is clearly dominated by the noise floor.

The loudspeaker signals calculated were processed with the same inversion filters, accordingly. The differences in the coherence and phase response at high frequencies should be due to the programs themselves, see figure A.4. The spectra of the sounds at high frequencies could explain that the decrease in the coherence and phase difference is due to the low energy level of the recorded reference sound, because more noise

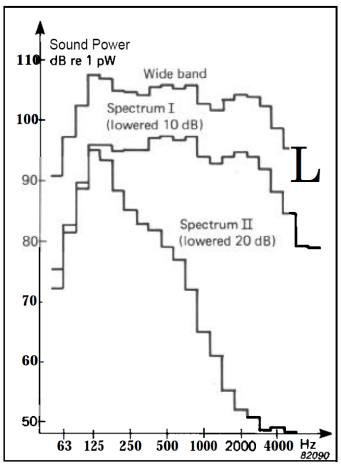


Figure 4.2: Figure and caption taken from the product data of the Sound source B&K type 4224 [22]: Sound power spectra with the 4224 operating at full power for a) Wide Band b) Spectrum I and c) Spectrum II. For the sake of clarity the curves for Spectrum I and Spectrum II have been lowered 10dB and 20dB respectively.

was induced in the measurement. In figure 2.11, it is shown that for high frequencies, the spectral content of different signals start to be more different between each other, having the less amplitude the program materials Pop music, Speech and Classical music, in that order. Reason for which some signals have decreased coherence and phase differences, and some others do not.

The z-score of the ratings for different program materials reproduced with different number of loudspeakers, figure 4.1(b), confirms that the tendency for higher ratings of Classical music is not significant with respect to other programs. In fact, the z-score shows that in the normalized range of similarity, all the subjects judged the three program materials in the same way.

4.4 Correction and post-processing of the inverse filter

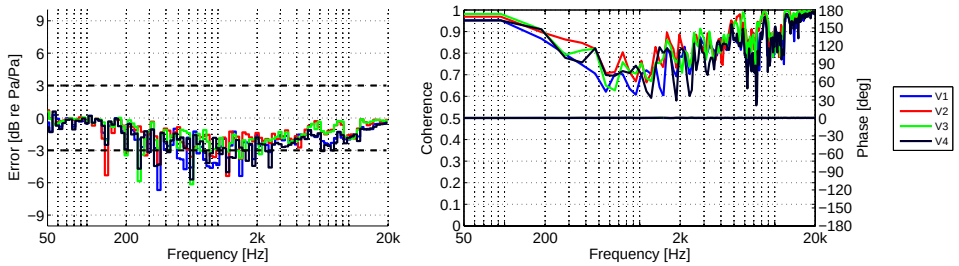
Two main differences exist between the two techniques proposed to correct the inverse filters. The correction of the inverse filter using the post-processing window in the impulse response is made before the evaluation of the error of the reproduction; and the error introduced depends on the way that each filter is calculated, i.e. the regularization parameter. The correction of the filter using the magnitude error at the validation positions takes into account the signal that is being reproduced, but the error introduced affects the target positions directly.

The correction after the magnitude error was measured at validation positions had no effect at low frequencies ($f < 100$Hz). This is due to that, even though at low frequencies an equalization filter should have reduced the amplitude of the error by approx. 3 to 5dB, the response of the loudspeakers at those frequencies is already very low to allow significant improvements.

Even though the correction using the equalization filter of the magnitude error at validation positions was qualified as non optimum due to the effect at target positions, the correction using the post-processing is not free of distortions in the performance either. The error, the coherence and the phase difference between the original and the post-processed impulse response of the filters derived using $\beta = 20$dB and $\beta = -20$dB are shown in figure 4.3.

It is evident how the effect of the post-processing is different for each regularization parameter, and it can also be concluded that the lower the regularization is, the more error is introduced. In spite of the error induced by the post-processing, the increase in the similarity, the improvement in coherence at low frequencies, and the null effect in the phase difference; are reasons to suggest the post-processing as a good tool to reduce the perception of audible artefacts, but deeper investigation needs to be done to maximize its effects.

(a) Error, $\beta = 20dB$. (b) Coherence and phase difference, $\beta = 20dB$.



(c) Error, $\beta = -20dB$. (d) Coherence, $\beta = -20dB$.

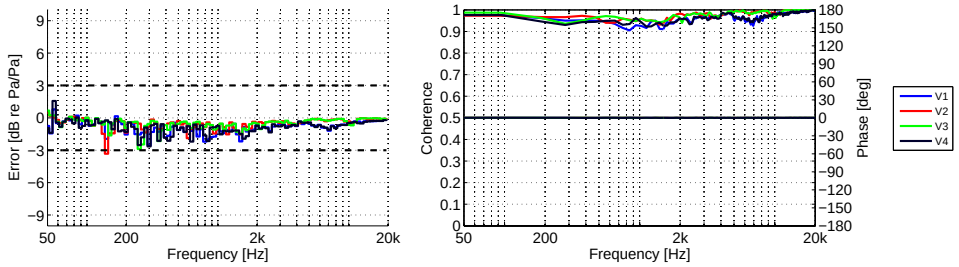


Figure 4.3: Error introduced by the post-processing window, coherence and phase difference between the original and post-processed impulse response.

By comparing the z-scores calculated for different program materials and different number of loudspeakers, the post-processed samples were indeed rated at higher similarity than the counterparts when no post-processing was applied.

Summary and conclusions

A method of realistic reproduction of sound was implemented for the application of testing telecommunications devices. This was done using the deconvolution of a matrix of transfer functions between loudspeaker- and target microphone positions. Different variables of the method were analysed: the regularization parameter, the number of channels used for the reproduction, different background noise programs, and two correction techniques to improve the performance of the reproduction.

This work was based on the proposal for a new standard of reproduction of sound for testing terminals using HATS. This project provides an insight into that proposal, mainly on the procedure suggested to determine the regularization parameter, and the corrections applied for improving the performance of the method. Subjective evaluation of the similarity between the reproduced- and reference sounds, complements the objective analysis of the error, coherence and phase difference.

A commonly recommended procedure to fix the regularization parameter was not a very efficient methodology in this project. That procedure starts by selecting a value for the regularization, and reducing it until the loudspeaker signals saturate the system. It was found that, within the selected values, the lowest regularization was the one that led to more errors and audible artefacts, even though the loudspeakers signals did not saturate the system. The reverberation time measured was relatively high, which made the inversion problem more challenging to solve. Objectively and subjectively, it was found an agreement in the tendency that, highest regularization produced less errors and more similarity, but it is needed more investigation in this matter to find the inflexion point in which the similarity starts to decay and the errors are not acceptable any more.

Objectively, there was no difference in the reproduction when 4 or 8 loudspeakers were used for the reproduction. However, due to the reduction of resolution at the validation positions, there were more errors when 2 loudspeakers were used. Subjectively, there was no difference among the setups, and the low confidence intervals of the ratings suggests a good agreement among the subjects.

No differences in the reproduction errors were found for different program materials, and the subjective evaluation provided the same conclusion. At higher frequencies than 1.5kHz, though, the coherence and phase differences measured between the reproduced- and reference sounds at the target microphone positions decreased, and a possible explanation for this was that the low energy of the signals in this frequency range introduced noise for the comparison.

The main problem of the method using deconvolution is the creation of audible artefacts in the form of pre-echoes, due to the regularization. It was found an agreement with respect to the literature, in a way that lower regularization values would ensure a more identical reproduction of sound, but the creation of audible artefacts make it actually more different, subjectively.

Two techniques to correct the error of the reproduction, and to ensure that for optimization with low

regularization parameters, low error and low level of the audible artefacts occur, were investigated in this project. The first was a proposal by HEAD Acoustics GmbH, that consisted in the calculation of an equalization filter based on the error at the validation positions. The advantage of this correction is that it takes into account the error for the particular background noise, and the features of the inverse filter; but it damages the response at other positions. The second proposal was the post-processing of the inverse filters' impulse responses, with a window to attenuate the pre- and post-ringing created by the regularization. This correction was found to enhance the similarity, improving the coherence in the low frequency range, with small refinement in the error, and no change in the phase difference between the reproduced- and reference sounds.

The advantage of the deconvolution method with respect to other methods, is that it takes into account the response of the loudspeakers and the crosstalk cancellation. However, the reproduction of sound is only local to the points selected for optimization. The reproduction of sound at those positions was found to have errors within ± 3dB in the frequency range from 100Hz to 20kHz, and the same performance was maintained at different positions up to 1.5kHz. Improvements in the perception of similarity at target positions with respect to the perception at validation positions, were only obtained for some values of regularization parameter. For the highest and the lowest regularization parameters selected, the similarity was equally judged for both positions.

In the subjective evaluation process, it was unknown whether the phase differences, between the two microphones used for recording the headphones signals, would help to identify more audible artefacts, or not. The listening test included an experiment, in which diotic presentation of the stimuli was compared to binaural, and the results did not show any improvement of one presentation mode with respect to the other.

The method of deconvolution applied to reproduction of sound was found to be a good technique for testing telecommunications devices. By defining particular points in space where the optimization is desired, low cost setups and fast implementation can ensure low errors between the reproduced- and reference sounds. However, the method is very sensitive to introduce audible artefacts, and a procedure for defining the regularization parameter in multichannel reproduction sound systems that considers the perceptual performance is not yet well defined in literature.

Even though in this project a real telecommunications device was not used, a customized design of the microphone array specific for the device could be a good idea to optimize the reproduction of sound, just in the points of interest, i.e. at the microphones of the device. Better performance could be obtained in a wider frequency range; however, the inversion process might become more challenging, because positioning the microphones closer to each other can potentially make the problem more ill-conditioned.

For future works, it is necessary to approach the determination of the regularization parameter from a more psycho-acoustical point of view. A better characterization of the audible artefacts created by the regularization, could provide tools to compensate for them, or to the determination process of the regularization itself. Better listening tests, with an optimal panel of listeners would also give more reliable results. The subjective evaluations could involve listening in the actual room, or include the signal processing of the telecommunications device, not only recordings presented over headphones. Different rooms can also be compared in the analysis, and real background noise recordings made with the designed microphone array need to be tried.

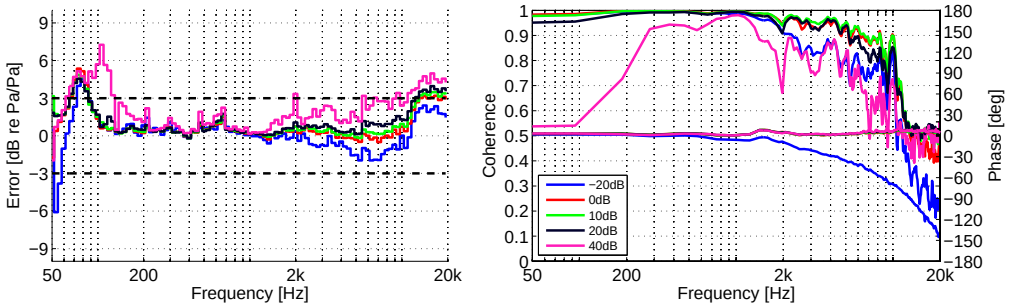
Appendices

APPENDIX **A**

Complementary figures

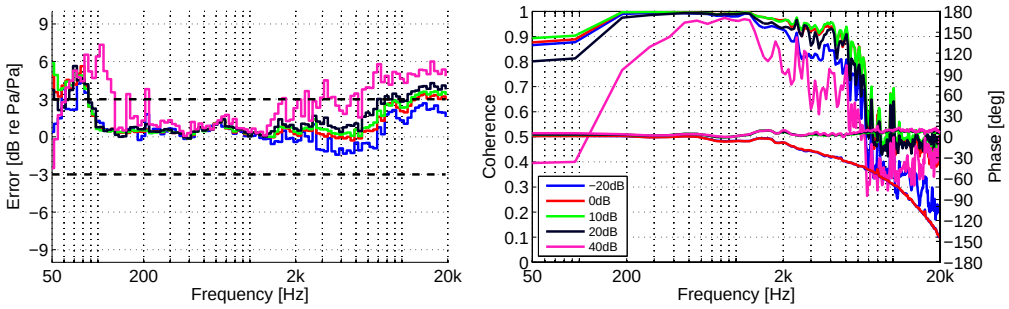
(a) Error, Pop music.

(b) Coherence and phase difference, Pop music.



(c) Error, Classical music.

(d) Coherence and phase difference, Classical music.



(e) Error, Speech.

(f) Coherence and phase difference, Speech.

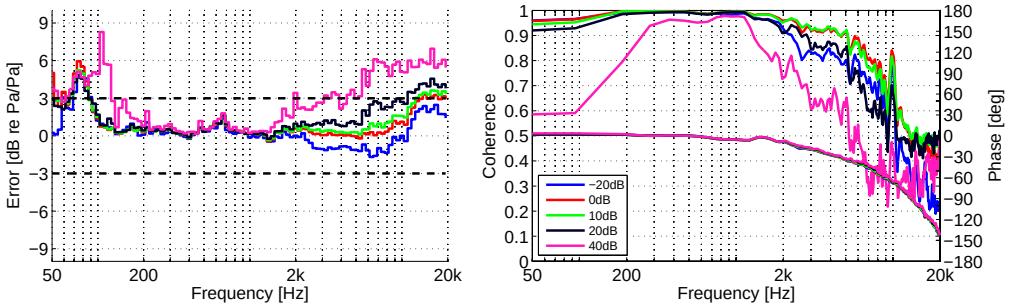
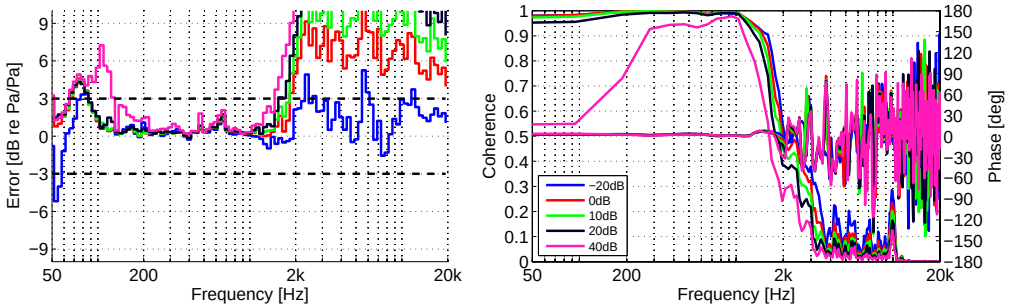


Figure A.1: Maximum error, average coherence and average phase difference between the reproduced- and reference sounds, across the target microphones positions. Regularization parameters: $\beta = -20$dB, 0dB, 10dB, 20dB and 40dB. Reproduction using 8 loudspeakers of the program materials Pop music, Classical music, and speech. No correction or post-processing applied to the inverse filter.

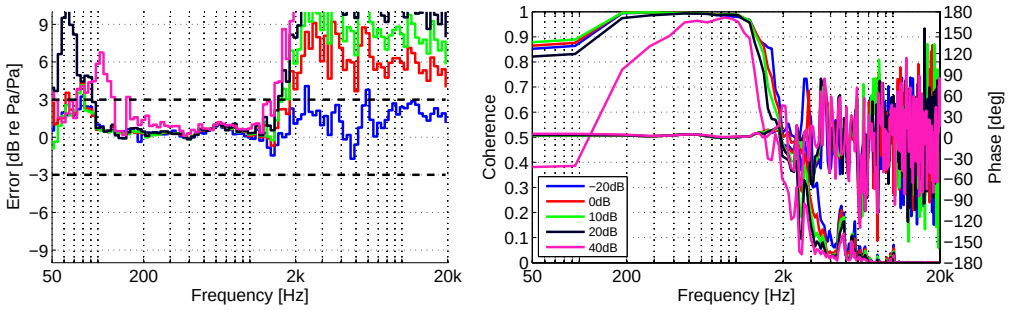
(a) Error, Pop music.

(b) Coherence and phase difference, Pop music.



(c) Error, Classical music.

(d) Coherence and phase difference, Classical music.



(e) Error, Speech.

(f) Coherence and phase difference, Speech.

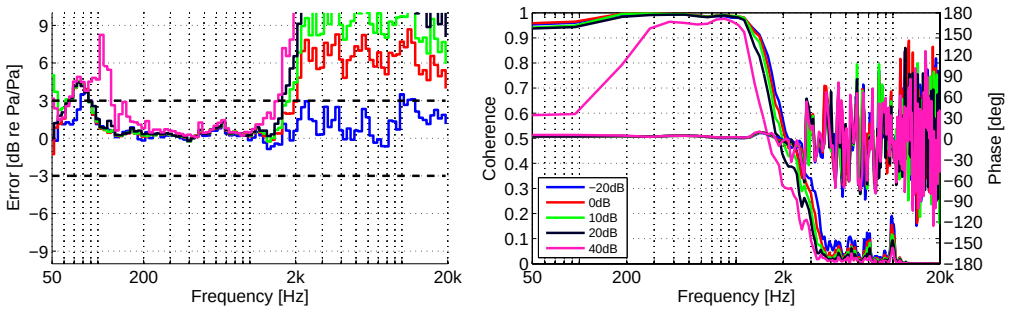
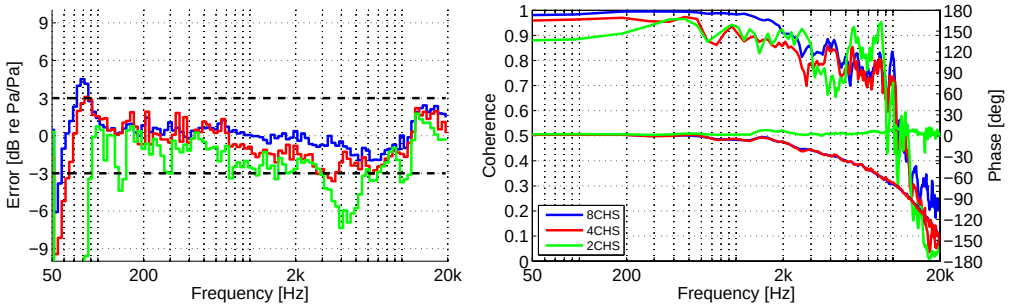


Figure A.2: Maximum error, average coherence and average phase difference between the reproduced- and reference sounds, across the validation microphones positions. Regularization parameters: $\beta = -20$dB, 0dB, 10dB, 20dB and 40dB. Reproduction using 8 loudspeakers of the program materials Pop music, Classical music, and speech. No correction or post-processing applied to the inverse filter.

(a) Error, target positions.

(b) Coherence and phase difference, target positions.



(c) Error, validation positions.

(d) Coherence and phase difference, validation positions.

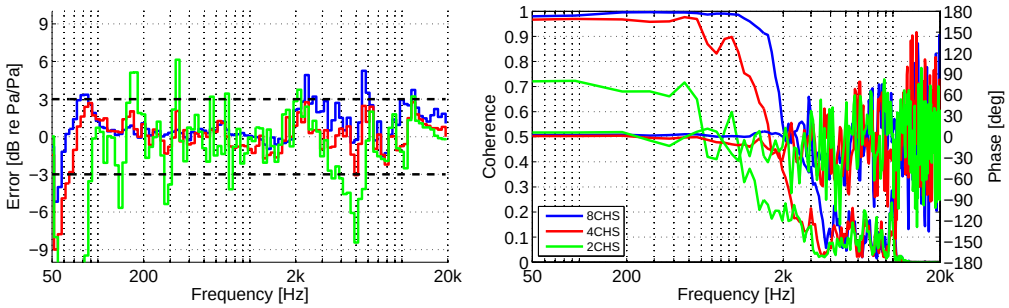


Figure A.3: Maximum error, average coherence and average phase difference between the reproduced- and reference sounds, across the target and the validation microphones positions. Regularization parameter: $\beta = -20$dB. Reproduction using 8, 4 and 2 loudspeakers of the program material Pop music. No correction or post-processing applied to the inverse filter.

(a) Error, $\beta = 0dB$.

(b) Coherence and phase difference, $\beta = 0dB$.

(c) Error, $\beta = 10dB$.

(d) Coherence and phase difference, $\beta = 10dB$.

(e) Error, $\beta = 20dB$.

(f) Coherence and phase difference, $\beta = 20dB$.

(g) Error, $\beta = 40dB$.

(h) Coherence and phase difference, $\beta = 40dB$.

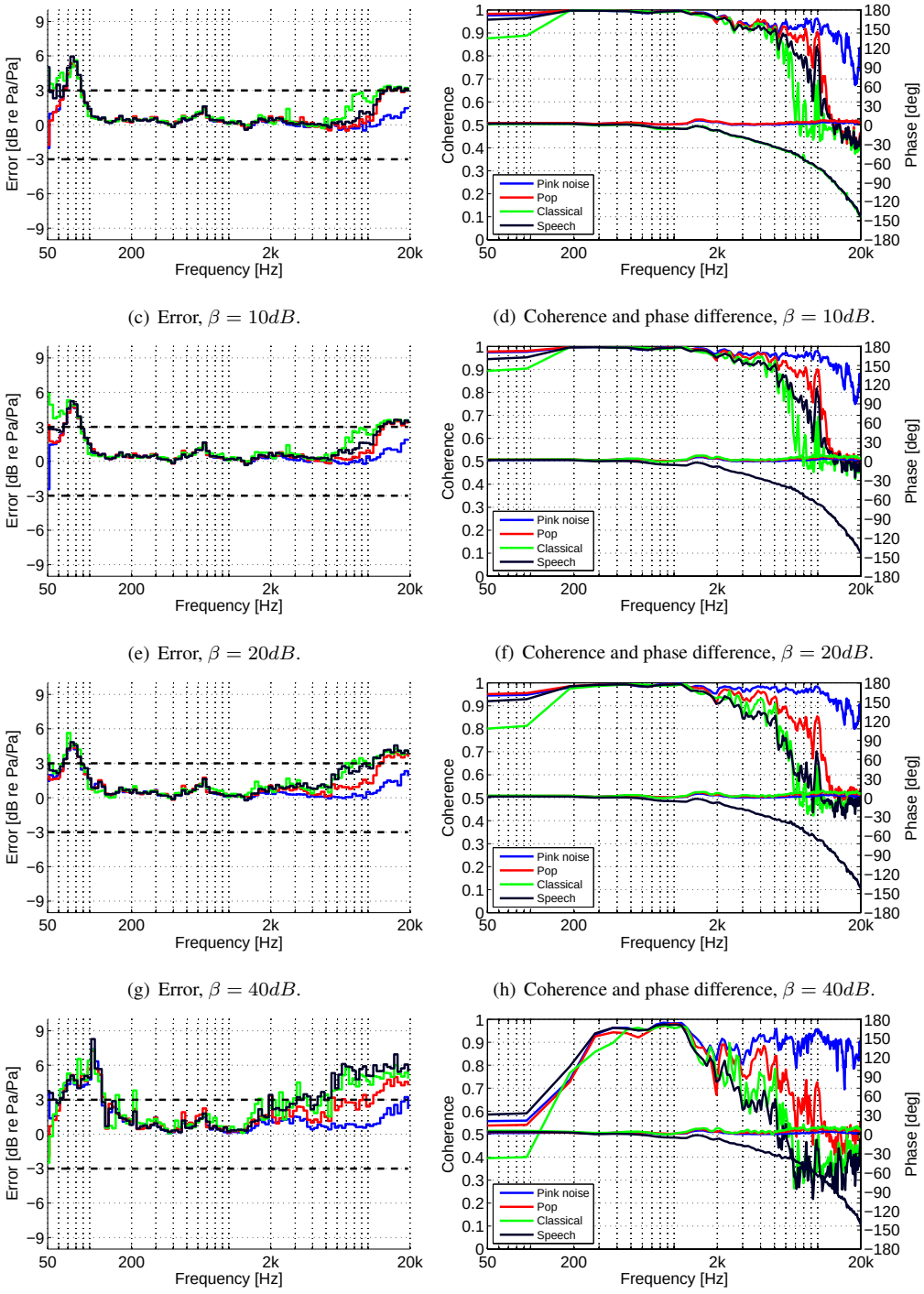


Figure A.4: Maximum error, average coherence and average phase difference between the reproduced- and reference sounds, across the target microphone positions. Regularization parameters: $\beta = 00$dB, 10dB, 20dB and 40dB. Reproduction using 8 loudspeakers of the program materials Pink noise, Pop music, Classical music and Speech. No correction or post-processing applied to the inverse filter.

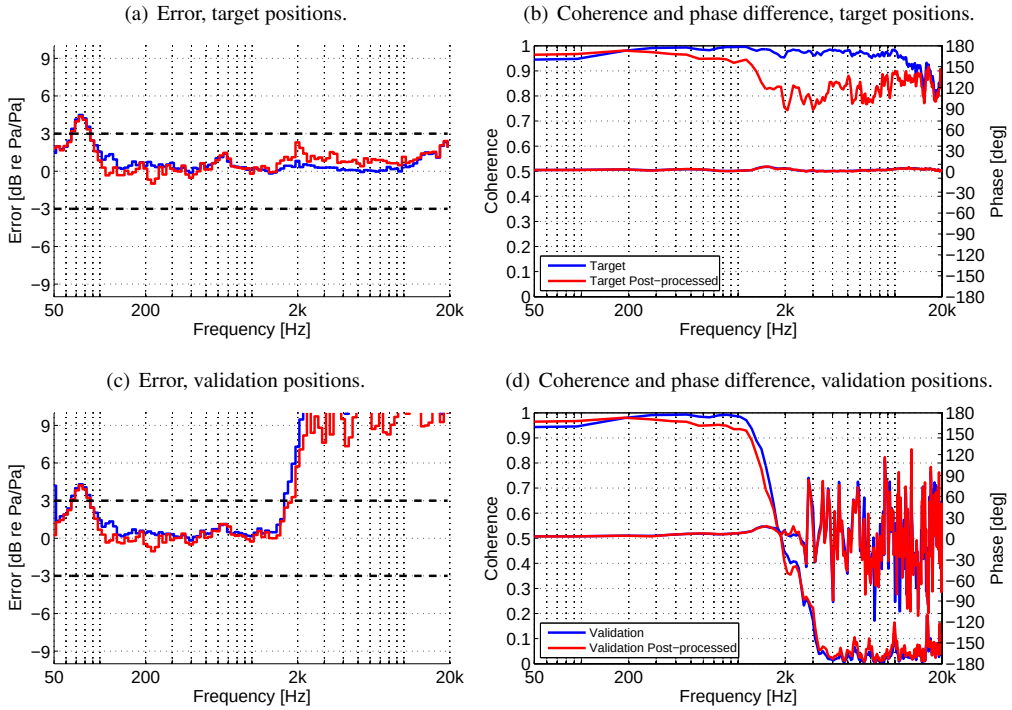


Figure A.5: Maximum error, average coherence and average phase difference between the reproduced- and reference sounds, across the target and the validation microphone positions. Regularization parameter: $\beta = 20$dB. Reproduction using 8 loudspeakers of the program material Pink noise. No correction but post-processing applied to the inverse filter..

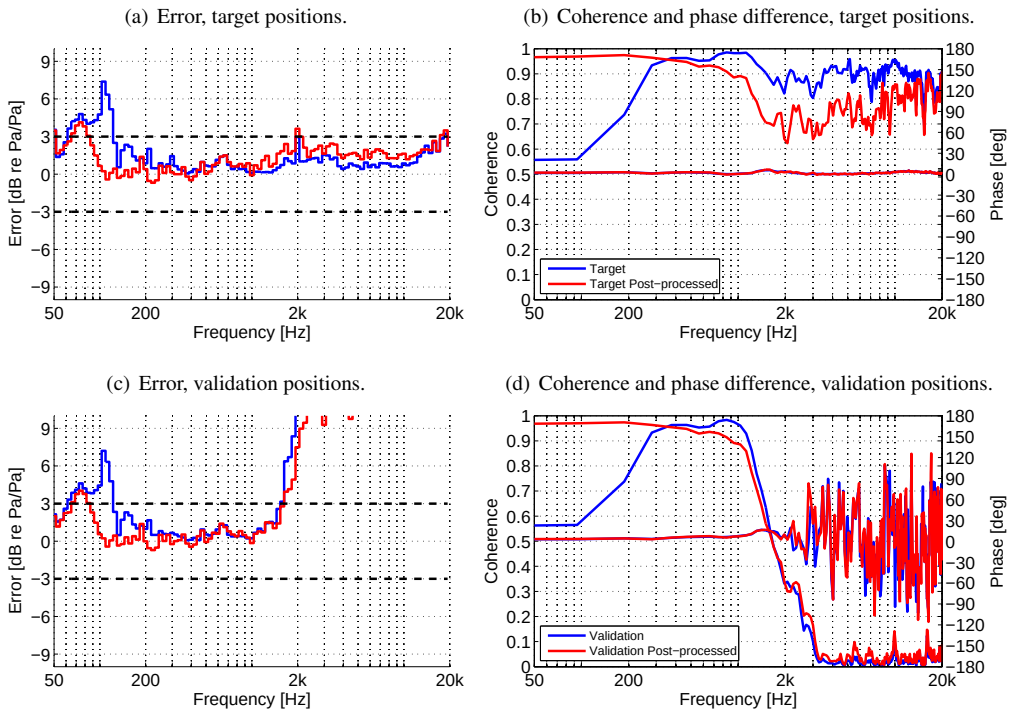


Figure A.6: Maximum error, average coherence and average phase difference between the reproduced- and reference sounds, across the target and the validation microphone positions. Regularization parameter: $\beta = 40$dB. Reproduction using 8 loudspeakers of the program material Pink noise. No correction but post-processing applied to the inverse filter..

APPENDIX **B**

Listening test platform

The platform in which the listening test was programmed and designed is called Pure Data (Pd) in its version Pd-0.43.4-extended. Pd is an open source visual programming language, oriented to developments using sound, video and images. It was chosen due to the versatility for implementing processing and/or control of audio in real time. This property made it easy to implement the MUSHRA test with smooth changes between audio samples. However, the graphical appearance is not very appealing, and it is not straight forward to make a Graphic User Interface (GUI) in it; so, a touch interface was designed in the iPad. An application called TouchOSC by `hexler.net`, allowed easy prototyping of the GUI, and by sending Open Sound Control (OSC) messages through Wi-Fi to Pd, the audio samples (previously recorded) could be changed with zero- or no perceivable latency.

The GUI designed for the iPad, is shown in figure B.1. This figure was given to the subjects as part of the instructions for the test. An excerpt of the instructions letter given to the subjects is shown below:

> Your task is to rate how similar you perceive groups of sounds with respect to a reference.
>
> 1. The tests starts with a welcome screen, you should tap the red button "TAP HERE TO START" to continue.
> 2. Next, you should press the button "ON/OFF" to turn on the sound.
> 3. The progress bar at the left of this button will start moving, showing the position of the song. You will listen to 18 seconds of the sound.
> 4. In the bottom of the screen there are buttons named "A, B, C, D, E, F, G, H, I and REF" the activation of each will enable the reproduction of a different sound.
> 5. In the central part of the screen, there are 9 faders, corresponding to each sound sample in the buttons "A" to "I". After listening to each of the sound samples you should rate how similar you perceive each of them with respect to the sound in "REF".
>
>   a) Be aware that when you have activated any of the buttons "A" to "I", and then the "REF", the previous button will remain active in order to remind you which was the sample you were comparing with, but the sound that you listen to is the one in "REF".
>
> 6. After you have listened to the song 6 times, rated and listened to all the sound samples, the button "NEXT" will appear below the button "ON/OFF". When you are finished evaluating the trial you can continue to the next one.

7. The test consists of two training trials, one break, and three groups of five trials separated by breaks. It is very important that you take theses breaks since your hearing can get tired.

8. If you have any questions or feel uncomfortable, you can interrupt the test at any time.

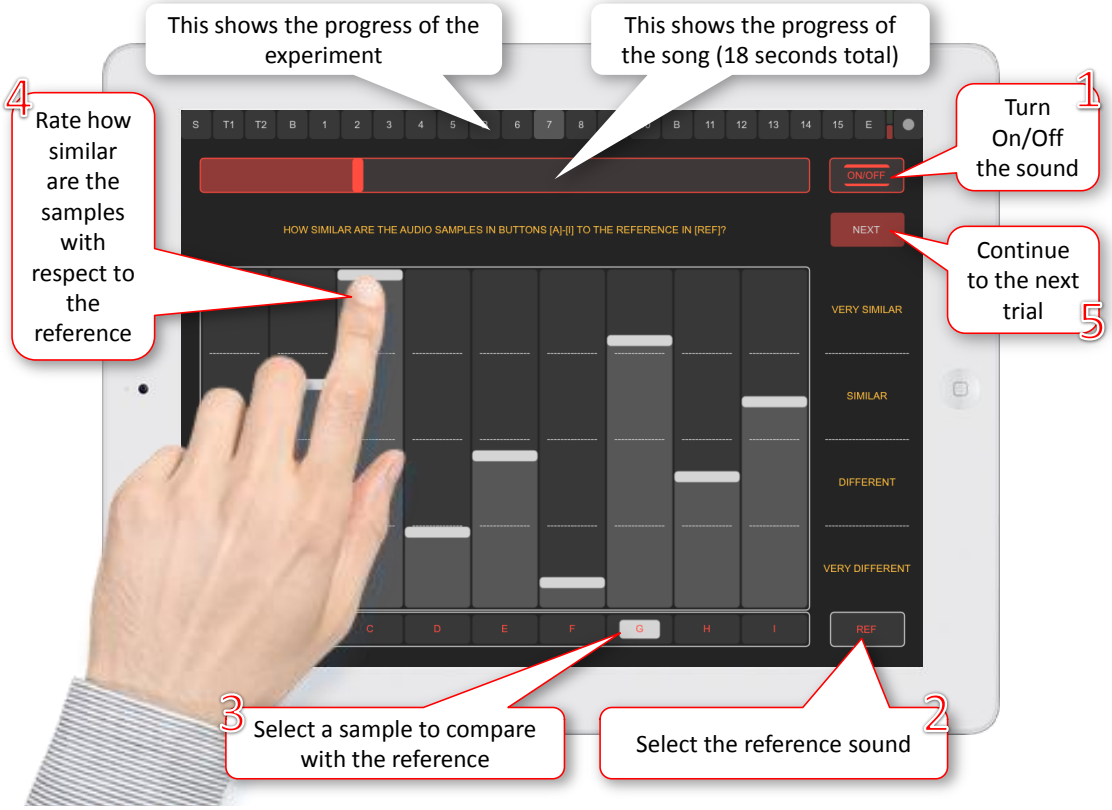


Figure B.1: Interface in a touch screen (iPad) for the listening test.

Bibliography

[1] European Telecommunication Standards Institute. Speech processing, transmission and quality aspects - Part 1: Background noise simulation technique and background noise database. *ETSI EG 202 396-1*, 2:1–58, 2008.

[2] Marton Marschall, Sylvain Favrot, and Jörg Buchholz. Robustness of a mixed-order Ambisonics microphone array for sound field reproduction. *Audio Engineering Society Convention 132*, pages 1–11, 2012.

[3] Ole Kirkeby and Philip A. Nelson. Local sound field reproduction using digital signal processing. *Journal of the Acoustical Society of America*, 100(February):1584–1593, 1996.

[4] Pauli Minnaar, Signe Frø lund Albeck, Christian Stender Simonsen, Boris Sø ndersted, Sebastian Alex Dalgas Oakley, and Jesper Bennedbæ k. Reproducing Real-Life Listening Situations in the Laboratory for Testing Hearing Aids. *Audio Engineering Society Convention 135*, 2013.

[5] Angelo Farina and Emanuele Ugolotti. Spatial Equalization of sound systems in cars. *15th International Conference: Audio, Acoustics & Small Spaces (October 1998)*, 1998.

[6] HEAD Acoustics GmbH. A sound field reproduction method adapted to terminal testing with HATS. *Not published yet*, pages 1–10, 2014.

[7] Filippo M. Fazi and Philip A. Nelson. The ill-conditioning problem in sound field reconstruction. *Audio Engineering Society Convention 123*, 2007.

[8] Scott G. Norcross, Gilbert A. Soulodre, and Michel C. Lavoie. Subjective effects of regularization on inverse filtering. *Audio Engineering Society Convention 114*, 2003.

[9] Scott G. Norcross, Gilbert A. Soulodre, and Michel C. Lavoie. Distortion Audibility in Inverse Filtering. *Audio Engineering Society Convention 117*, 2004.

[10] Scott G. Norcross, Gilbert A. Soulodre, and Michel C. Lavoie. Subjective investigations of inverse filtering. *Journal of the Audio Engineering Society*, 52(10):1003–1028, 2004.

[11] International Telecomunication Union. Recommendation ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems. 2003.

[12] Philip A. Nelson and Stephen J. Elliott. *Active control of sound*. Academic Press, 1993.

[13] Ole Kirkeby, Philip A. Nelson, Felipe Orduna-Bustamante, and Hareo Hamada. Fast deconvolution of multichannel systems using regularization. *IEEE Transactions on Speech and Audio Processing*, 6(2):189–194, March 1998.

[14] Ole Kirkeby, Philip A. Nelson, Per Rubak, and Angelo Farina. Design of cross-talk cancellation networks by using fast deconvolution. *Audio Engineering Society Convention 106*, 1999.

[15] Hironori Tokuno, Ole Kirkeby, Philip A. Nelson, and Hareo Hamada. Inverse filter of sound reproduction systems using regularization. *IEICE Trans. Fundamentals*, E80-A(5):809–820, 1997.

[16] Scott G. Norcross and Martin Bouchard. Multichannel Inverse Filtering with Minimal-Phase Regularization. *Audio Engineering Society Convention 123*, pages 1–8, 2007.

[17] M. R. Schroeder. New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 2005.

[18] Swen Müller and Paulo Massarani. Transfer function measurement with sweeps. *Journal of the Audio Engineering Society*, pages 443–471, 2001.

[19] Ryo Mukai, Shoko Araki, Hiroshi Sawada, and Shoji Makino. Evaluation of separation and dereverberation performance in frequency domain blind source separation. *Acoustical Science and Technology*, 25(2):119–126, 2004.

[20] IEEE Standards Association. IEEE Standard Methods for Measuring Transmission Performance of Analog and Digital Telephone Sets, Handsets, and Headsets. *IEEE 269-2010*, 2010.

[21] Wikipedia. Standard score, 2014.

[22] Bruel & Kjær Sound and Vibration A/S. Product data: Sound source type 4224.