Speaker localization and separation in teleconferences

Thomas Grasser, Martin Rothbucher, Klaus Diepold

Technical Report

Speaker localization and separation in teleconferences

Thomas Grasser, Martin Rothbucher, Klaus Diepold

March 30, 2013

Lehrstuhl für Datenverarbeitung
Technische Universität München

Thomas Grasser, Martin Rothbucher, Klaus Diepold. Speaker localization and separation in teleconferences. . Technische Universität München, LDV. March 30, 2013.

Abstract

At the Institute for Data Processing at the Technical University of Munich, a lot of research effort is spent in the development of a conference phone. In this thesis the focus is on the recording, the speaker localization and the source separation. In the past, several algorithms were tested in the conference scenario, but there was no comparison of the different algorithms with the same test data. Firstly, a suitable microphone array for each of the three treated algorithms was designed and built with a 3D plotter. Afterwards, the impulse-responses for the three microphone arrays were estimated for 0°, 10°, 20° elevation and in 5° steps in azimuth to make a statement about the properties of the different microphone arrays. Furthermore, more than 1000 recordings were made in the audiolab and in the videolab to evaluate the different algorithms. As well as the subjective examination of the auditory impression the SIR, SAR and SDR-values were determined for all the recordings. From the results of the experiments, the best localization and the best separation algorithms were chosen, which will be used in the following processing steps, for example speaker recognition.

Am Lehrstuhl für Datenverarbeitung der TU München wird nun schon seit längerem an der Entwicklung eines Telekonferenzsystems gearbeitet. In dieser Arbeit wird die Aufnahme, die Sprecherlokalisierung und die Quellentrennung näher untersucht. In der Vergangenheit wurden schon mehrere Ansätze auf ihre Tauglichkeit im Telekonferenzszenario untersucht, jedoch wurde noch nie ein Vergleich der verschiedenen Ansätze mit den gleichen Testdaten durchgeführt. Am Anfang dieser Arbeit wurde für die drei zu untersuchenden Ansäzte jeweils ein passendes Mikrophonarray entworfen und mit einem 3D Plotter gedruckt. Anschließend wurden für diese drei Mikrophonarrays die Kanalimpulsworten für vierschiedene Winkel bestimmt, um Aussagen über die Eigenschaften der Arrays machen zu können. Desweiteren wurden mehr als 1000 Aufnahmen im Audiolabor und im Videolabor zur Evaluierung der verschiedenen Ansätze angefertigt. Neben der Überprüfung des subjektiven Höreindrucks wurden für sämtliche Aufnahmen die SIR, SAR und SDR-Werte bestimmt. Anhand der Versuchsauswertungen wurde der am besten geeignete Lokalisierungs- und Trennalgorithmus ausgewählt, das für die weitere Verarbeitung z.B. die Sprechererkennung verwendet werden soll.

Contents

1. Introduction

1.1. Motivation for this thesis

In the era of globalization, it is important for many companies to have locations all over the world and to take advantage of the various sites or to sell in as many countries as possible their own products. The affected companies must find solutions across their multiple sites for diverse problems. In order to inform another location about a task, it is sufficient to write an email, but to jointly solve a problem together you need a more flexible communication channel. Probably the best way to communicate is to hold a meeting, but that is very time consuming and expensive. Since the digitization and prevalence of the internet, even conference calls are possible. The existing conference phones record either a mixture of the individual speakers or each speaker has its own microphone. If a conference participant who is via a phone, listening to a mixture of different people speaking simultaneously, it can be difficult for him to follow the conversation. This is the so-called “cocktail party effect” [3]. If several people are talking at the same time in one room, the listener can concentrate on one speaking person on the basis of directional perception. When each speaker is on a separate channel, one can give each speaking person a direction with Head Related Transfer Functions (HRTFs), on the receiving side in order to solve the “cocktail party problem”. For this reason, it would be nice to have a conference phone that does not need an individual microphone for each speaker, but one which can separate the individual speakers, from the mixture recorded by the microphones within the conference phone.

1.2. Objectives and contents of this thesis

At the Institute for Data Processing a lot of research effort is spent on an innovative conference phone [17]. Many students have written their theses on this topic [7, 11, 18, 14]. This work is concerned with the comparison of the existing approaches, which were all tested but not also with the same test data. The content of this thesis is to find that algorithm, which delivers the best localization and separation results, under the same conditions. The aim is not to find the algorithm, which delivers the best evaluation results, but we are searching for the best compromise, which descripes the teleconference scenario optimally.

1.3. System overview

In Figure 1.1 one can see the content of this thesis at a glance. The sound sources are recorded by a microphone array with eight microphones. Subsequently, the eight sound streams can be processed in various manners, in order to get the seperated sources.

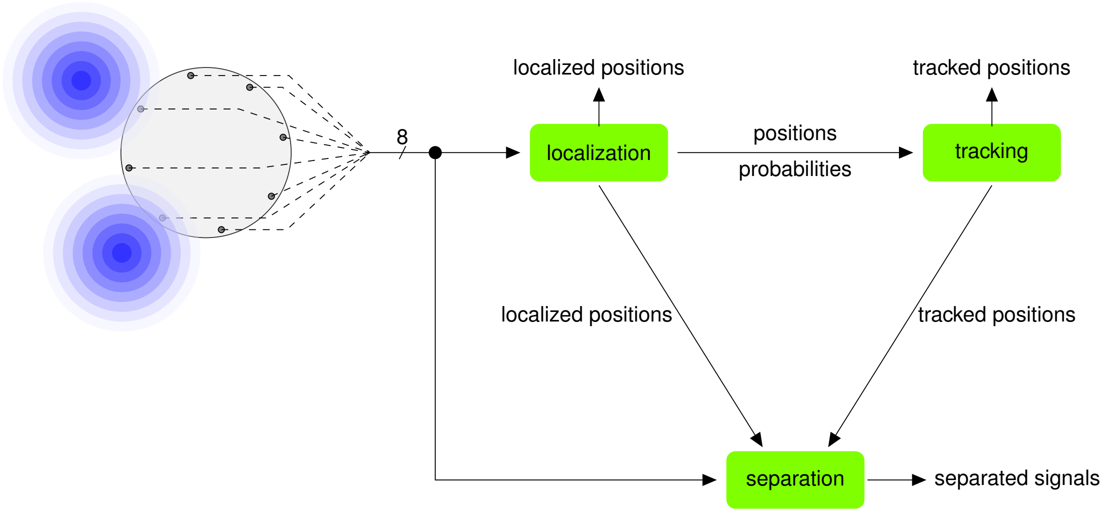


Figure 1.1.: System overview

Localization A localization algorithm should find the coordinates of the source positions from the eight sound streams from the microphone array. In both localization approaches the radius $1.3\mathrm{m}$ is known as meaning that the algorithms have to find the azimuth and the elevation angles.

Tracking The job of the tracking part of the algorithms is normally to follow moving sources. In this thesis, we want to find out how well the localization and the separation, in a simple conference scenario where all speakers are sitting around a table, can work. Nevertheless, the tracking part solves the permutation problems of the localizer, which means that the localizer can not always match the located positions with the corresponding source.

Separation The function of this part is to separate the mixed recordings, so that one receives a own channel, for each source. As shown in Figure 1.1, there are three possibilities to receive separated sources. In the case of Blind Source Separation (BSS), the separation is done without preprocessing directly from the recordings, but some separators need the positions of the sources to operate as well.

The following algorithms are treated in this thesis.

- *SRP-PHAT*: This program delivers the positions of the sound sources. This code also includes the particle filter, which should compensate localization errors or permutation problems. [localization and tracking]

- *GSS*: This software performs a geometric-source-separation and uses therefore the previous located positions. [separation after localization and tracking]
- *COMPaSS*: The COMPaSS-algorithm localizes the sources with the assistance of previously determined Transfer Functions (TFs). The results are the indexes of the suitable TFs. To solve the permutation problem, a particle filter is also implemented. [localization and tracking]
- *Binary Masking*: Binary Masking requires the source positions and the appropriate TFs for the separation process. [separation after localization and tracking]
- *IVA*: IVA belongs to Blind Source Separation (BSS). Hence, it requires no information about the positions of the speaker. [only separation]

2. Previous work

Many students have written their thesis in connection with telephone conferences at the Institute for Data Processing of the Technical University of Munich. For this reason, there are many existing software implementations and related hardware. In this chapter the existing work is presented assorted by the authors and not by localization or separation algorithm. The various approaches are explained briefly in this chapter, for further informations, please look in the references.

2.1. Universal relevant definitons

First, some basic definitons are made which are relevant in all localization and tracking approaches. We have N different, time-dependent *source signals*

$$\mathbf{s}_i(t) = (s_1(t), s_2(t), \ldots, s_N(t))^T, \tag{2.1}$$

M *microphone sinals*

$$\mathbf{x}_j(t) = (x_1(t), x_2(t), \ldots, x_M(t))^T \tag{2.2}$$

and for each microphone channel additional noise

$$\mathbf{n}_j(t) = (n_1(t), n_2(t), \ldots, n_M(t))^T, \tag{2.3}$$

where i is the index for a source signal and j is the index for a microphon channel. The instantaneous mixture model

$$x_j(t) = \sum_{i=1}^{N} a_{ji} \cdot s_i(t) + n_j(t) \tag{2.4}$$

or in matrix vector notation

$$\mathbf{x}_j(t) = \mathbf{A}_{ji} \cdot \mathbf{s}_i(t) + \mathbf{n}_j(t), \tag{2.5}$$

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_M(t) \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1N} \\ a_{21} & a_{22} & \ldots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \ldots & a_{MN} \end{pmatrix} \cdot \begin{pmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_N(t) \end{pmatrix} + \begin{pmatrix} n_1(t) \\ n_2(t) \\ \vdots \\ n_M(t) \end{pmatrix} \tag{2.6}$$

assumes that all source signals arrive at the same time, at the microphones and the differences between the signals are weighting factors and noise factors.

The convolutive mixture model also takes the TF (speed of sound, spectral effects) from the source, to the microphone into account and therefore describes the behaviour of sounds better. The formula for the convolutive mixture model is

$$\mathbf{x}_j(t) = \mathbf{A}_{ji}(t) * \mathbf{s}_i(t) + \mathbf{n}_j(t) \tag{2.7}$$

and can be written in detail as

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_M(t) \end{pmatrix} = \begin{pmatrix} a_{11}(t) & a_{12}(t) & \dots & a_{1N}(t) \\ a_{21}(t) & a_{22}(t) & \dots & a_{2N}(t) \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1}(t) & a_{M2}(t) & \dots & a_{MN}(t) \end{pmatrix} * \begin{pmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_N(t) \end{pmatrix} + \begin{pmatrix} n_1(t) \\ n_2(t) \\ \vdots \\ n_M(t) \end{pmatrix}. \tag{2.8}$$

Because the existing algorithms work in frequency domain, the signals must be windowed and transformed. The windowing is done with $L = 1024$ samples and with 50% overlapping as shown in Figure 2.1 with cosine windows. The drawback of the frequency domain is the processing in frames of 1024 samples. Regardless of the calculating time the algorithms have a timedelay of $t = 1024/48\text{kHz} \approx 0.02\text{s}$. Because GSS and binary masking need a localization, the algorithmic delay is $2 * 1024 = 2048$ samples.

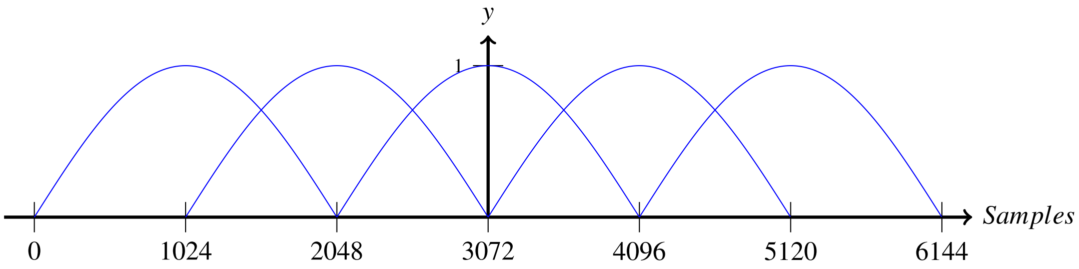


Figure 2.1.: Cosine window function with 50% overlapping

The recorded mixture for M microphones and for each frequency bin f after the transformation is described by

$$\mathbf{x}_j^f = (x_1^f, x_2^f, ..., x_M^f)^T. \tag{2.9}$$

In Figure 2.2 one can see the possible paths for receiving the separated signals from the recorded mixtures.

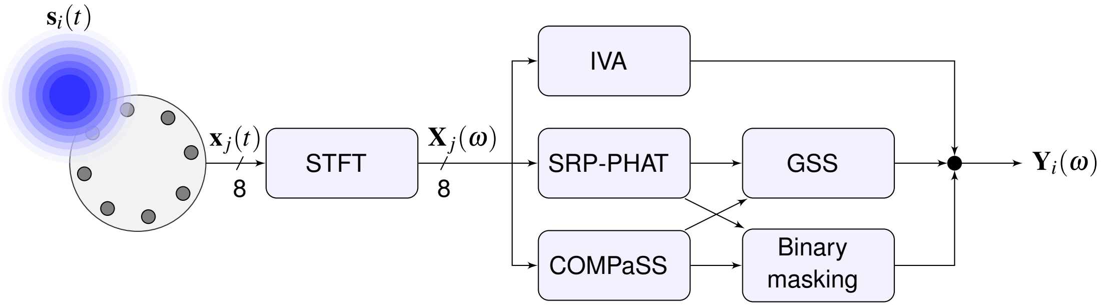


Figure 2.2.: Flow chart of the treated algorithms

2.2. Blind Source Separation for Speaker Recognition Systems

Here only the Blind Source Separation (BSS) part from Michael Unverdorben's thesis [18] is treated. BSS means the separation of different sources without prior knowledge of the sources, the mixing process and the position of the microphones. The following performed separation is IVA and is only based on the supposition that the different sources are statistically independent.

2.2.1. Independent Component Analysis (ICA)

ICA [10] is the basic demixing algorithm of IVA. It works with the instantaneous mixing model. From this point on, the noise will not be considered for simplicity. The aim of the independent component analysis is the extraction of independent components in $\mathbf{s}$. To distinguish the source signal from the mixed microphone signals, one needs a demixing matrix $\mathbf{W}$, which is in the ideal case the inverse of $\mathbf{A}$,

$$\mathbf{W} = \mathbf{A}^{-1}. \tag{2.10}$$

The equation

$$\mathbf{y}_j(t) = \mathbf{W} \cdot \mathbf{x}(t) \tag{2.11}$$

provides the estimated value for the source signals. $\mathbf{y}$ should preferably be an exact estimation of $\mathbf{s}$. The problem in this case, is that we have no information about $\mathbf{A}$ and $\mathbf{s}$. If we were to know the mixing matrix $\mathbf{A}$, we could solve the problem simply through their inversion. For simplification three assumptions via $\mathbf{s}$ are made:

- The independent components in $\mathbf{s}$ (and thus in $\mathbf{x}$) are free from averaging, if not, then $\mathbf{x}$ can be easily centered by subtracting the mean value.
- The variance of each component in $\mathbf{s}$ is one. Normalizing of the variance to one only causes a change in the coefficients in A, wherein the modeling is still valid.
- The solution to the problem is that the individual components of $\mathbf{s}$ are statistically independent.

Therefore, we have a standardized random variable. The first step of ICA is to project the values of $\mathbf{x}$ on a different basis, so that the new resulting components $\mathbf{u}$ are then statistically independent.

$$\mathbf{u} = \mathbf{W} \cdot \mathbf{x} \tag{2.12}$$

$\mathbf{W}$ must be chosen so that the individual components in $\mathbf{u}$ are statistically independent in pairs. The single base vectors of the base change, are the rows of $\mathbf{W}$.

$$\mathbf{W} = \begin{pmatrix} w_1^T \\ w_2^T \\ \vdots \\ w_n^T \end{pmatrix} \tag{2.13}$$

For finding the matrix $\mathbf{W}$ there are various approaches, including PCA, which is explained in chapter A.2.

2.2.2. ICA for audio signals

As ICA only works for instantaneous mixtures, it can not be directly applied to audio signals, which are based on the convolutive mixture model. The solution to this problem is the transformation in the frequency domain:

$$\begin{gathered} \mathbf{x}_j(t) = \mathbf{A}_{ji}(t) * \mathbf{s}_i(t) \\ \begin{array}{c} \circ \\ | \\ \bullet \end{array} \\ \mathbf{X}(\omega) = \mathbf{A} \cdot \mathbf{S}(\omega). \end{gathered} \tag{2.14}$$

To get the estimated source signal

$$\mathbf{Y}(\omega) = \mathbf{W}(\omega) \cdot \mathbf{X}(\omega) \tag{2.15}$$

one could now actually apply ICA, but there is still one problem. ICA works only on stationary[1] frames and speech is non-stationary. The solution to this problem is to divide the source signal, before the Short-Time-Fourier-Transformation (STFT) into short frames, which are stationary. This dissection is performed by a window function.

After the windowing of the signal, the Discrete-Fourier-Transformation (DFT) can be performed, so that we get a frequency representation for each windowed frames.

From the transformation into the frequency domain, we now have for each frequency bin, the separation problem

$$\mathbf{y}^f = \mathbf{W}^f \cdot \mathbf{x}^f, \tag{2.16}$$

which leads us to the next problem, the permutation problem. To solve this problem IVA is presented in the next chapter.

2.2.3. Independent Vector Analysis (IVA)

As ICA is performed in every frequency bin of each frame, one distinguishes the separated channels for each frequency bin, but the separated channels are interchanged from fequency bin to frequency bin. The expansion of ICA to IVA [8, 4] is the solution and the main part of IVA and can be summarized as follows:

- The components from different sources within a frequency bin are independent of each other.
- The components from the same sources across all frequency bins are dependent on each other.

Figure 2.3 shows the schematic structure of IVA of a 2 x 2 mixture with the fragmentation of each source $\mathbf{s}_i = (s_i^1, s_i^2, \ddots s_i^F)^T$ and microphone signal $\mathbf{x}_j = (x_j^1, x_j^2, \ddots x_j^F)^T$ in freqeuncy bins from 1 to F.

[1] Stationary means one can calculate an arithmetic mean value and make a statement about the deviations from the mean value, which are not dependent on the time.

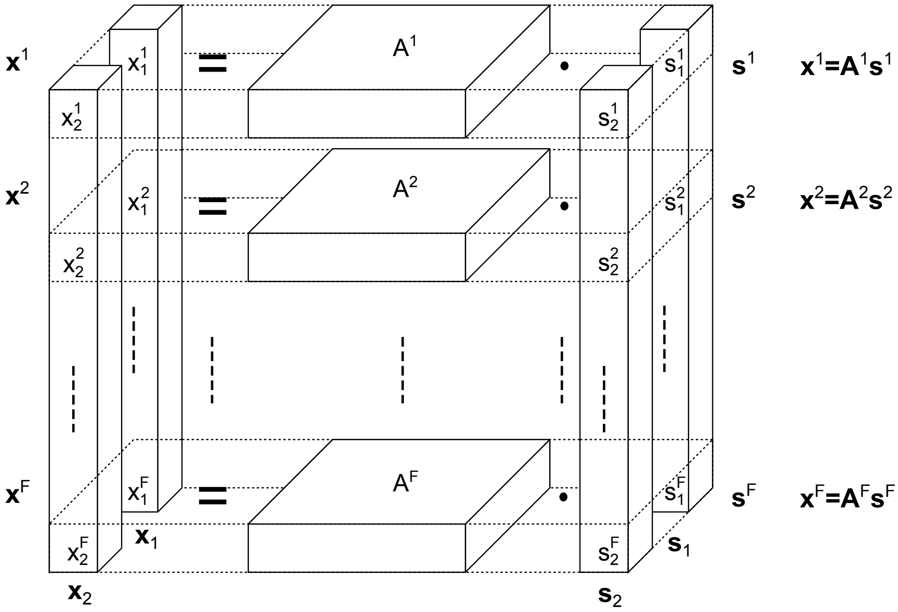


Figure 2.3.: IVA-mixture for two sources and two microphones [18]

2.3. Extension of a binaural localization and separation Algorithm

The dissertation of Dr.-Ing. Marko Durcković [5] is concerned with sound localization, tracking and separation in the binaural case of a robot. The localization and the tracking was transferred by Tobias Plutka [14] to the teleconference scenario with a microphone arrray consisting of eight microphones. The localization is performed by a modification of the Cross Convolution Localization (CCL)-algorithm, the tracking by a particle filter and binary masking is implemented for source separation.

2.3.1. COMPaSS (A modification of the CCL-algorithm)

The problem of the original CCL-algorithm is that it only works in the single source case, because multiple active sound source signals would overlap in time domain. The solution is to transform the recorded signals in a domain where the signals have a sparse representation and do not overlap. The typical methods to find a new base in statistical signal processing, for example Principle Component Analysis (PCA) are deliberatly not used because they need samples from the data to create the new base. In [16] one can check that a windowed (Hann, Hamming or triangular window shape of 1024 sampels) Fourier transformation provides a sparse representation of speech. The main idea of the CCL-algorithm is to find the suitable pair of TFs from a previously recorded database. If the index from the proper TFs is noted, one can look in the database for the corresponding direction. Mathematically one has to deconvolve the recorded signals $\mathbf{x}_1(t)$ und $\mathbf{x}_2(t)$ with each filter pair η (η denotes the TF index in the database). If the right filter pair is used, the resulting signals are equal

$$\hat{\mathbf{s}}_{1,\eta}(t) = \hat{\mathbf{s}}_{2,\eta}(t) = \mathbf{s}(t), \quad \Longleftrightarrow \quad \eta = \eta_0. \tag{2.17}$$

The deconvolution in the frequency domain becomes instable if any of the filter entries is close to zero. For this reason, the CCL-algorithm [19] suggests to convolve each observation with the opposite TF, so that one gets two new signals

$$\tilde{\mathbf{s}}_{1,\eta}(t) = \mathbf{h}_{2,\eta}(t) * \mathbf{x}_1(t) \tag{2.18}$$

and

$$\tilde{\mathbf{s}}_{2,\eta}(t) = \mathbf{h}_{1,\eta}(t) * \mathbf{x}_2(t). \tag{2.19}$$

At this algorithm the letter $\mathbf{h}$ is used instead of letter $\mathbf{a}$, because this is the typical letter for TFs. Due to the associative law of the convolution, one can calculate $\tilde{\mathbf{s}}_{2,\eta}(t)$ from $\tilde{\mathbf{s}}_{1,\eta}(t)$ if the right filter pair is used:

$$\begin{aligned} \tilde{\mathbf{s}}_{1,\eta}(t) &= \mathbf{h}_{2,\eta}(t) * \mathbf{x}_1(t) \\ &= \mathbf{h}_{2,\eta}(t) * \mathbf{h}_{1,\eta_0}(t) * \mathbf{s}(t) \\ &= \mathbf{h}_{1,\eta}(t) * \mathbf{h}_{2,\eta_0}(t) * \mathbf{s}(t) \\ &= \mathbf{h}_{1,\eta}(t) * \mathbf{x}_2(t) \\ &= \tilde{\mathbf{s}}_{2,\eta}(t) \quad \Longleftrightarrow \quad \eta = \eta_0. \end{aligned} \tag{2.20}$$

This theoretical approach has to be adapted to the real environment. Real audio recordings contain noise and distortions from the hardware used. For this reason, one does not receive ever the same signals, even though the suitable filter pair is used. The solution for this problem is a maximization of a similarity measure over all η. The similarity of the cross-convolved signals is calculated with a cross-correlation function

$$(\tilde{\mathbf{s}}_{1,\eta}(t) \star \tilde{\mathbf{s}}_{2,\eta}(t)) := \int_{-\infty}^{\infty} \tilde{\mathbf{s}}_{1,\theta,\varphi}(\tau)\tilde{\mathbf{s}}_{1,\theta,\varphi}(\tau)\,d\tau \tag{2.21}$$

at time delay zero, where $\star$ means the crosscorrelation operator. Afterwards, the maximum of the similarity measure

$$\hat{\eta} = \arg\max_{\eta} \tilde{\mathbf{s}}_{1,\eta}(t) \star \tilde{\mathbf{s}}_{2,\eta}(t) \tag{2.22}$$

is determined. After equation 2.20, there are differences between the CCL-algorithm and the loCalization Of MultiPle Sound Sources (COMPaSS)-algorithm [6]. At the COMPaSS-algorithm the similarity measurement is done in Fourier domain for each frequency bin f of each frame separately as one can read in [5]. The winner TF in every single frequency bin, gives a point in the histogram. Before the algorithm starts the maximum number of sources must be set, so that the algorithm can choose as many TFs as the maximum number of sources. For each chosen TF, the probability is calculated depending on the points in the histogram. Not all sources are active over the whole time, so the algorithm has to decide which sources are currently active. This is done by setting a threshold depending on the existing amount of noise. The extension from the binaural case in [5] to the teleconference is a very simple step. The histogram is not built over one microphone pair, but it is built over all four microphone pairs [1 5; 2 6; 3 7; 4 8] to get better results. The next problem is that the chosen TFs are interchanged. The assignment is done by the particle filter.

2.3.2. Particle filter

The output of the COMPaSS-algorithm gives noisy informations about the positions of the sources, because the localization is done for each sound frame (here: 1024 samples) individually. So it can happen that a actual active source is not detected because the other sources are too dominant. Furthermore, this algorithm can not match the localized positions to the right sources and is also not able to follow moving sources. The problems are solved by a particle filter simillary as in [20]. Particle filtering, also known as Sequential Monte Carlo simulations, is appropriate for nonlinear and non-Gaussian Bayesian tracking [2]. A particle filter has a state model and some measurements. It tries to estimate the posterior status, at each time step k, with the informations from all measurements to k, it is the so called prediction step. At the next time step $k+1$, the status for time step $k+1$ is updated with the measurements from time step $k+1$ and a prediction for time step $k+2$ is done. A particle filter consists of N weighted particles, which represent the probability distribution. Every sound source is represented by a set of particles. For further informations, please look in [2].

2.3.3. Binary masking

The difference to most of the other separation algorithms is that binary masking also operates in the underdetermined case. This means that the number of sources can be higher than the number of recorded sound mixtures. In [5] binary masking is used for source separation, because of the binaural case of a robot. This separation algorithm performs also in Fourier domain, where human speech is sparse. A very famous representative of binary masking is the Degenerate Unmixing Estimation Technique (DUET)-algorithm [15]. Binary masking starts from the premise that the source signals dominate different frequency bins in the recordings. A binary matrix is established which marks the dominant sound source in each frequency bin. Afterwards, the sources can be separated from both two sound streams by partitioning the time-frequency representation. This separation is a simple multiplication of one channel of the binary matrix with the time-frequency representation of one channel and then transform back into time domain. Better results are available from the channel which is nearer to the sound source.

2.4. Sound Localization and Separation for Teleconferencing Systems

The thesis of Johannes Feldmaier [7] contains a source localization with beamforming and subsequent particle filtering. Furthermore, a geometric source separation is performed.

2.4.1. Acoustic Beamforming

Acoustic Beamforming means the recording of a sound event at a location, where one can not place a microphone. Instead of one microphone a whole microphone array is used, from which the single signals are combined in such a way that one gets constructive interference for the desired direction and destructive interference for all other directions. The combination of the single signals is done by computing the time delays from the point of interest to each microphone. The dime delay

$$t_{\text{delay}} = \frac{s}{c_{\text{air}}} \tag{2.23}$$

can be calculated with the geometrical distance s and the speed of sound in air

$$c_{\text{air}} = (331.3 + (0.606°C^{-1} \cdot \theta))\frac{m}{s}, \tag{2.24}$$

where θ is the temperature in degrees Celsius. Here the Steered Response Power - Phase Transform (SRP-PHAT) [20] is applied because of its robustness towards noise and room effects. This algorithm belongs to the filter-and-sum beamformer and performs the beamforming in the frequency domain, because the energy calculation is more efficient and the whitening is easier. For this purpose the signal is windowed (Hamming-window, L=1024 samples, 50% overlapping). For each microphone pair of the microphone array the time delay for each point of the search region is calculated. For eight microphones and a search region of a halfsphere with 1861 points, the computer has to calculate 52108 delays ($delays = M \cdot (M-1)/2 \cdot points = 8 \cdot 7/2 \cdot 1861 = 52108$). The whitened cross-correlation between microphone pair j' and j'' is calculated by

$$\mathbf{R}_{j'j''}(\tau) \approx \sum_{k=1}^{L-1} \frac{\mathbf{X}_{j'}(k)\mathbf{X}_{j''}(k)^*}{||\mathbf{X}_{j'}(k)||\,||\mathbf{X}_{j''}(k)^*||} e^{j2\pi k\tau/L} \tag{2.25}$$

where $\mathbf{X}(k)$ is one frame with length L in the frequency domain and * denotes the complex conjugate. With formula (2.25) an energy map is created with peaks for the source positions. The informations from the SRP-PHAT are forwarded to the particle filter, which uses those for the update step. As basis implementation the ManyEars [20] [21] algorithm of the University of Sherbrooke (Canada) is used.

2.4.2. Particle Filter

The particle filter after the beamforming has the same task as the particle filter after the COMPaSS-algorithm (please look at 2.3.2).

2.4.3. Geometric Source Separation

Geometric Source Separation (GSS) [13] combines Blind Source Separation (BSS) with acoustic beamforming. A cross power spectral minimization is done provided that all sources are localized in space before. From putting equation (2.14) in equation (2.15) one receives

$$\tilde{\mathbf{S}}(\omega) = \mathbf{W}(\omega)\mathbf{A}(\omega)\mathbf{S}(\omega) = \mathbf{P}\mathbf{Q}(\omega)\mathbf{S}(\omega), \tag{2.26}$$

where $\mathbf{P}$ is a arbitrary permutation matrix and $\mathbf{Q}(\omega)$ is a scaling matrix per frequency. The minimization of the correlation between each channel can be done by diagonalization of

$$\mathbf{R_{yy}}(t,\tau) = E[\mathbf{y}(t)\mathbf{y}^H(t+\tau)]. \tag{2.27}$$

To directly estimate the separation matrix $\mathbf{W}(\omega)$ the two constraint

$$\mathbf{R_{yy}}(t,\tau) - diag[\mathbf{R_{yy}}(t,\tau)] = 0, \tag{2.28}$$

which includes the minimization problem of (2.27) and

$$\mathbf{W}(\omega)\mathbf{A}(\omega) = \mathbf{I}, \tag{2.29}$$

which contains the geometric part are to strong together, but they can be used as cost functions. The cost function

$$J_1(\mathbf{W}(\omega)) = ||\mathbf{R_{yy}}(t,\tau) - diag[\mathbf{R_{yy}}(t,\tau)||^2 \tag{2.30}$$

expresses the crass-talk minimazation of the ouput signals and the second cost function

$$J_2(\mathbf{W}(\omega)) = ||\mathbf{W}(\omega)\mathbf{A}(\omega) - \mathbf{I}||^2 \tag{2.31}$$

contains the geometric information. The matrix norm is defined as $||\mathbf{M}||^2 = [\mathbf{M}\mathbf{M}^H]$. $\mathbf{A}(\omega)$ includes the estimated linear transfer function between the sources and the microphones on the basis of the results of the localizer. The difference between the original GSS-algorithm [13] and the here used algorithm [21] is the instantaneous estimation of the correlation matrices

$$\mathbf{R_{xx}}(t,\tau) = \mathbf{x}(t,\tau)\mathbf{x}(t,\tau)^H \tag{2.32}$$

and

$$\mathbf{R_{yy}}(t,\tau) = \mathbf{y}(t,\tau)\mathbf{y}(t,\tau)^H \tag{2.33}$$

instead of the estimation on several data. The gradients of the cost functions

$$\frac{\delta J_1(\mathbf{W}(\omega))}{\delta \mathbf{W}^*(\omega)} = 4\mathbf{E}(\omega)\mathbf{W}(\omega)\mathbf{R_{xx}}(t,\tau) = 4[\mathbf{E}(\omega)\mathbf{W}(\omega)\mathbf{x}(t,\tau)]\mathbf{x}(t,\tau)^H \tag{2.34}$$

and

$$\frac{\delta J_2(\mathbf{W}(\omega))}{\delta \mathbf{W}^*(\omega)} = 2[\mathbf{W}(\omega)\mathbf{A}(\omega) - \mathbf{I}]\mathbf{A}(\omega) \tag{2.35}$$

are calculated with respect to $\mathbf{W}(\omega)$ and where $\mathbf{E}(\omega) = \mathbf{R}_{\mathbf{yy}}(t,\tau) - diag[\mathbf{R}_{\mathbf{yy}}(t,\tau)$. With the both gradients of the cost functions (2.34) and (2.35) the separation matrix

$$\mathbf{W}^{n+1}(\omega) = \mathbf{W}^{n}(\omega) - \mu[\alpha(\omega)\frac{\delta J_1(\mathbf{W}(\omega))}{\delta \mathbf{W}^*(\omega)} + \frac{\delta J_1(\mathbf{W}(\omega))}{\delta \mathbf{W}^*(\omega)}] \tag{2.36}$$

is calculated, where μ is the adaptation rate and

$$\alpha(\omega) = [\|\mathbf{x}(t,\tau)\|^2]^-2 \tag{2.37}$$

is a normalization factor. With the separation matrix one can compute the estimated source signal with the help of equation (2.15).

3. The new microphone arrays

Due to the drawbacks of the existing microphone array a new one was built with a 3D-Plotter (look at Figure 3.1, `http://www.reprap.org`). The material for the 3D-printer is Acrylnitril-Butadien-Styrol [1], which is on a roll. The individual components were designed with the open source software *Google Sketchup 8.0.16845*. With the free license of *Google Sketchup* STL-files can not be exported, so one needs a suitable plugin for this. The exported STL-files were opened with the next open source software *netfabb Studio Basic 4.9.4*. This software shows the problematic areas of the single components in red and the functioning components in green. This software also has a repair-function, which can repair “non-waterproof" bodies. The complexity in designing the microphone arrays is that the 3D-Plotter can not build overhangs. For this reason, the components may have no overhangs. Additionally, the dimensions of the individual parts must be smaller than $150mm\ x\ 150mm\ x\ 150mm$. The images of the models are in the appendix A.14. In both arrays microphones of the company *Cui* and Phantom Power Adapter of the company *IMG Stage Line* (A.9) are used.

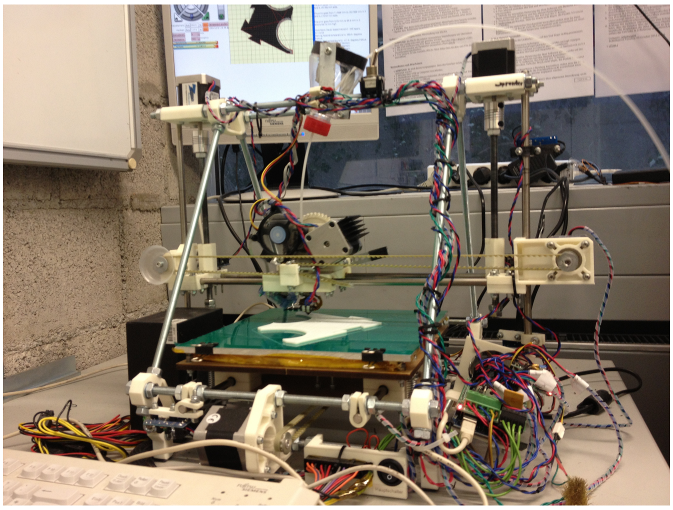


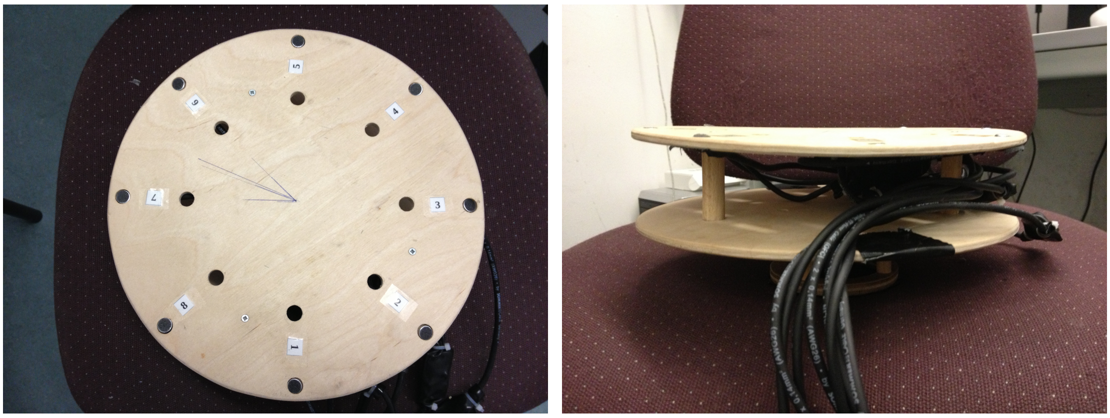
Figure 3.1.: 3D-Plotter at the institute for Data Processing

3.1. The old microphone array in a new form

The array in Figure 3.4 is a replication of the array in Figure 3.2 which is developed in [7]. Because of the planar order of the microphones, the sound sources are in the line of sight of

[1] *Acrylnitrile-Butadiene-Styrene* (ABS): $(C_8H_8C_4H_6C_3H_3N)_n$, melting temperature: $220-250°C$

the microphones and that is exactly suitable for beamforming, which is performed in [7]. The old array is made of wood and now it is a little warped. The holes, in which the microphones sit, are a bit too large, so that the microphones do not sit tightly and they can slip. Therefore, a new array with the same measurements which one can see in Figure 3.3 was created with the 3D-Plotter.


(a) Top view on the old microphone array

(b) Sideview on the old microphone array

Figure 3.2.: The old microphone array

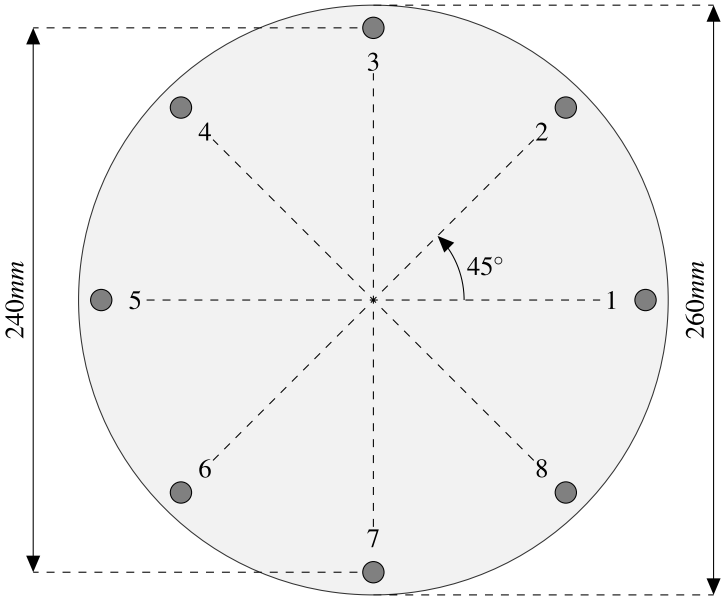
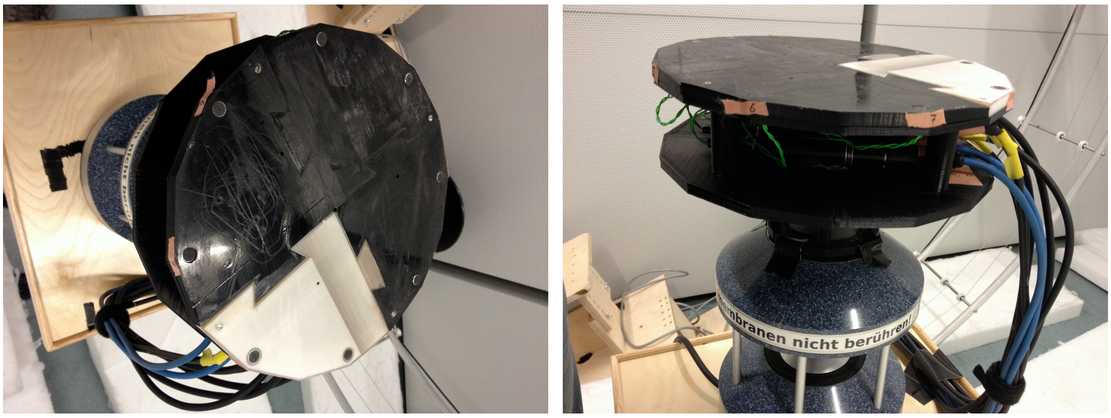


Figure 3.3.: Technical drawing of the old and the black array.

Due to the number of the microphone arrays, we need names for distinguishing them. The planar array is called the “black" one.



(a) Top view of the black microphone array

(b) Sideview of the black microphone array

Figure 3.4.: The black microphone array

3.1.1. A new conchiform attachment

In this thesis, not only beamforming is performed but also other algorithms are used. The COMPaSS-algorithm needs previously determined TFs to localize the sound sources. To distinguish individual TFs for each direction, a new conchiform part for each microphone was designed. As the shell-shaped attachment [12] is mathematically equivalent to an Archimedean spiral (Figure 3.5), it could not be created with *Google Sketchup*. For this, the open source software *OpenSCAD* was used. The Archimedean spiral has the feature that each point of the helix has a different distance from the center, because the radius

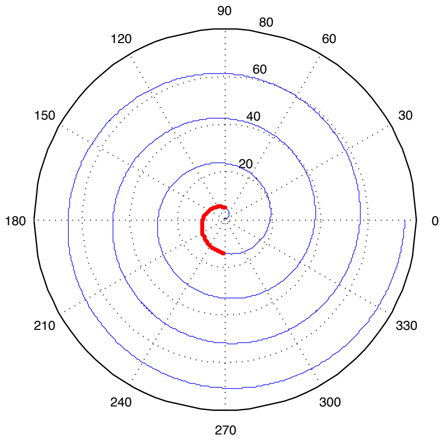


Figure 3.5.: Archimedean Spiral

$$r = b \cdot \gamma \tag{3.1}$$

is a function of the angle γ. To get the red part in Figure 3.5, the angle γ has to run from $\frac{\pi}{2}$ to $\frac{3\cdot\pi}{2}$ and to get the same dimensions as in Figure 3.6 the constant b must have the value 1.0122254.

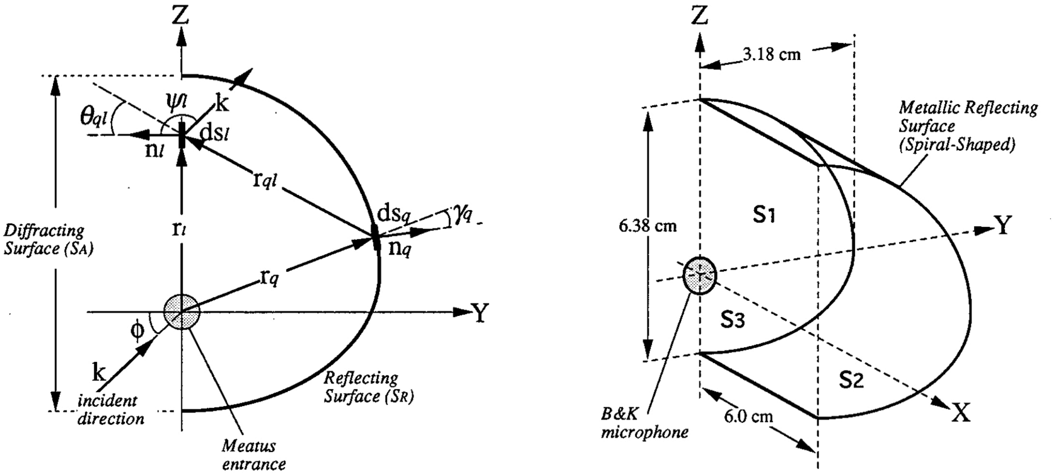
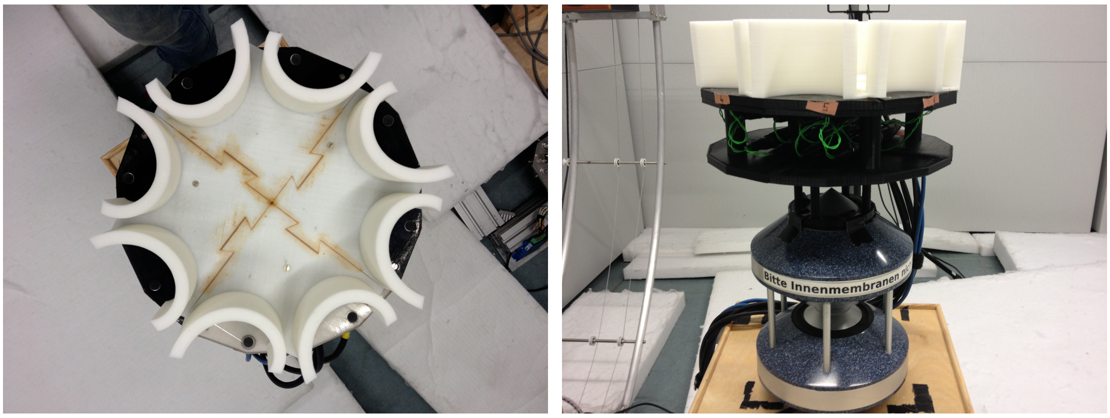


Figure 3.6.: Technical drawing of the shell-shaped attachment [12]



(a) Top view of the black-concha microphone array

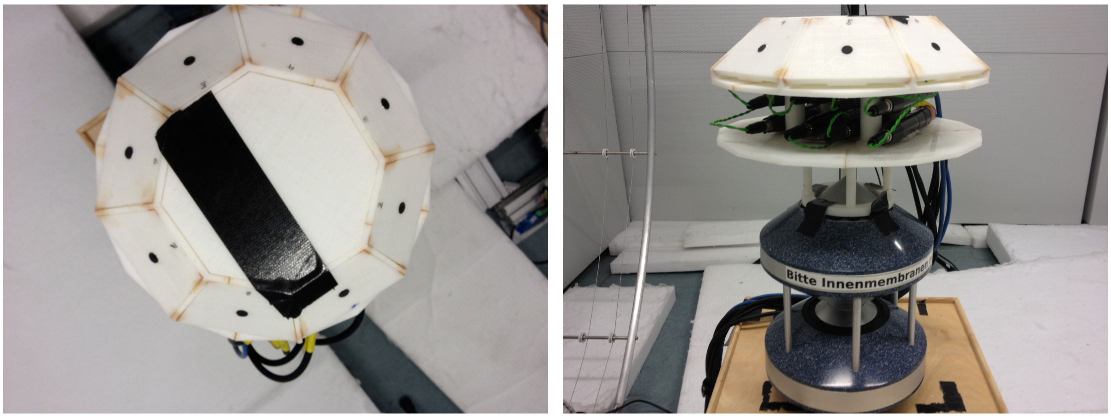
(b) Sideview of the black-concha microphone array

Figure 3.7.: The black-concha microphone array

The black microphone array, with the shell-shaped attachment is called “black-concha" in this thesis.

3.2. The white microphone array

The idea for this array arose after the observation of the elevation depending amplitude of the TFs in the black array and in the black array with the shell-shaped attachment (A.1, A.2, A.3, A.4, A.5, A.6). The arrangement of the microphones in the direction of the sources supplies more uniform amplitudes for different elevations. A further advantage of this microphone arrangement is the natural source separation because of the mechanical shielding for sources on the opposite. In the case of BSS this array should deliver better results. How well this array works effecively one can read in chapter 4.4.



(a) Top view of the white microphone array

(b) Sideview of the white microphone array

Figure 3.8.: The white microphone array

The name for this array, which is used in this thesis is “white".

4. Experiments

4.1. Evaluation

4.1.1. Quality of source separation

To examine the quality of a source separation, one needs to compare an estimated source $\hat{\mathbf{s}}_j$ with the original source $\mathbf{s}_j$. For this purpose the *BSS Eval toolbox* is used [23]. The estimated source signal is computed by

$$\hat{s}_j = s_{target} + e_{interf} + e_{noise} + e_{artif}. \tag{4.1}$$

$\mathbf{s}_{target}$ is a version of $\mathbf{s}_j$ with an allowed distortion $f \in \mathscr{F}$. The other three terms stand for the interference, noise and artifact errors. $\mathbf{s}_j$ is the required source and $\mathbf{s}_{j'}$ is one of the other sources. In [22] a decompostion is proposed, which is based on orthogonal projections. $\prod\{\mathbf{y}_1, \dots, \mathbf{y}_k\}$ is the orthogonal projector onto the subspace spanned by the vectors $\mathbf{y}_1, \dots, \mathbf{y}_k$. The three projectors

$$P_{s_j} := \prod\{s_j\}, \tag{4.2}$$

$$P_{\mathbf{s}} := \prod\{(s_{j'})_{1 \le j' \le n}\}, \tag{4.3}$$

$$P_{\mathbf{s},\mathbf{n}} := \prod\{(s_{j'})_{1 \le j' \le n}, (n_i)_{1 \le i \le m}\} \tag{4.4}$$

are needed to define the four terms of the estimated source signal $\hat{s}_j$

$$s_{target} := P_{s_j}\hat{s}_j, \tag{4.5}$$

$$e_{interf} := P_{\mathbf{s}}\hat{s}_j - P_{s_j}\hat{s}_j, \tag{4.6}$$

$$e_{noise} := P_{\mathbf{s},\mathbf{n}}\hat{s}_j - P_{\mathbf{s}}\hat{s}_j \tag{4.7}$$

and

$$e_{artif} := \hat{s}_j - P_{\mathbf{s},\mathbf{n}}\hat{s}_j. \tag{4.8}$$

From the decompostion of $\hat{\mathbf{s}}_j$ and the computation of different energy ratios, one can calculate the Source to Distortion Ratio (SDR)

$$\text{SDR} := 10\log_{10}\frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2}, \tag{4.9}$$

the Source to Interference Ratio (SIR)

$$\text{SIR} := 10\log_{10}\frac{||s_{target}||^2}{||e_{interf}||^2}, \tag{4.10}$$

and the Sources to Artifacts Ratio (SAR)

$$\text{SAR} := 10\log_{10}\frac{\|s_{target}+e_{interf}+e_{noise}\|^2}{\|e_{artif}\|^2} \tag{4.11}$$

in decibel (dB).

4.1.2. Quality of the localization algorithms

The localization accuracy is denoted with the same parameters as in [5]. The recordings for this quality measurement include no breaks, because with breaks one can not give a statement about the localization accuracy.

Exact accuracy This value means the number of correctly localized sources to all detected sources in percent. A correctly detected source is nearer to the recognized grid point as to another grid point.

Tolerance region Not for all application it makes sense to look at the exact accuracy, for example if the grid distances are very small. The tolerance region is a predefined region around the true source position, which counts as correctly detected in this case. This number is also in percent.

Mean Angular Error (MAE) In contrast to the both other numbers the MAE shows the real deviation in degree, so one can compare different algorithms with different grid distances better.

4.2. Simulations

To test the several algorithms for sound source localization and source seperation, one can make simulations with the recorded impulse responses [5], but this makes no sense for the COMPaSS algorithm. The COMPaSS algorithm uses the impulse responses for calculating the source location, so one gets a perfect localization if the same impulse responses are used for simulation and calculation [5]. The maximum in all samples of the impulse responses is searched for scaling the impulse responses to one before convolving a recorded speaker with an impulse responses. In the case of more than one speaker the convolved files are summed up for each channel separately and devided by three, to avoid clipping in the wav-files.

4.3. Recordings

In this thesis a lot of attention is paid to the reproducibility of the tests. The used Phantom Power Adaptors deliver such a low output level, that the gain controls of the microphone preamps can be cranked up completely. The used microphones are not high-end ones, so one does not receive the same output power, whereas the preamps are adjusted to the same value. The level adaptation is proceeded on the computer while a sinus signal is played with a *Pro 100 (NewTec*

Figure 4.1.: Pro 100 loudspeaker from NewTec [1]

Design : Audio, Figure 4.1) loudspeaker. The special characteristic of this loudspeaker is the consistent emission of sound waves for every direction from $0°$ to $360°$. The microphone arrays were put on this loudspeaker, so that each microphone has the same distance to the center of this loudspeaker. The microphones of both arrays have a number and they were plugged into the corresponding numbers of the preamplifiers.

4.3.1. Impulse Response (IR) and Transfer Function (TF)

First, the impulse responses were determined for the three microphone arrays in five degree steps in azimuth and for three elevations ($0°$, $10°$, $20°$) on one day with the same settings in the anechoic chamber of the Institute for Data Processing, so one can compare the transfer functions and the impulse responses between them. The distance from the center of the array to the front side of the speaker was 1.3m. A further facility is a computer controlled turntable on which the microphone array is put as one can see in Figure 4.2. So one has to measure the distances and angles one time and for the measurement of the different azimuth angles the array is turned by the turntable. In Table 4.1 are the measured temperatures during the recordings for the IRs. They are needed in the SRP-PHAT-algorithm. The impulse responses can not be measured directly, but one can calculate them via cross-correlation from the recorded signal with the source signal. As a source signal, a special Maximum Length Sequence (MLS) signal is used.

Black-array The impulse responses look nearly the same for the different azimuth angles and also for the three elevations (A.1, A.2, A.3), because the sound source and the microphones are in the line of sight, so that the sound waves arrive almost unchanged at the microphones. The transfer functions for one elevation have the same amplitude, and one can see the time-of-arrival difference between the microphone at the loudspeaker side

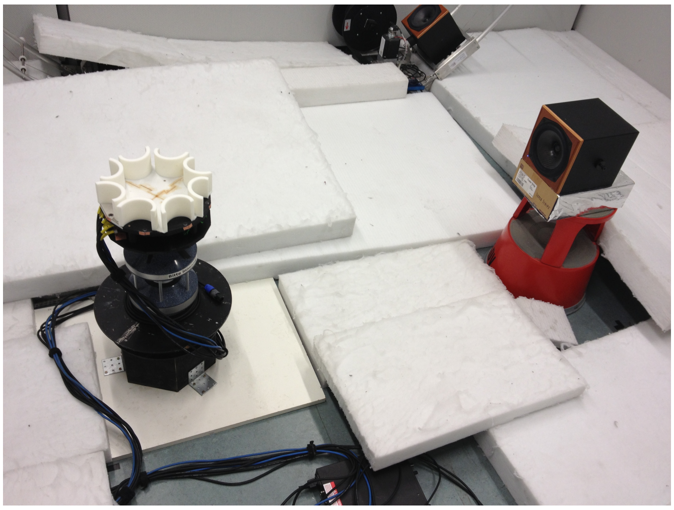


Figure 4.2.: Impulse response measurement for the conference phone with concha

and the microphone on the opposite (A.1, A.2, A.3). The amplitude of the transfer functions increases with the elevation angle, because of the microphone characteristics. At elevation 0° the direction of propagation of the sound waves is parallel to the microphone membrane which explains the difference in the amplitude.

Black-concha-array The transfer funcions (A.4, A.5, A.6) show the time delays between the microphones on the speaker side and the microphones on the far side. The amplitude at the microphones of the back side is obviously smaller because the shielding of the shell-shaped attachment. Furthermore, the amplitude of the TFs increases with the elevation angle, because of the angle between the microphone membrane and the direction of the sound waves. The IRs contain the expected angle dependent spectral effects.

White-array As seen earlier, in both array configurations the amplitude is dependent on the angle between the sound waves and the microphone membrane. For this reason, a new array with no planar disposed microphones was designed for which nearly no difference in the amplitude exists (A.7, A.8, A.9). The disposal of the microphones in the speaker direction, should support the localization and the source separation. The TFs show again the time delays between the microphones in the front and in the back. Moreover, one can see the smaller amplitude at the shielded microphones.

MLS and IR-calculation

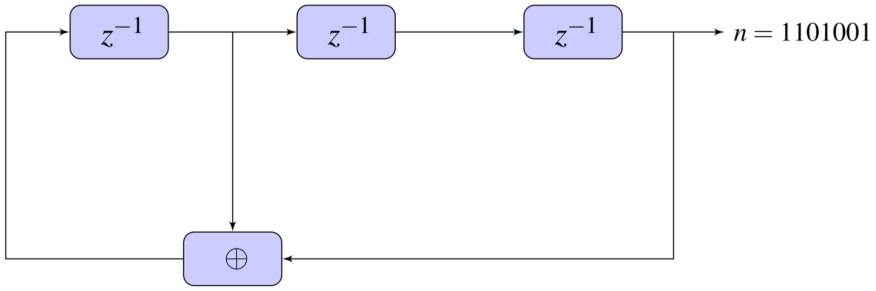
An MLS is a binary, pseudo-random noise signal of length $P = 2^N - 1$, whereas N is an integer. This definition ensures that the length of the MLS is odd. To generate an MLS signal a recursive formula is used, where k is the index and $\oplus$ is the XOR operator. The first N digits of n have to be set with 1 and 0, but not all N digits with 0, because then all digits are 0.

array-typ	elevation	temperature in degree
black_ concha	0	22.8
	10	24.0
	20	23.1
black	0	24.5
	10	21.5
	20	22.6
white	0	24.2
	10	23.6
	20	23.7

Table 4.1.: Temperature during the measurement of the impulse responses.

Example for one version of an MLS with 7 digits

$$\begin{gathered} n(k+3) = n(k+2) \oplus n(k) \\ n(1) = 1,\, n(2) = 1,\, n(3) = 0 \\ P = 2^3 - 1 = 7 \\ n(4) = n(3) \oplus n(1) = 0 \oplus 1 = 1 \\ n(5) = n(4) \oplus n(2) = 1 \oplus 1 = 0 \\ n(6) = n(5) \oplus n(3) = 0 \oplus 0 = 0 \\ n(7) = n(6) \oplus n(4) = 0 \oplus 1 = 1 \\ n = 1101001 \end{gathered}$$



If the MLS-signal is used in signal processing, it may not have a steady component. For this reason the components of the MLS-signal are not 1 and 0, but 1 and -1.

MLS characteristics:

- The number of 1's is exactly one more than the number of 0's.
- The autocorrelation of an MLS-signal is a perfect impulse.

4.4. Teleconference scenario

To evaluate the different algorithms a lot of recordings were made in the audiolab (Figure 4.3) and in the videolab (Figure 4.4).

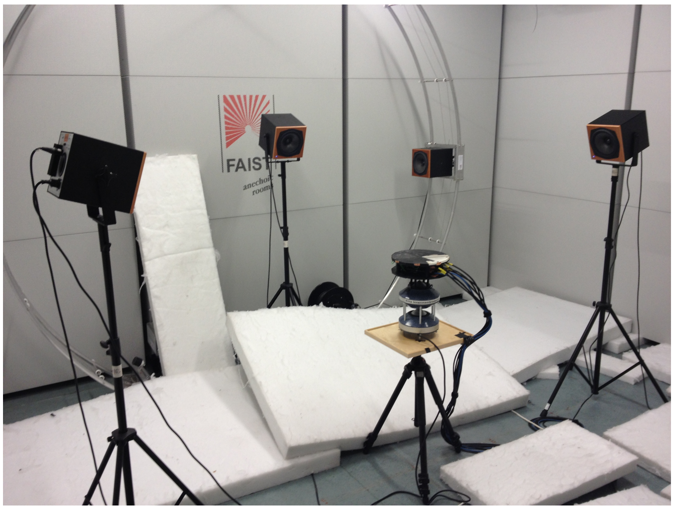


Figure 4.3.: Recording setup in the audiolab

Sound Source	KS Digital C5 Tiny
Sound Card	RME Multiface II
Microphone preamplifiers	Focusrite Sapphire Pro 40
Audiolab dimensions	4.7m x 3.7m x 2.84m
Audiolab noise level	<30dBA
Audiolab reverberation time t_{60}	0.08s

Table 4.2.: Experimental setup in the audiolab

Sound Source	KS Digital C5 Tiny
Sound Card	RME Multiface II
Microphone preamplifiers	Focusrite Sapphire Pro 40
Videolab dimensions	6.3m x 4m x 2.8m
Videolab reverberation time t_{60}	0.64s

Table 4.3.: Experimental setup in the videolab

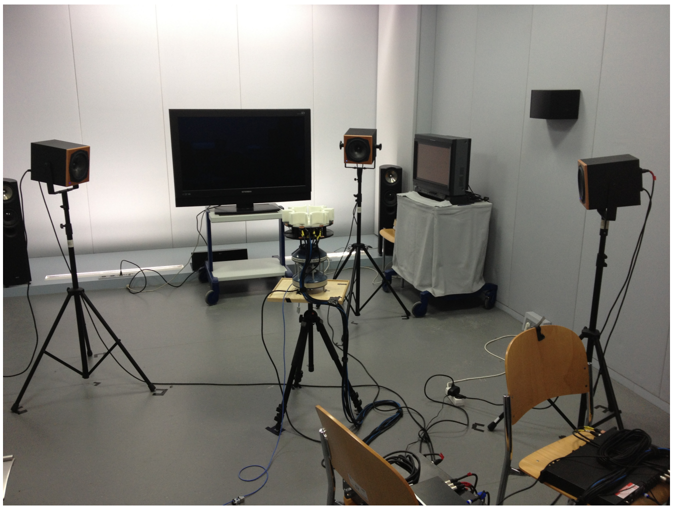


Figure 4.4.: Recording setup in the videolab

4.4.1. Speaker recording

First, 8 male speakers and 4 female speakers were recorded in the anechoic chamber of the Institute for Data Processing, as one can see in Figure 4.5. They had to read a passage from a book over 6 minutes. In this thesis sections with only a duration of 10 seconds are used, but in other research areas like speaker recognition the whole recordings are needed. The speakers who are used in this thesis you can see in the appendix A.1.

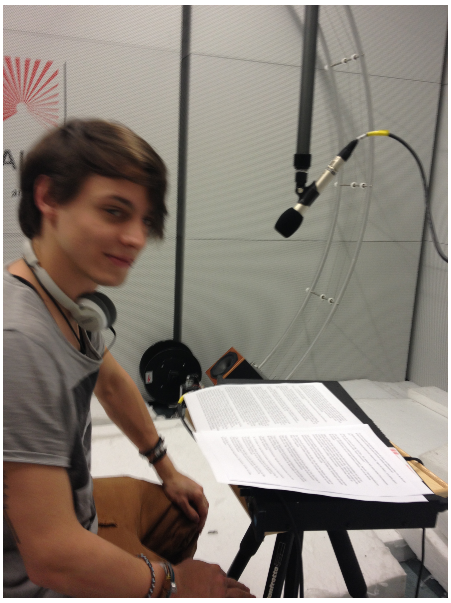


Figure 4.5.: Recording a speaker

4.4.2. Conference Recording

In order to evaluate the algorithms, which are described in section 2, a lot of recordings were made in the anechoic audiolab and in the videolab, a room with office attributes. As we evaluate three microphone arrays with two elevations (10° and 20°) in two different rooms with 108 combinations of speakers (Table A.2), at least $3 * 2 * 2 * 108 = 1296$ recordings, we decide for one recording setup (Figure 4.6). The participants of a teleconference also do not sit in one row, but they sit on a table on the opposite or with 90° gap. For this reason, the recording setup represents a real teleconference. The previously recorded speakers were played from one of the three loudspeakers. The recordings are wav-files with 8 channels and a duration of 10s. In one wav-file either one, two or three speakers are talking simultaneously over the whole duration.

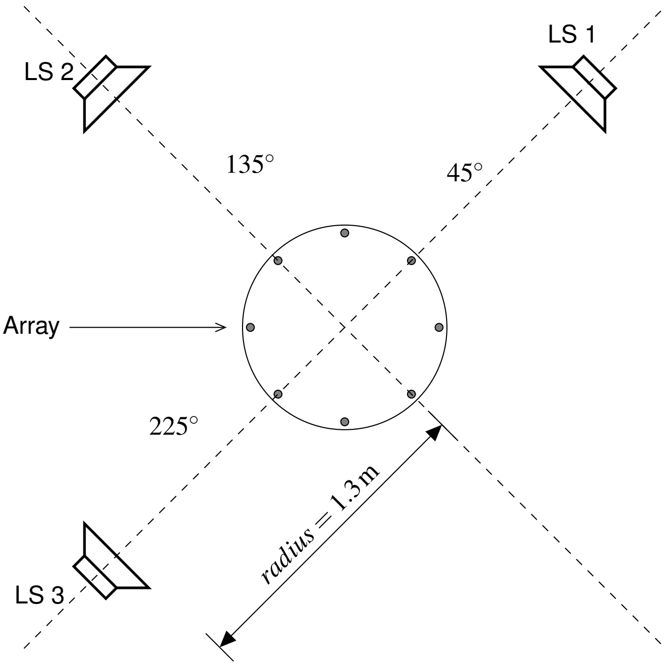


Figure 4.6.: The recording configuration for the audiolab and for the videolab

4.4.3. Localization results

The detailed localization results are in the appendix A.7. For both localization algorithms following values are calculated:

- **MAEazi:** MAE in azimuth direction
- **MAEele:** MAE in elevation direction
- **TOLazi:** Percentage of recordings which are in the tolerance region of 5° in azimuth direction

	black				**black-concha**				**white**			
	audiolab		videolab		audiolab		videolab		audiolab		videolab	
	10°	20°	10°	20°	10°	20°	10°	20°	10°	20°	10°	20°
one source												
COMPaSS	-	-	-	-	✓	✓	-	✓	-	✓	-	-
SRP-PHAT	✓	✓	✓	✓	-	-	✓	-	✓	-	✓	✓
two sources												
COMPaSS	-	-	-	-	-	✓	-	-	-	-	-	-
SRP-PHAT	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	✓
three sources												
COMPaSS	-	-	-	-	-	-	-	-	-	-	-	-
SRP-PHAT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 4.4.: Comparison of the localization approaches

- **TOLele:** Percentage of recordings which are in the tolerance region of $5°$ in elevation direction

To find the best localization algorithm the case of two sources is used, because the particle filter implementation of the COMPaSS-algorithm produces errors for three sources. In Table 4.4 the MAEazi is the determining value to find the winner, because the search region of the COMPaSS algorithm is smaller and depending on the measured IRs. The calculated percentages of recordings in the tolerance region make a inaccurater statement then the MAE-values. Table 4.4 delivers an obvious winner, the SRP-PHAT-algorithm. A further advantage of the SRP-PHAT-algorithm that one does not have to determine the IRs.

4.4.4. Separation results

In this chapter the separation results are presented. Binary masking and GSS need the positions of the sources. In order to evlauate the separation algorithms independently from the localization results, the correct positions of the sources are put in the algorithms. For this reason, the values show what results with the algoritms are theoretical reachable. In appendix A.8 all separation results are presented. Binary masking provides good SIR-values indeed, but the SAR and SDR-values are worse than at the both other algorithms. The auditory impression is also worse, because of disturbing artifacts. IVA has the best results at the simulations and GSS at the recordings. The problem of the GSS-algorithm with the simulated files can be investigated in the future. In this thesis the winner is the GSS-algorithm, because it delivers the best results in the recordings in the audiolab and in the videolab. Simulations are good to test different algorithms, but they are irrelevant in practise. In Table 4.5 the separation is done with the positions of the SRP-PHAT localizer. Therefore the second column shows the results in the videolab, which are reachable in a real teleconference.

SRP-PHAT+GSS	Audiolab			Videolab		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
elevation: 10°						
one source						
black	15.9	Inf	15.9	5.1	Inf	5.1
black-concha	16.4	Inf	16.4	5.2	Inf	5.2
white	15.7	Inf	15.7	4.3	Inf	4.3
two sources						
black	7.7	15.0	9.1	2.4	11.7	3.4
black-concha	5.8	14.3	6.8	1.9	11.5	2.8
white	7.8	14.5	9.4	2.3	11.4	3.4
three sources						
black	2.2	11.3	3.3	-1.0	8.4	0.2
black-concha	0.5	8.8	2.0	-2.1	6.6	-0.3
white	2.3	10.2	3.7	-1.3	7.4	0.3
elevation: 20°						
one source						
black	14.4	Inf	14.4	5.9	Inf	5.9
black-concha	13.2	Inf	13.2	6.3	Inf	6.3
white	13.1	Inf	13.1	5.4	Inf	5.4
two sources						
black	7.3	12.9	9.3	2.7	11.4	3.8
black-concha	5.8	13.7	6.9	2.6	12.3	3.4
white	8.0	14.1	9.7	3.1	12.0	4.1
three sources						
black	1.0	8.7	2.8	-1.7	7.4	-0.1
black-concha	0.5	9.0	2.0	-1.8	7.0	-0.1
white	2.5	10.7	3.7	-0.7	8.4	0.7

Table 4.5.: The results of GSS after the SRP-PHAT

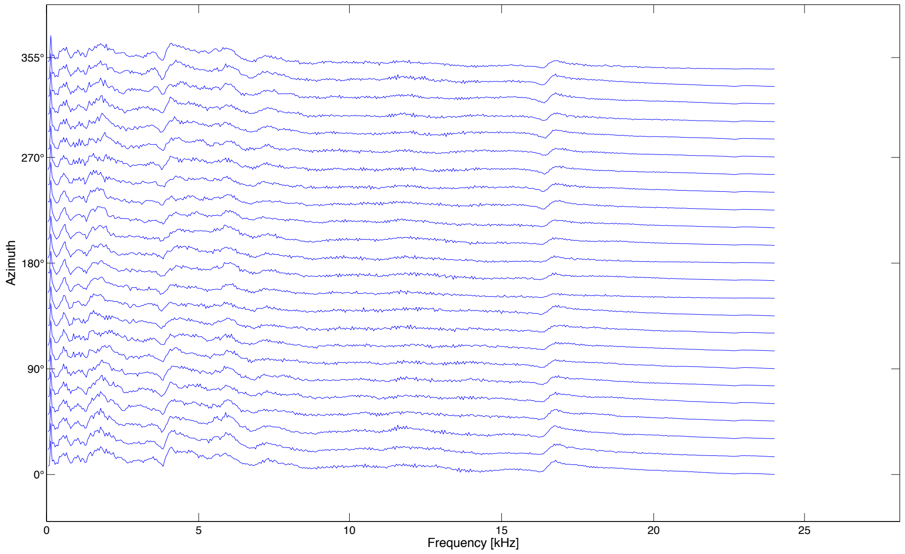
5. Conclusion

In this thesis different algorithms are presented and evaluated with the same test data. All algorithms have the right to exist, because they all deliver suitable results. In the teleconference case the SRP-PHAT-algorithm supplies the best localization results and the GSS the best SDR, SIR and SAR-values and also a subjective hearing test has confirmed this. The black-concha array, which was designed for the COMPaSS-algorithm does not operate with the SRP-PHAT and the GSS-algorithm as expected. The both other arrays nearly make no differnce in the performance. The next step is the adjustment of the algorithms on the teleconference scenario where the state is not uniform, as in the recordings of this thesis. The particle filter of the SRP-PHAT-algorithm recognizes the number of active speakers very well. Furthermore, the separation can be segmented in parts with the same condition. Before such a modern conference phone reaches market maturity, a lot of research effort will be needed, but a good basis is already existing. In future it will be seen, when such a modern conference phone can be bought and which algorithms are implemented.

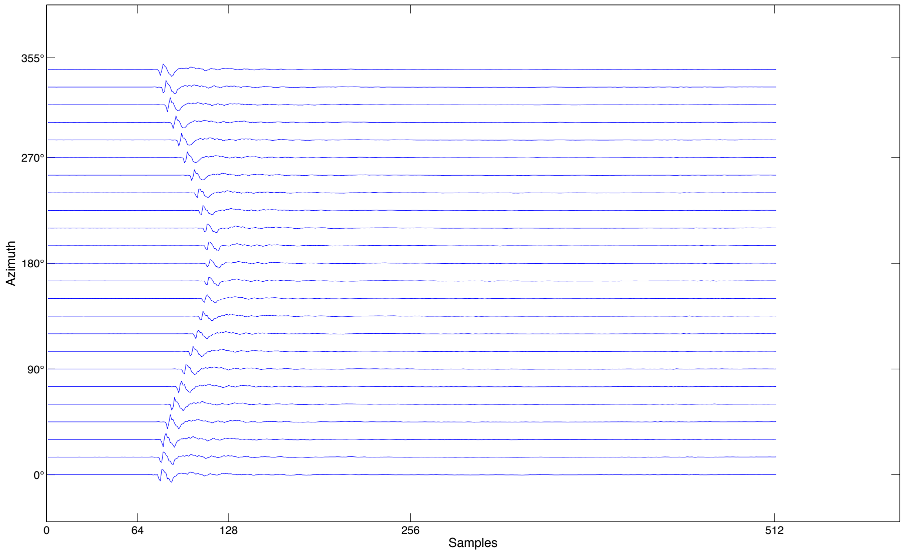
A. Appendix

A.1. Impulse responses

A.1.1. Black microphone array

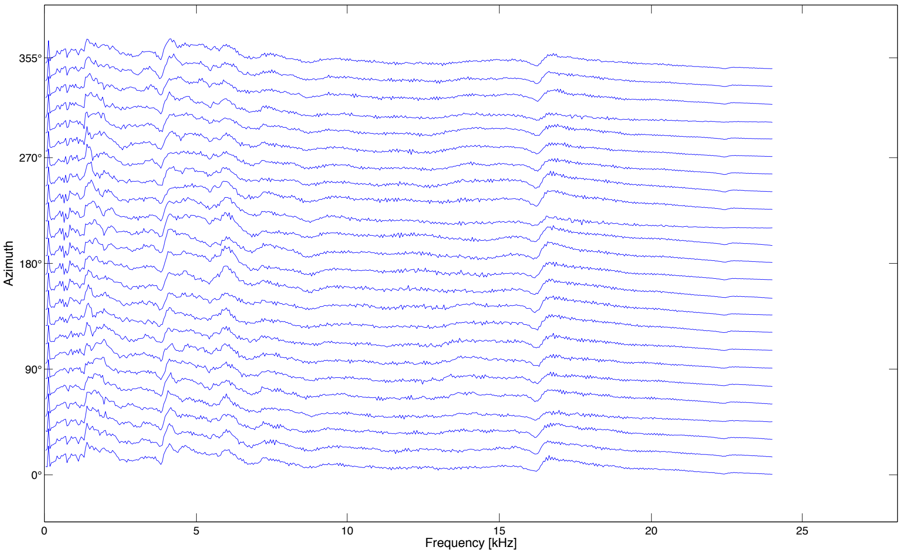


(a) Impulse responses for every 15°

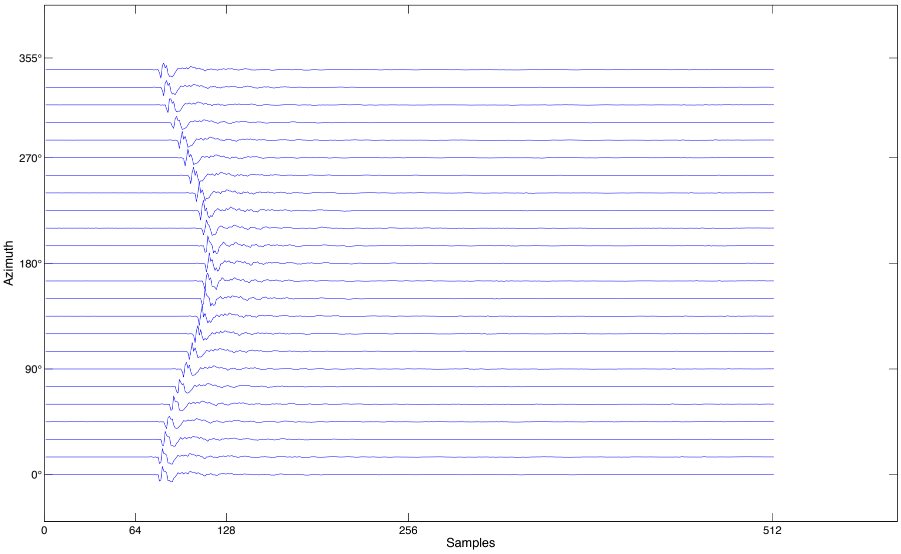


(b) Transfer-function for every 15°

Figure A.1.: Elevation 0° for microphone 1 of the black microphone array

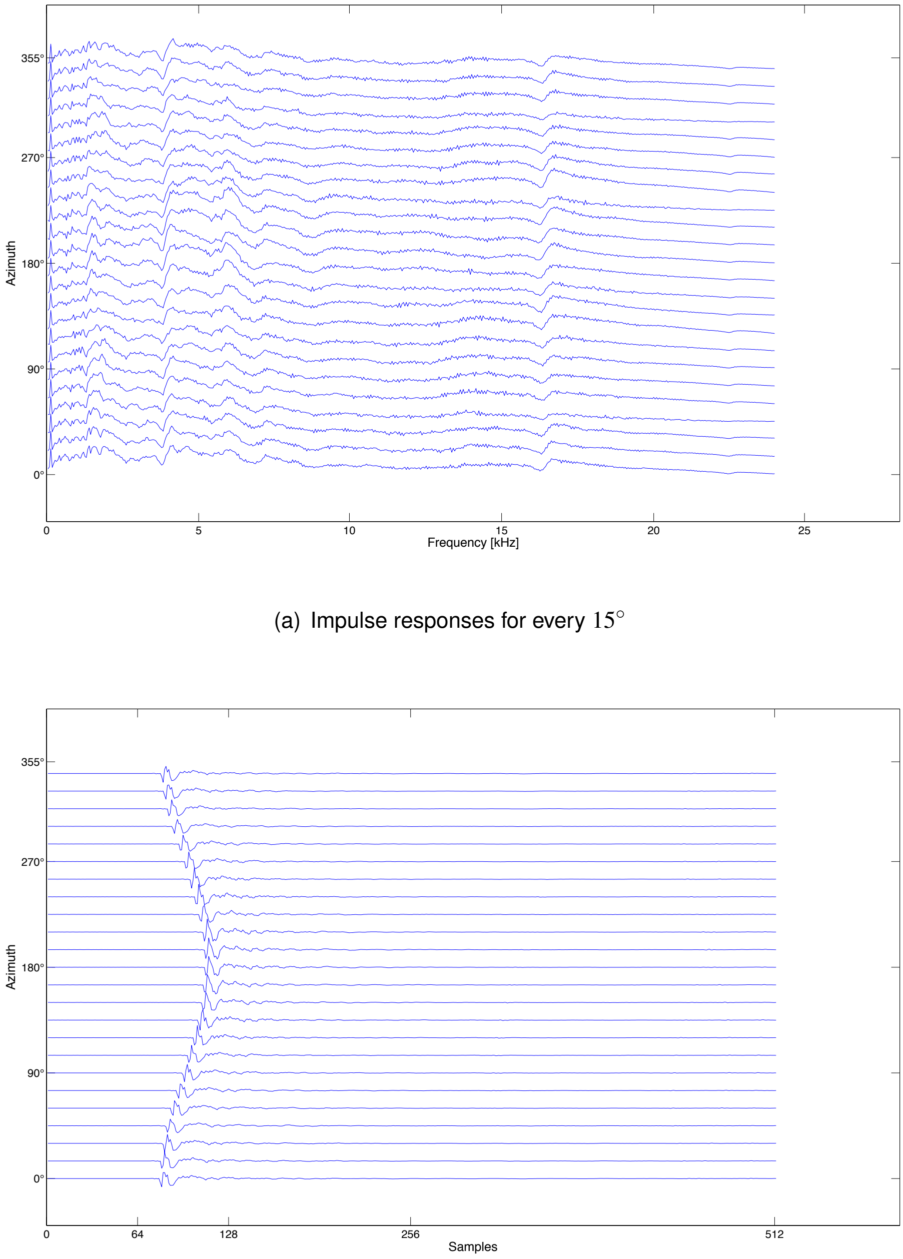


(a) Impulse responses for every 15°



(b) Transfer-function for every 15°

Figure A.2.: Elevation 10° for microphone 1 of the black microphone array

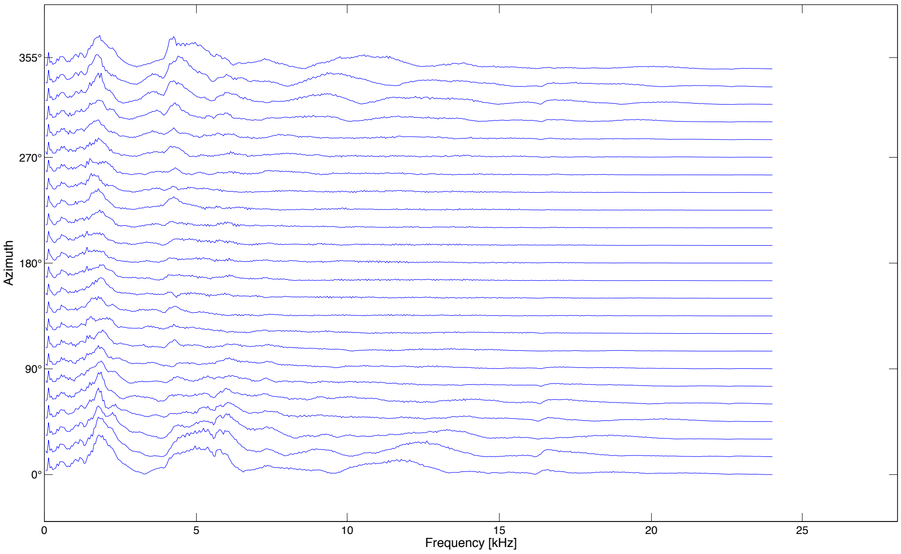


(a) Impulse responses for every 15°

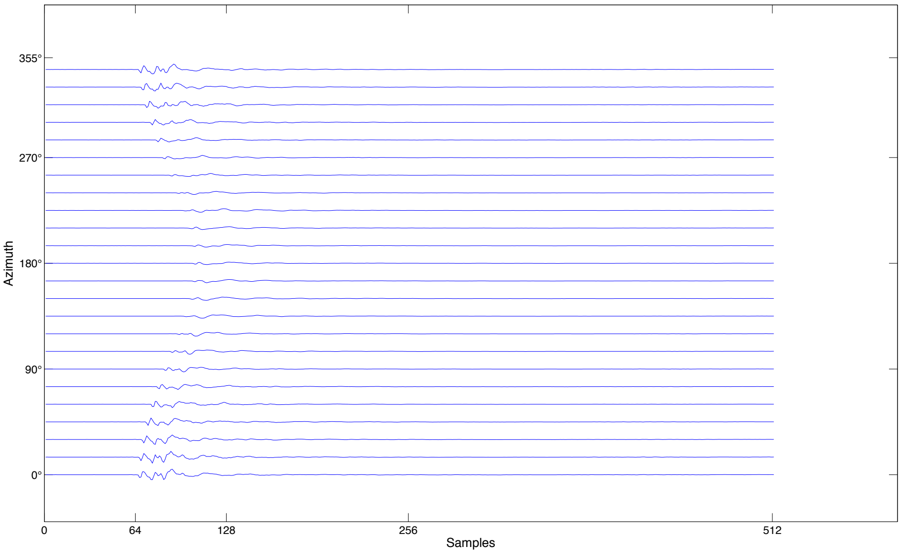
(b) Transfer-function for every 15°

Figure A.3.: Elevation 20° for microphone 1of the black microphone array

A.1.2. Black-concha microphone array

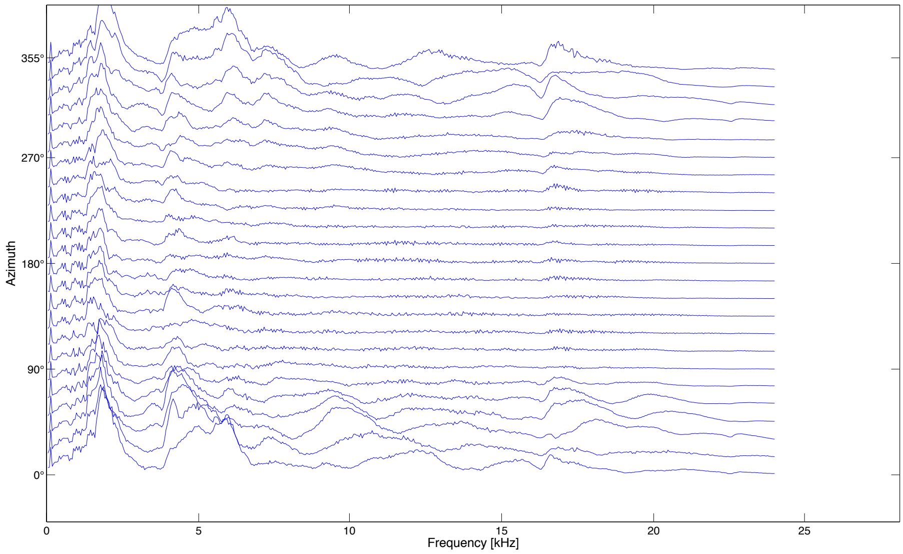


(a) Impulse responses for every 15°

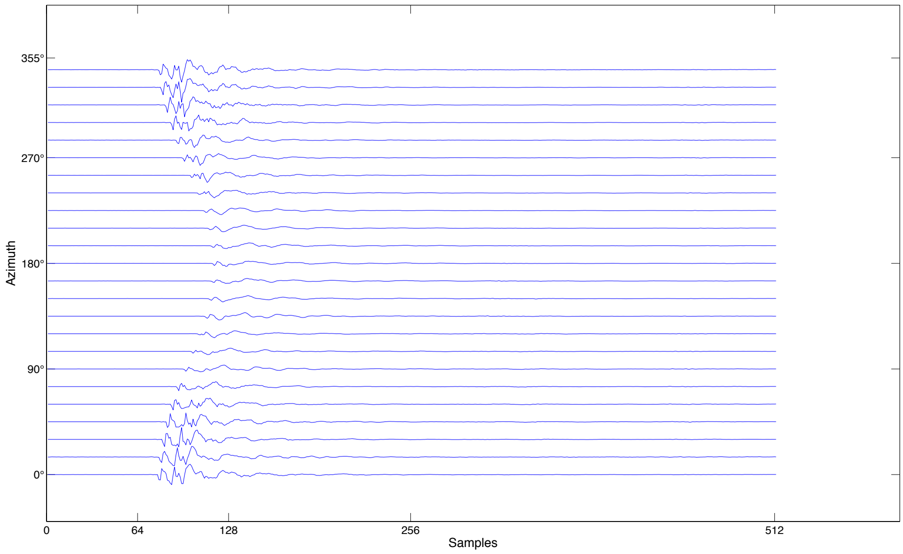


(b) Transfer-function for every 15°

Figure A.4.: Elevation 0° for microphone 1 of the black-concha microphone array

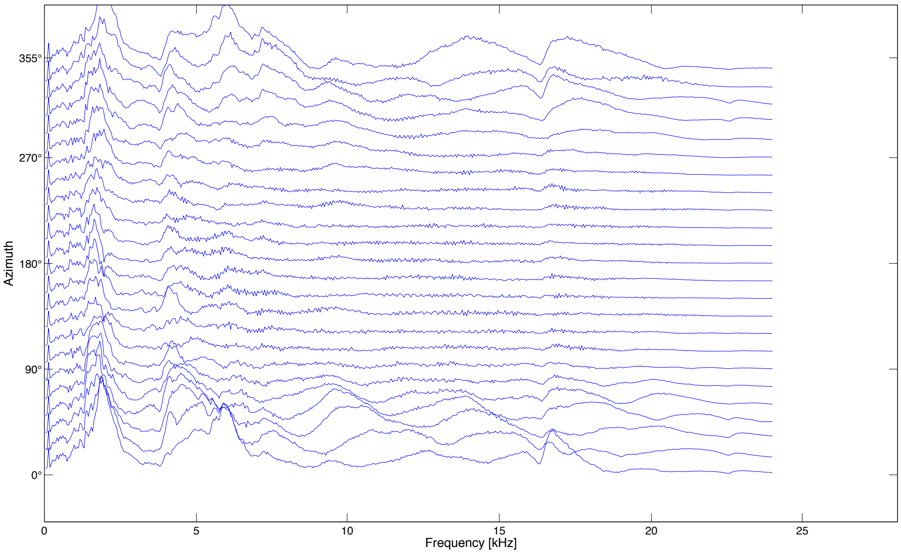


(a) Impulse responses for every 15°

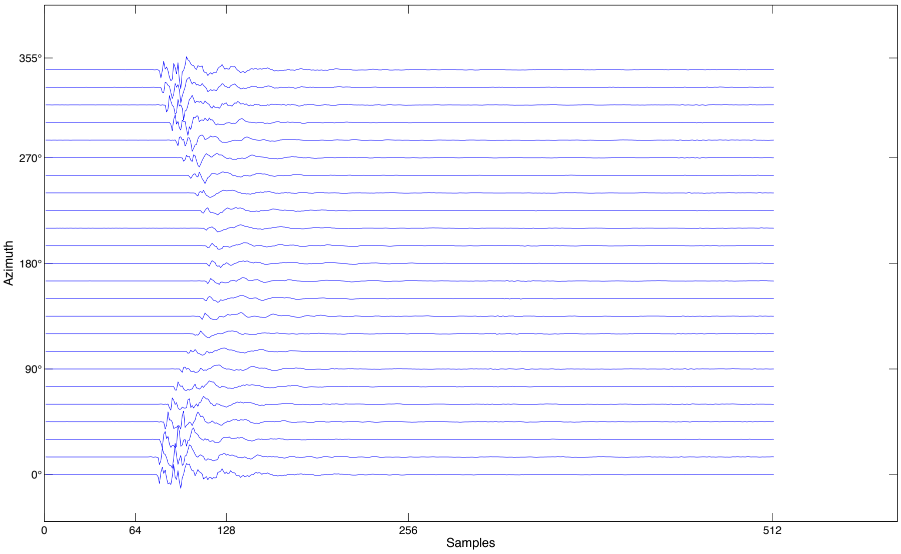


(b) Transfer-function for every 15°

Figure A.5.: Elevation 10° for microphone 1 of the black-concha microphone array



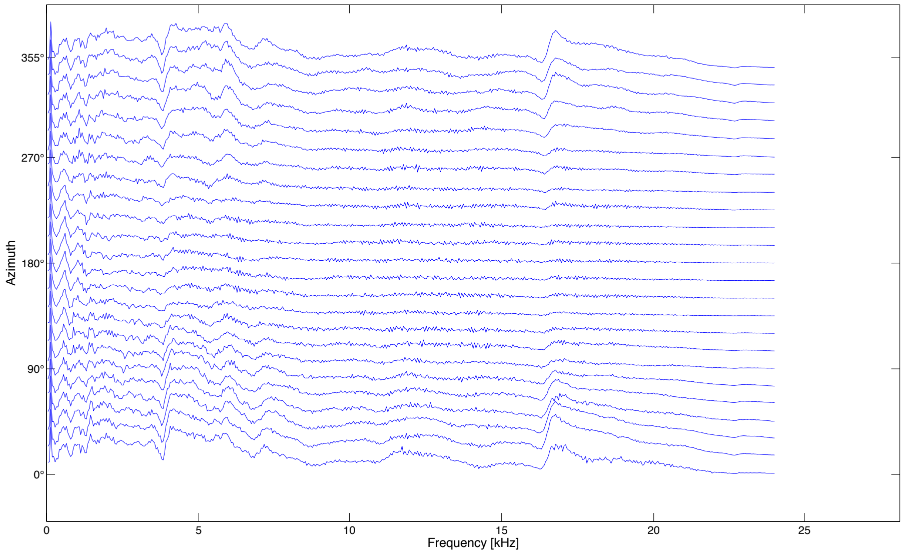
(a) Impulse responses for every 15°



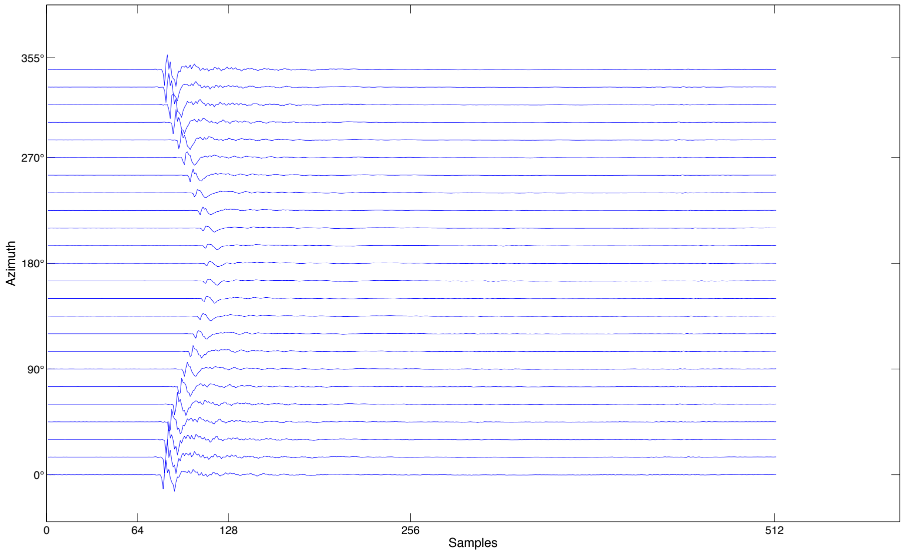
(b) Transfer-function for every 15°

Figure A.6.: Elevation 20° for microphone 1 of the black-concha microphone array

A.1.3. White microphone array

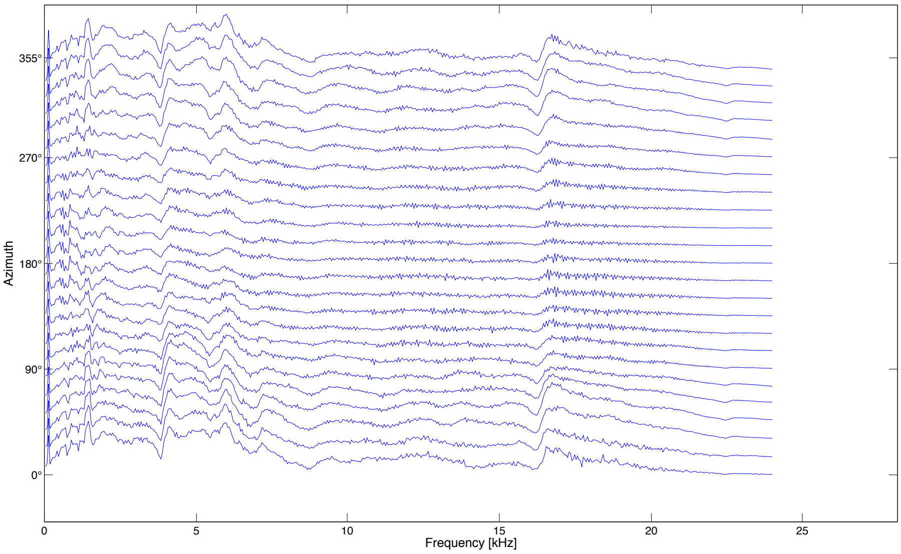


(a) Impulse responses for every 15°

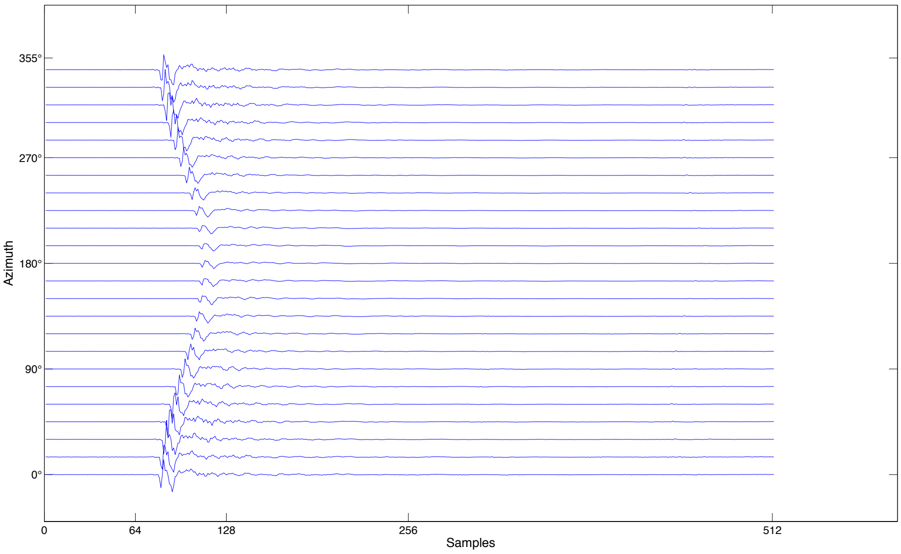


(b) Transfer-function for every 15°

Figure A.7.: Elevation 0° for microphone 1 of the white microphone array

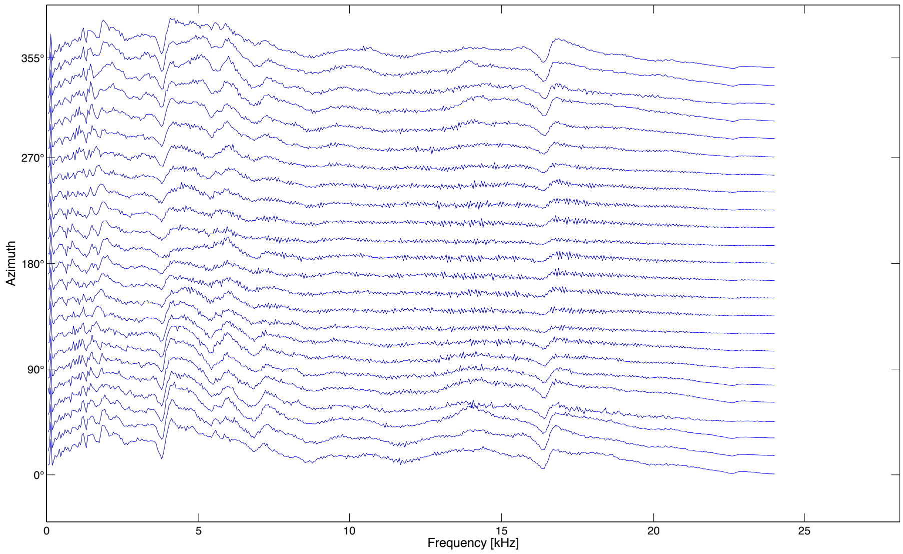


(a) Impulse responses for every 15°

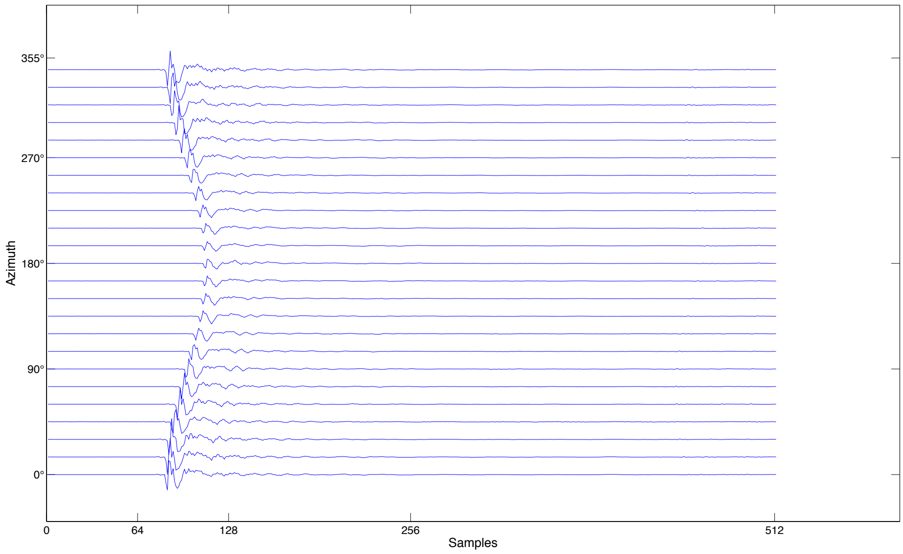


(b) Transfer-function for every 15°

Figure A.8.: Elevation 10° for microphone 1 of the white microphone array



(a) Impulse responses for every 15°



(b) Transfer-function for every 15°

Figure A.9.: Elevation 20° for microphone 1 of the white microphone array

A.2. Principle Component Analysis (PCA)

The goal of PCA is to transform a data set in a new base, so that the noise is filtered out and the important data will be disclosed. One seeks a new base in the direction in which the data will have maximum variance, since it is assumed that the direction with the greatest variance will also describe the interesting dynamic of the system. The following table with the related Figures shows the PCA procedure.

1. Subtraction of the mean values (figure: A.11)
2. Calculating the covariance matrix
3. Computing the eigenvectors and eigenvalues of the covariance matrix
4. Calculation of the new data (figure: A.12)

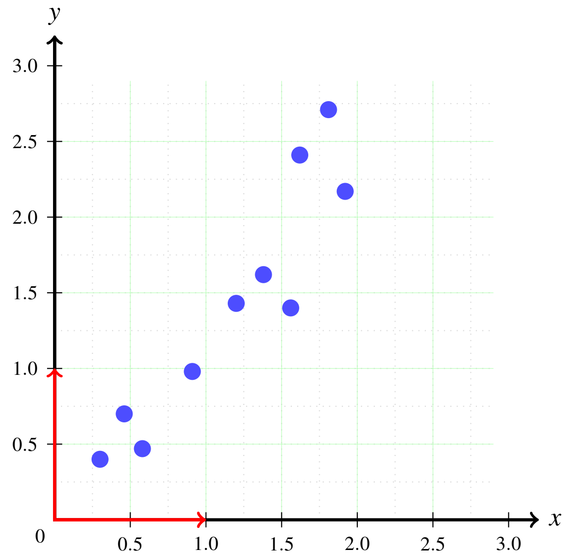


Figure A.10.: Dataset X

PCA is also used for the subspace method [9] to find the number of active sources. The anaylsis of the eigenvalues delivers the dominant eigenvalues. These dominant eigenvalues indicate how many sources are active.

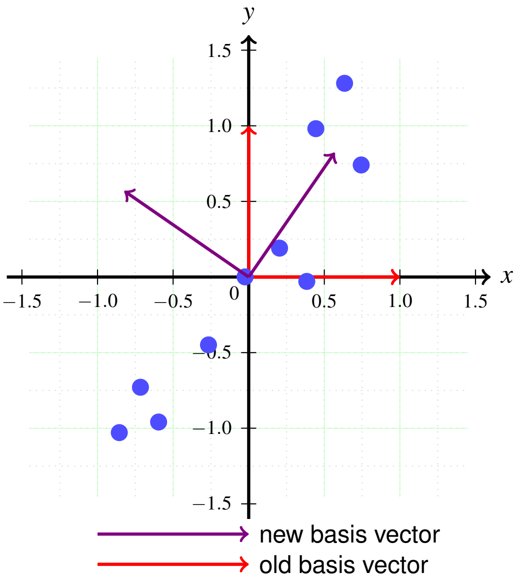


Figure A.11.: Dataset free from averaging

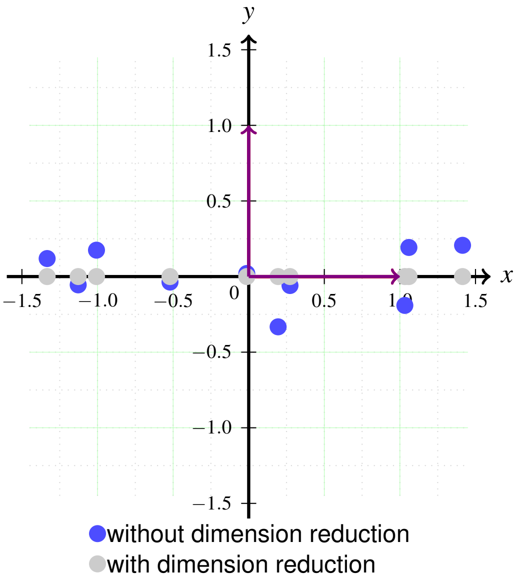


Figure A.12.: Dataset free form averaging with the new base

A.3. Recorded speaker

Speaker male	*b_ Thomas.wav*,	*d_ Thomas.wav*
	b_ Alex.wav,	*d_ Alex.wav*
	b_ Tom.wav,	*d_ Tom.wav*
	b_ Andre.wav,	*d_ Andre.wav*
	b_ Benedikt.wav,	*d_ Benedikt.wav*
	b_ Martin.wav,	*d_ Martin.wav*
	b_ Jonas.wav,	*d_ Jonas.wav*
	b_ Richard.wav,	*d_ Richard.wav*
Speaker female	*b_ Ricarda.wav*,	*d_ Ricarda.wav*
	b_ Imen.wav,	*d_ Imen.wav*
	b_ Kathrin.wav,	*d_ Kathrin.wav*
	b_ Lisa.wav,	*d_ Lisa.wav*

Table A.1.: Speaker sources, *b* stands for the begin of a speaker and *d* stands for a part in the middle of the recording

A.4. Recorded files

One Source	Two Sources	Three Sources
b_ Alex_ 45	b_ Alex_ 45_ b_ Andre_ 225	b_ Alex_ 135_ d_ Andre_ 45_ b_ Benedikt_ 225
b_ Alex_ 135	b_ Alex_ 45_ b_ Benedikt_ 225	b_ Alex_ 135_ d_ Martin_ 45_ b_ Jonas_ 225
b_ Alex_ 225	b_ Alex_ 135_ b_ Andre_ 45	b_ Andre_ 135_ d_ Benedikt_ 45_ b_ Martin_ 225
b_ Andre_ 45	b_ Alex_ 135_ b_ Benedikt_ 45	b_ Andre_ 135_ d_ Richard_ 45_ b_ Kathrin_ 225
b_ Andre_ 135	b_ Andre_ 45_ b_ Benedikt_ 225	b_ Benedikt_ 135_ d_ Imen_ 45_ b_ Lisa_ 225
b_ Andre_ 225	b_ Andre_ 45_ b_ Martin_ 225	b_ Benedikt_ 135_ d_ Martin_ 45_ b_ Richard_ 225
b_ Benedikt_ 45	b_ Andre_ 135_ b_ Benedikt_ 45	b_ Imen_ 135_ d_ Jonas_ 45_ b_ Kathrin_ 225
b_ Benedikt_ 135	b_ Andre_ 135_ b_ Martin_ 45	b_ Imen_ 135_ d_ Lisa_ 45_ b_ Tom_ 225
b_ Benedikt_ 225	b_ Benedikt_ 45_ b_ Martin_ 225	b_ Jonas_ 135_ d_ Kathrin_ 45_ b_ Lisa_ 225
b_ Imen_ 45	b_ Benedikt_ 45_ b_ Richard_ 225	b_ Jonas_ 135_ d_ Thomas_ 45_ b_ Alex_ 225
b_ Imen_ 135	b_ Benedikt_ 135_ b_ Martin_ 45	b_ Kathrin_ 135_ d_ Lisa_ 45_ b_ Thomas_ 225
b_ Imen_ 225	b_ Benedikt_ 135_ b_ Richard_ 45	b_ Kathrin_ 135_ d_ Ricarda_ 45_ b_ Andre_ 225
b_ Jonas_ 45	b_ Imen_ 45_ b_ Jonas_ 225	b_ Lisa_ 135_ d_ Thomas_ 45_ b_ Ricarda_ 225
b_ Jonas_ 135	b_ Imen_ 45_ b_ Kathrin_ 225	b_ Lisa_ 135_ d_ Tom_ 45_ b_ Benedikt_ 225
b_ Jonas_ 225	b_ Imen_ 135_ b_ Jonas_ 45	b_ Martin_ 135_ d_ Jonas_ 45_ b_ Thomas_ 225
b_ Kathrin_ 45	b_ Imen_ 135_ b_ Kathrin_ 45	b_ Martin_ 135_ d_ Richard_ 45_ b_ Imen_ 225
b_ Kathrin_ 135	b_ Jonas_ 45_ b_ Kathrin_ 225	b_ Ricarda_ 135_ d_ Andre_ 45_ b_ Richard_ 225
b_ Kathrin_ 225	b_ Jonas_ 45_ b_ Lisa_ 225	b_ Ricarda_ 135_ d_ Tom_ 45_ b_ Alex_ 225
b_ Lisa_ 45	b_ Jonas_ 135_ b_ Kathrin_ 45	b_ Richard_ 135_ d_ Imen_ 45_ b_ Jonas_ 225
b_ Lisa_ 135	b_ Jonas_ 135_ b_ Lisa_ 45	b_ Richard_ 135_ d_ Kathrin_ 45_ b_ Ricarda_ 225
b_ Lisa_ 225	b_ Kathrin_ 45_ b_ Lisa_ 225	b_ Thomas_ 135_ d_ Alex_ 45_ b_ Martin_ 225
b_ Martin_ 45	b_ Kathrin_ 45_ b_ Thomas_ 225	b_ Thomas_ 135_ d_ Ricarda_ 45_ b_ Tom_ 225
b_ Martin_ 135	b_ Kathrin_ 135_ b_ Lisa_ 45	b_ Tom_ 135_ d_ Alex_ 45_ b_ Andre_ 225
b_ Martin_ 225	b_ Kathrin_ 135_ b_ Thomas_ 45	b_ Tom_ 135_ d_ Benedikt_ 45_ b_ Imen_ 225
b_ Ricarda_ 45	b_ Lisa_ 45_ b_ Ricarda_ 225	
b_ Ricarda_ 135	b_ Lisa_ 45_ b_ Thomas_ 225	
b_ Ricarda_ 225	b_ Lisa_ 135_ b_ Ricarda_ 45	
b_ Richard_ 45	b_ Lisa_ 135_ b_ Thomas_ 45	
b_ Richard_ 135	b_ Martin_ 45_ b_ Imen_ 225	
b_ Richard_ 225	b_ Martin_ 45_ b_ Richard_ 225	
b_ Richard_ 45	b_ Martin_ 135_ b_ Imen_ 45	
b_ Thomas_ 45	b_ Martin_ 45_ b_ Richard_ 45	
b_ Thomas_ 135	b_ Ricarda_ 45_ b_ Alex_ 225	
b_ Thomas_ 225	b_ Ricarda_ 45_ b_ Tom_ 225	
b_ Tom_ 45	b_ Ricarda_ 135_ b_ Alex_ 45	
b_ Tom_ 135	b_ Ricarda_ 135_ b_ Tom_ 45	
b_ Tom_ 225	b_ Richard_ 45_ b_ Imen_ 225	
	b_ Richard_ 45_ b_ Jonas_ 225	
	b_ Richard_ 135_ b_ Imen_ 45	
	b_ Richard_ 135_ b_ Jonas_ 45	
	b_ Thomas_ 45_ b_ Ricarda_ 225	
	b_ Thomas_ 45_ b_ Tom_ 225	
	b_ Thomas_ 135_ b_ Ricarda_ 45	
	b_ Thomas_ 135_ b_ Tom_ 45	
	b_ Tom_ 45_ b_ Alex_ 225	
	b_ Tom_ 45_ b_ Andre_ 225	
	b_ Tom_ 135_ b_ Alex_ 45	
	b_ Tom_ 135_ b_ Andre_ 45	

Table A.2.: This combinations of the speakers were recorded in the audiolab and in the videolab with the three microphone arrays and for 10° and 20° elevation.

A.5. Content of the matlab files

framesize	1024 samples
radius	1.3m
tolerance	5°
location	'lab/' or 'office/'
elevation	10° or 20°
spider	'black' or 'black_ concha' or 'white'
nam	*file-name*, e.g. 'b_ Alex_ 135_ b_ Benedikt_ 45.wav'
fs	sampling frequency: 48000kHz
temperature	room temperature
sources	this vector contains the played source files
maxlocs	number of sources
beamformertime	duration of beamforming
locs_ pf	carthesian coordinates after beamformer and particle filter
meanxyz	mean values over all frames in carthesian coordinates
meansph	spherical meanvalue, calculated from *meanxyz*
notlocs	number of frames, in which the beamformer recognizes no source
index	order of the localized positions
angles	angles from the filename
maeazi	mean angular error azimuth over all frames
maeele	mean angular error elevation over all frames
tolazi	percentage of angles which are in the tolerance region in azimuth
tolele	percentage of angles which are in the tolerance region in elevation
output	contains the separated channels
gsstime	duration of the GSS
SDR	Source to Distortion ratio
SIR	Source to Interference ratio
SAR	Source to Artifacts ratio

Table A.3.: These data are the results of SRP-PHAT and GSS

fft_ points	1024 (as big as the framesize)
cost_ type	function to calculate the costs of iva
stft_ fun	function to calculate the transformation
iter	300 (maximum of iterations)
auto_ stop	1 or 0 (on or off)
info	index 1 belongs to all microphones, index 2 belongs to microphones [2 4 6 8]
location	'lab/' or 'office/'
elevation	10° or 20°
spider	'black' or 'black_ concha' or 'white'
nam	*file-name*, e.g. 'b_ Alex_ 135_ b_ Benedikt_ 45.wav'
fs	sampel frequency (48000kHz)
temperature	room temperature
sources	this vector contains the played source files
maxlocs	number of sources
iva_ time_ 8	duration of iva calculation with all 8 microphones
output_ 1	contains the separated channels with all 8 microphones
SDR_ 1	Source to Distortion ratio calculated with output_ 1
SIR_ 1	Source to Interference ratio calculated with output_ 1
SAR_ 1	Source to Artifacts ratio calculated with output_ 1
iva_ time_ 4	duration of iva calculation with 4 microphones
output_ 2	contains the separated channels with 4 microphones
SDR_ 2	Source to Distortion ratio calculated with output_ 2
SIR_ 2	Source to Interference ratio calculated with output_ 2
SAR_ 2	Source to Artifacts ratio calculated with output_ 2
cost_ 1	cost function for all 8 microphones
cost_ 2	cost function for 4 microphones

Table A.4.: These data are the results of IVA

radius	1.3m
Localization	'Localization with the angles from the file name
location	'lab/' or 'office/'
elevation	10° or 20°
spider	'black' or 'black_ concha' or 'white'
nam	*file-name*, e.g. 'b_ Alex_ 135_ b_ Benedikt_ 45.wav'
fs	sampel frequency $(48000$kHz$)$
temperature	room temperature
sources	this vector contains the played source files
maxlocs	number of sources
angles	angles from the filename
locs	coordinates from the file name
compass_ mask_ time	time of calculating binary masking
output	contains the separated channels
SDR_ 1	Source to Distortion ratio calculated with channel 2
SIR_ 1	Source to Interference ratio calculated with channel 2
SAR_ 1	Source to Artifacts ratio calculated with channel 2
SDR_ 2	Source to Distortion ratio calculated with channel 6
SIR_ 2	Source to Interference ratio calculated with channel 6
SAR_ 2	Source to Artifacts ratio calculated with channel 6

Table A.5.: These data are the results of binary masking

mics	[1,5;2,6;3,7;4,8]
location	'lab/' or 'office/'
elevation	$10°$ or $20°$
spider	'black' or 'black_ concha' or 'white'
nam	*file-name*, e.g. 'b_ Alex_ 135_ b_ Benedikt_ 45.wav'
fs	sampling frequency: 48000kHz
temperature	room temperature
sources	this vector contains the played source files
maxlocs	number of sources
loctime	duration of processing COMPaSS
locs	indices of IRs of the localized positions
coords	cartesian coordinates of the localized positions
probs	probabilities of the localized positions
parfitime	duration of processing the particle filter
tracked_ new	cartesian coordinates after the particle filter
meanxyz	mean values over all frames in carthesian coordinates
meansph	spherical meanvalue, calculated from *meanxyz*
notlocs	number of frames, in which the beamformer recognizes no source
index	order of the localized positions
angles	angles from the filename
maeazi	mean angular error azimuth over all frames
maeele	mean angular error elevation over all frames
tolazi	percentage of angles which are in the tolerance region in azimuth
tolele	percentage of angles which are in the tolerance region in elevation

Table A.6.: These data are the results of COMPaSS

A.6. 3D models

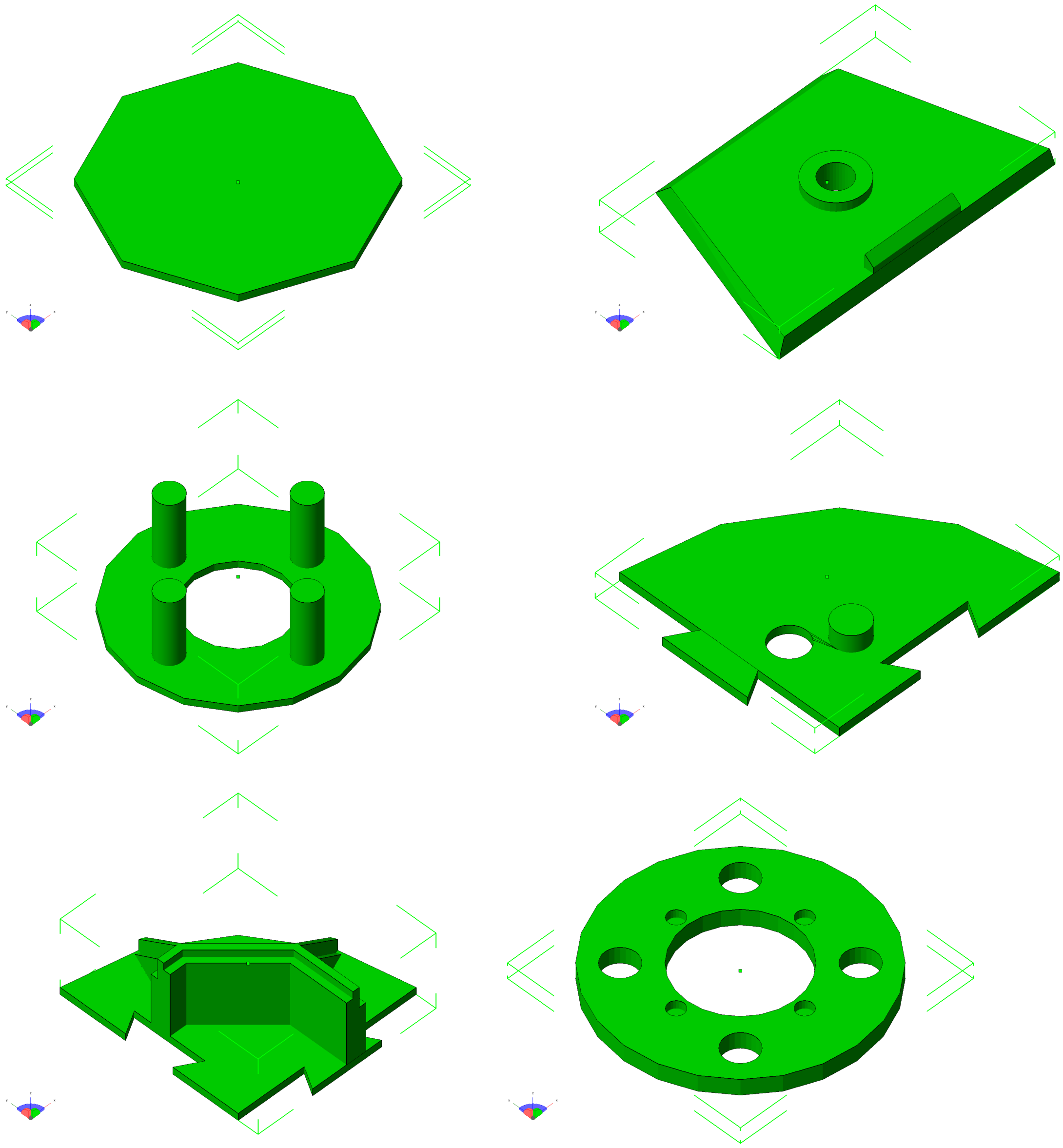


Figure A.13.: 3D-models of the white microphone array

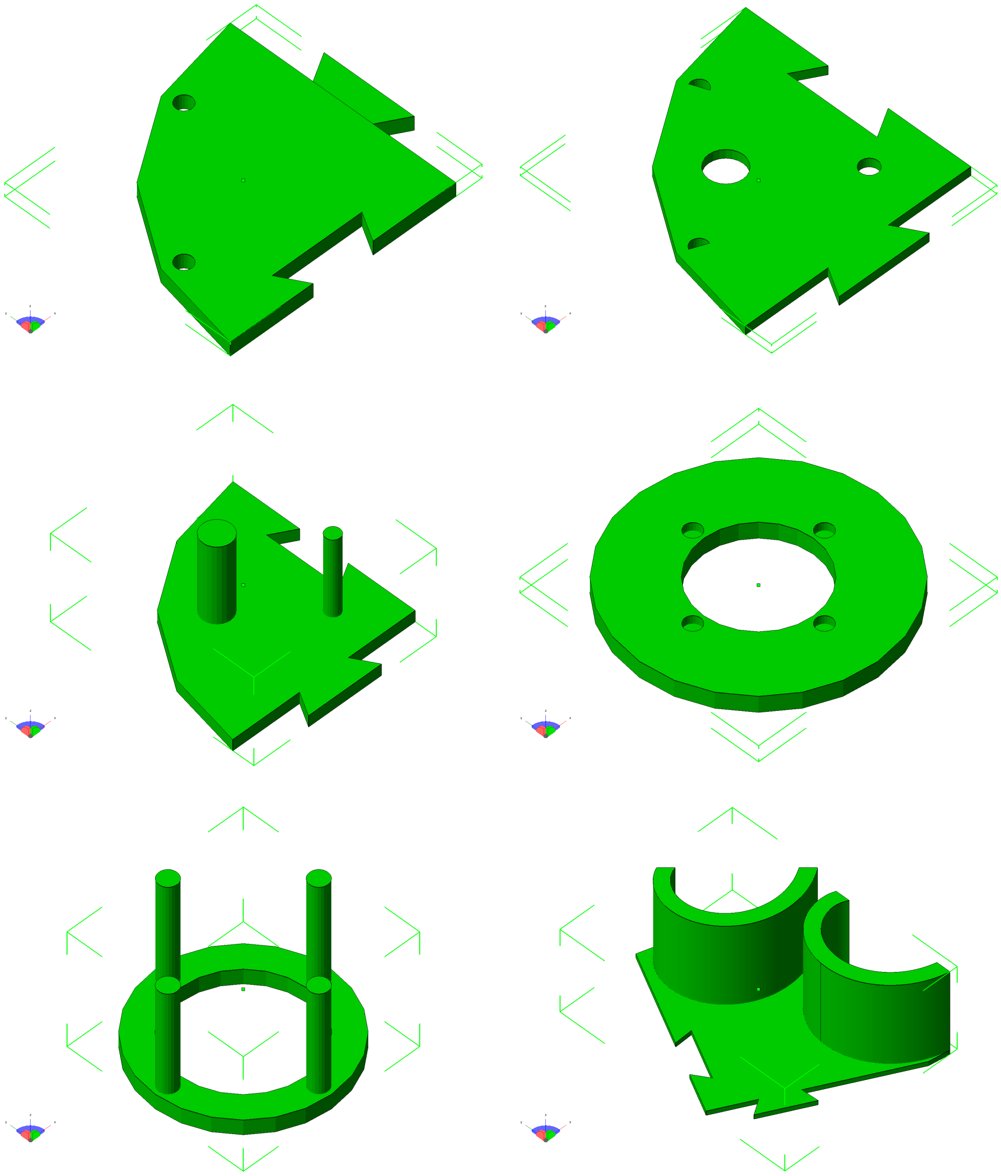


Figure A.14.: 3D-models of the black and of the black-concha microphone array

A.7. Localization results

Algorithm	Audiolab				Videolab				Simulation			
	MAEazi	**MAEele**	**TOLazi**	**TOLele**	**MAEazi**	**MAEele**	**TOLazi**	**TOLele**	**MAEazi**	**MAEele**	**TOLazi**	**TOLele**
spider: black, **elevation:** 10°												
one source												
COMPaSS	0.7°	1.4°	98.2%	91.4%	1.3°	6.6°	96.6%	32.2%	0.5°	0.5°	99.0%	99.0%
SRP-PHAT	0.22°	4.3°	99.3%	84.9%	1.0°	4.6°	97.6%	74.9%	0.8°	3.6°	99.3%	84.4%
two sources												
COMPaSS	1.8°	2.8°	95.6%	74.5%	6.8°	7.7°	90.8%	21.7%	0.6°	0.6°	94.7%	94.7%
SRP-PHAT	0.3°	3.8°	95.2%	82.3%	0.5°	4.1°	93.5%	74.5%	0.8°	6.0°	95.1%	64.0%
three sources												
COMPaSS	$-^\circ$	$-^\circ$	−%	−%	$-^\circ$	$-^\circ$	−%	−%	-	-	-	-
SRP-PHAT	5.8°	3.9°	82.5%	73.0%	2.9°	3.8°	83.3%	68.6%	9.5°	5.4°	82.2%	62.6%

Table A.7.: Comparison of the different localization approaches with recordings from the black microphone array at 10° elevation.

Algorithm	Audiolab				Videolab				Simulation			
	MAEazi	**MAEele**	**TOLazi**	**TOLele**	**MAEazi**	**MAEele**	**TOLazi**	**TOLele**	**MAEazi**	**MAEele**	**TOLazi**	**TOLele**
spider: black-concha, **elevation:** 10°												
one source												
COMPaSS	2.2°	0.9°	87.0%	95.7%	11.7°	4.3°	77.9%	59.6%	0.6°	0.6°	99.0%	99.0%
SRP-PHAT	4.2°	12.1°	65.4%	12.9%	6.4°	9.7°	47.1%	20.9%	5.2°	12.5°	39.5%	35.9%
two sources												
COMPaSS	7.3°	1.9°	74.1%	83.7%	50.7°	5.2°	46.4%	52.4%	0.6°	0.6°	92.8%	92.8%
SRP-PHAT	6.0°	10.0°	39.8%	18.7%	6.0°	9.1°	40.9%	21.3%	19.3°	21.5°	21.8%	20.3%
three sources												
COMPaSS	$-^\circ$	$-^\circ$	−%	−%	$-^\circ$	$-^\circ$	−%	−%	-	-	-	-
SRP-PHAT	11.8°	9.0°	30.6%	19.8%	14.1°	8.5°	24.3%	21.6%	35.0°	20.6°	19.1%	16.1%

Table A.8.: Comparison of the different localization approaches with recordings from the black-concha microphone array at 10° elevation.

Algorithm	Audiolab				Videolab				Simulation			
	MAEazi	**MAEele**	**TOLazi**	**TOLele**	**MAEazi**	**MAEele**	**TOLazi**	**TOLele**	**MAEazi**	**MAEele**	**TOLazi**	**TOLele**
spider: white, **elevation:** 10°												
one source												
COMPaSS	2.1°	3.8°	87.9%	64.5%	9.6°	8.6°	85.0%	12.5%	0.6°	0.6°	99.0%	99.0%
SRP-PHAT	0.2°	5.6°	99.3%	44.8%	1.8°	7.2°	96.4%	15.0%	1.9°	6.4°	99.3%	0.2%
two sources												
COMPaSS	13.6°	3.6°	73.3%	66.9%	16.5°	8.8°	77.9%	11.4%	0.6°	0.6°	94.4%	94.3%
SRP-PHAT	0.6°	6.4°	95.2%	31.2%	0.8°	7.0°	91.9%	14.6%	2.7°	6.6°	88.5%	9.2%
three sources												
COMPaSS	$-^\circ$	$-^\circ$	−%	−%	$-^\circ$	$-^\circ$	−%	−%	-	-	-	-
SRP-PHAT	6.5°	7.2°	82.6%	18.7%	10.3°	6.7°	74.3%	17.7%	9.9°	7.0°	73.0%	5.5%

Table A.9.: Comparison of the different localization approaches with recordings from the white microphone array at 10° elevation.

Algorithm	Audiolab				Videolab				Simulation			
	MAEazi	**MAEele**	**TOLazi**	**TOLele**	**MAEazi**	**MAEele**	**TOLazi**	**TOLele**	**MAEazi**	**MAEele**	**TOLazi**	**TOLele**
spider: black, **elevation:** 20°												
one source												
COMPaSS	1.1°	1.5°	96.2%	89.6%	1.7°	6.9°	94.1%	44.6%	0.6°	0.6°	99.0%	99.0%
SRP-PHAT	0.8°	2.5°	97.1%	96.5%	0.8°	2.3°	98.1%	97.1%	1.8°	4.8°	98.5%	54.7%
two sources												
COMPaSS	11.5°	3.0°	87.3%	77.9%	10.7°	10.9°	88.0%	30.9%	0.6°	0.6°	93.8%	93.8%
SRP-PHAT	0.6°	2.5°	94.4%	90.3%	0.6°	2.3°	92.7%	89.3%	2.4°	10.2°	91.8%	43.0%
three sources												
COMPaSS	–°	–°	–%	–%	–°	–°	–%	–%	-	-	-	-
SRP-PHAT	7.1°	2.3°	80.9%	83.4%	8.5°	2.4°	77.6%	80.6%	5.3°	12.6°	80.7%	41.2%

Table A.10.: Comparison of the different localization approaches with recordings from the black microphone array at 20° elevation.

Algorithm	Audiolab				Videolab				Simulation			
	MAEazi	**MAEele**	**TOLazi**	**TOLele**	**MAEazi**	**MAEele**	**TOLazi**	**TOLele**	**MAEazi**	**MAEele**	**TOLazi**	**TOLele**
spider: black-concha, **elevation:** 20°												
one source												
COMPaSS	0.7°	0.6°	97.9%	97.9%	1.1°	0.7°	97.3%	97.9%	0.6°	0.6°	99.0%	99.0%
SRP-PHAT	3.9°	20.2°	85.2%	0.39%	4.3°	19.1°	74.3%	2.5%	4.8°	16.3°	46.4%	10.4%
two sources												
COMPaSS	0.8°	0.7°	95.6%	95.6%	7.0°	1.3°	88.4%	92.6%	0.6°	0.6°	93.9%	93.9%
SRP-PHAT	4.0°	18.6°	63.2%	2.8%	5.1°	17.2°	45.5%	4.5%	8.9°	16.3°	41.7%	12.3%
three sources												
COMPaSS	–°	–°	–%	–%	–°	–°	–%	–%	-	-	-	-
SRP-PHAT	11.1°	18.3°	45.1%	2.7%	13.3°	17.1°	29.1%	3.9%	29.8°	18.7°	30.5%	6.5%

Table A.11.: Comparison of the different localization approaches with recordings from the black-concha microphone array at 20° elevation.

Algorithm	Audiolab				Videolab				Simulation			
	MAEazi	**MAEele**	**TOLazi**	**TOLele**	**MAEazi**	**MAEele**	**TOLazi**	**TOLele**	**MAEazi**	**MAEele**	**TOLazi**	**TOLele**
spider: white, **elevation:** 20°												
one source												
COMPaSS	1.1°	1.7°	94.9%	88.1%	9.4°	7.9°	85.5%	41.6%	0.6°	0.6°	99.0%	99.0%
SRP-PHAT	2.1°	16.3°	96.9%	0.47%	1.6°	16.2°	97.7%	0.57%	1.3°	15.3°	99.3	0%
two sources												
COMPaSS	4.7°	3.0°	94.2%	82.1%	9.7°	10.2°	83.6%	39.5%	0.6°	0.6°	95.3%	95.3%
SRP-PHAT	1.0°	17.3°	95.4%	0.1%	1.2°	17.2°	93.8%	0.1%	1.5°	15.8°	95.9%	0%
three sources												
COMPaSS	–°	–°	–%	–%	–°	–°	–%	–%	-	-	-	-
SRP-PHAT	2.5°	17.8°	85.2%	0.01%	7.5°	17.4°	78.5%	0.1%	8.9°	16.7°	78.8%	0%

Table A.12.: Comparison of the different localization approaches with recordings from the white microphone array at 20° elevation.

A.8. Separation results

Algorithm	Audiolab			Videolab			Simulation		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
spider: black, **elevation:** 10°									
one source									
GSS	17.3	Inf	17.3	5.7	Inf	5.7	-4.6	Inf	-4.6
IVA [1 2 3 4 5 6 7 8]	9.7	Inf	9.7	2.0	Inf	2.0	16.0	Inf	16.0
IVA [2 4 6 8]	9.7	Inf	9.7	1.9	Inf	1.9	15.9	Inf	15.9
Binary Masking	9.3	Inf	9.3	1.5	Inf	1.5	15.4	Inf	15.4
two sources									
GSS	12.1	15.7	15.0	4.2	12.1	5.3	-6.0	10.7	-5.3
IVA [1 2 3 4 5 6 7 8]	5.3	11.6	8.5	-0.7	7.8	1.4	10.2	15.6	13.3
IVA [2 4 6 8]	5.9	11.6	8.8	-1.1	7.1	1.0	9.8	15.1	13.1
Binary Masking	3.2	14.6	3.9	-1.9	12.4	5.5	5.4	16.3	6.1
three sources									
GSS	9.8	13.4	12.9	2.9	9.8	4.5	-6.7	7.0	-5.5
IVA [1 2 3 4 5 6 7 8]	4.4	9.2	8.0	-1.7	4.9	1.4	6.7	10.0	12.0
IVA [2 4 6 8]	4.1	8.3	8.2	-2.0	4.4	1.2	6.2	9.2	11.4
Binary Masking	0.3	10.4	1.3	-3.7	8.4	-2.4	2.4	12.4	3.4

Table A.13.: Comparison of the different separation approaches with recordings from the black microphone array at 10° elevation. All values are given in dB.

Algorithm	Audiolab			Videolab			Simulation		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
spider: black-concha, **elevation:** 10°									
one source									
GSS	18.2	Inf	18.2	6.8	Inf	6.8	-4.5	Inf	-4.5
IVA [1 2 3 4 5 6 7 8]	10.0	Inf	10.0	3.3	Inf	3.3	16.6	Inf	16.6
IVA [2 4 6 8]	9.9	Inf	9.9	3.4	Inf	3.4	16.6	Inf	16.6
Binary Masking	9.6	Inf	9.6	2.7	Inf	2.7	16.4	Inf	16.4
two sources									
GSS	12.0	15.2	15.1	5.0	12.2	6.3	-5.8	10.5	-5.2
IVA [1 2 3 4 5 6 7 8]	6.9	14.6	9.4	0.8	9.5	2.9	10.8	16.7	13.9
IVA [2 4 6 8]	6.3	12.2	9.2	0.2	8.2	2.3	10.0	14.6	14.0
Binary Masking	3.8	15.6	4.4	-0.4	13.6	0.2	6.0	17.2	6.6
three sources									
GSS	9.1	11.6	13.2	3.4	9.2	5.3	-7.1	6.0	-5.6
IVA [1 2 3 4 5 6 7 8]	5.5	11.1	8.6	-0.1	7.1	2.4	9.8	12.9	13.9
IVA [2 4 6 8]	4.6	9.2	8.3	-1.1	5.3	2.0	7.3	10.0	12.7
Binary Masking	0.6	10.9	1.6	-2.7	8.9	-1.5	2.7	12.9	3.5

Table A.14.: Comparison of the different separation approaches with recordings from the black-concha microphone array at 10° elevation. All values are given in dB.

Algorithm	Audiolab			Videolab			Simulation		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
spider: white, **elevation:** 10°									
one source									
GSS	16.8	Inf	16.8	5.2	Inf	5.2	-3.9	Inf	-3.9
IVA [1 2 3 4 5 6 7 8]	9.5	Inf	9.5	1.8	Inf	1.8	15.7	Inf	15.7
IVA [2 4 6 8]	9.6	Inf	9.6	2.0	Inf	2.0	16.0	Inf	16.0
Binary Masking	9.0	Inf	9.0	1.4	Inf	1.4	15.3	Inf	15.3
two sources									
GSS	11.1	15.1	13.8	4.0	11.9	5.1	-5.3	11.6	-4.8
IVA [1 2 3 4 5 6 7 8]	6.0	13.7	8.7	-0.6	8.1	1.6	10.3	16.3	13.5
IVA [2 4 6 8]	6.2	12.6	9.1	-0.8	7.4	1.3	10.1	15.6	13.5
Binary Masking	2.8	14.0	3.6	-2.1	11.8	-1.3	5.3	16.2	5.9
three sources									
GSS	9.0	12.3	12.2	2.9	9.7	4.5	-6.4	7.7	-5.4
IVA [1 2 3 4 5 6 7 8]	4.0	9.1	8.0	-1.9	4.8	1.3	9.6	12.8	13.8
IVA [2 4 6 8]	4.2	8.7	8.2	-2.2	4.1	1.1	8.1	11.4	12.5
Binary Masking	-0.1	9.9	1.1	-4.1	7.6	-2.6	2.1	12.2	3.0

Table A.15.: Comparison of the different separation approaches with recordings from the white microphone array at 10° elevation. All values are given in dB.

Algorithm	Audiolab			Videolab			Simulation		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
spider: black, **elevation:** 20°									
one source									
GSS	16.0	Inf	16.0	6.5	Inf	6.5	-4.6	Inf	-4.6
IVA [1 2 3 4 5 6 7 8]	9.2	Inf	9.2	2.8	Inf	2.8	16.9	Inf	16.9
IVA [2 4 6 8]	8.9	Inf	8.9	2.7	Inf	2.7	17.0	Inf	17.0
Binary Masking	8.4	Inf	8.4	2.1	Inf	2.1	16.6	Inf	16.6
two sources									
GSS	10.8	13.6	14.5	4.7	11.9	6.0	-6.0	10.9	-5.4
IVA [1 2 3 4 5 6 7 8]	5.2	10.6	8.1	0.0	8.3	2.0	10.8	16.0	14.2
IVA [2 4 6 8]	4.4	9.4	7.9	-0.5	7.4	1.5	9.8	14.6	13.6
Binary Masking	3.0	14.5	3.6	-1.0	12.8	-0.4	5.6	16.3	6.3
three sources									
GSS	8.7	11.4	13.0	3.3	10.0	5.0	-6.9	7.0	-5.7
IVA [1 2 3 4 5 6 7 8]	3.7	8.6	7.7	-1.0	5.7	1.8	8.1	11.7	12.8
IVA [2 4 6 8]	4.1	9.1	7.6	-1.8	4.4	1.3	7.5	10.6	12.4
Binary Masking	0.2	10.5	1.2	-3.1	8.5	-1.8	2.6	12.5	3.4

Table A.16.: Comparison of the different separation approaches with recordings from the black microphone array at 20° elevation. All values are given in dB.

Algorithm	Audiolab			Videolab			Simulation		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
spider: black-concha, **elevation:** 20°									
one source									
GSS	17.1	Inf	17.1	7.6	Inf	7.6	-5.5	Inf	-5.5
IVA [1 2 3 4 5 6 7 8]	9.5	Inf	9.5	3.7	Inf	3.7	17.7	Inf	17.7
IVA [2 4 6 8]	9.1	Inf	9.1	3.7	Inf	3.7	17.9	Inf	17.9
Binary Masking	8.6	Inf	8.6	3.0	Inf	3.0	17.6	Inf	17.6
two sources									
GSS	11.5	14.7	14.8	5.7	12.9	6.9	-6.9	10.5	-6.3
IVA [1 2 3 4 5 6 7 8]	6.0	11.9	8.6	1.3	9.7	3.3	11.9	17.5	15.2
IVA [2 4 6 8]	6.3	12.4	8.8	0.4	8.2	2.8	11.9	16.9	15.3
Binary Masking	2.9	14.3	3.6	-0.6	12.5	0.1	6.0	16.9	6.6
three sources									
GSS	9.2	11.7	13.1	3.9	10.0	5.7	-7.6	6.7	-6.4
IVA [1 2 3 4 5 6 7 8]	6.8	12.8	9.4	1.3	9.2	3.2	11.0	14.1	15.3
IVA [2 4 6 8]	3.6	8.0	7.6	-1.7	4.8	1.4	7.0	9.6	12.8
Binary Masking	-0.2	9.9	1.0	-2.9	8.2	-1.5	2.8	13.0	3.6

Table A.17.: Comparison of the different separation approaches with recordings from the black-concha microphone array at 20° elevation. All values are given in dB.

Algorithm	Audiolab			Videolab			Simulation		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
spider: white, **elevation:** $20°$									
one source									
GSS	15.6	Inf	15.6	6.2	Inf	6.2	-4.3	Inf	-4.3
IVA [1 2 3 4 5 6 7 8]	8.5	Inf	8.5	2.3	Inf	2.3	16.7	Inf	16.7
IVA [2 4 6 8]	8.6	Inf	8.6	2.6	Inf	2.6	17.0	Inf	17.0
Binary Masking	7.9	Inf	7.9	1.9	Inf	1.9	16.5	Inf	16.5
two sources									
GSS	10.9	14.6	13.6	4.7	12.4	5.9	-5.7	11.4	-5.3
IVA [1 2 3 4 5 6 7 8]	4.5	10.3	7.7	-0.2	8.2	2.1	10.3	15.5	14.0
IVA [2 4 6 8]	5.1	10.6	8.0	-0.3	7.4	2.0	9.4	13.9	13.6
Binary Masking	2.4	13.6	3.2	-1.6	11.7	-0.8	5.3	16.0	5.9
three sources									
GSS	8.9	12.1	12.2	3.3	10.1	4.9	-6.4	7.9	-5.5
IVA [1 2 3 4 5 6 7 8]	4.8	11.1	7.9	-0.1	8.0	2.0	8.9	12.2	13.3
IVA [2 4 6 8]	2.4	6.4	7.3	-2.3	5.0	0.6	6.4	8.9	12.1
Binary Masking	-0.6	9.6	0.7	-3.8	7.7	-2.3	2.2	12.1	3.1

Table A.18.: Comparison of the different separation approaches with recordings from the white microphone array at $20°$ elevation. All values are given in dB.

A.9. Data sheets

img Stage Line®

EMA-300P

Best.-Nr. 23.2410

D A CH Phantomspeisungsadapter

Bitte lesen Sie diese Bedienungsanleitung vor dem Betrieb gründlich durch und heben Sie sie für ein späteres Nachlesen auf.

1 Einsatzmöglichkeiten

Der Adapter EMA-300P ist speziell für den Betrieb der Elektret-Mikrofone ECM-300B und ECM-300L aus dem Programm von „img Stage Line“ konzipiert. Er ermöglicht den Anschluss dieser Mikrofone an ein Audiogerät (z. B. Mischpult, Verstärker), das über eine Phantomspeisung von 9 – 48 V⎓ verfügt.

2 Hinweise für den sicheren Gebrauch

Der Adapter entspricht allen relevanten Richtlinien der EU und ist deshalb mit CE gekennzeichnet.

- Setzen Sie den Adapter nur im Innenbereich ein und schützen Sie es vor Tropf- und Spritzwasser, hoher Luftfeuchtigkeit und Hitze (zulässiger Einsatztemperaturbereich 0 – 40 °C).
- Verwenden Sie für die Reinigung nur ein trockenes, weiches Tuch, auf keinen Fall Chemikalien oder Wasser.
- Wird der Adapter zweckentfremdet, falsch angeschlossen oder nicht fachgerecht repariert, kann keine Haftung für daraus resultierende Sach- oder Personenschäden und keine Garantie für das Mikrofon übernommen werden.

Soll der Adapter endgültig aus dem Betrieb genommen werden, übergeben Sie es zur umweltgerechten Entsorgung einem örtlichen Recyclingbetrieb.

3 Technische Daten

Eingang:	Mini-XLR-Stecker, asymmetrisch
Ausgang:	XLR-Stecker, symmetrisch
Einsatztemperatur:	0 – 40 °C
Abmessungen, Gewicht:	∅ 19 mm × 92 mm, 80 g
Stromversorgung:	Phantomspeisung 9 – 48 V⎓

Änderungen vorbehalten.

GB Phantom Power Adapter

Please read these operating instructions carefully prior to operation and keep them for later use.

3 Applications

The adapter EMA-300P is especially designed for operation of the electret microphones ECM-300B and ECM-300L of the “img Stage Line” range. It allows to connect these microphones to an audio unit (e. g. mixer, amplifier) which is provided with a phantom power of 9 – 48 V⎓.

2 Safety Notes

The adapter corresponds to all relevant directives of the EU and is therefore marked with CE.

- The adapter is suitable for indoor use only. Protect it against dripping water and splash water, high air humidity and heat (admissible ambient temperature range 0 – 40 °C).
- For cleaning only use a dry, soft cloth, never use chemicals or water.
- No guarantee claims for the adapter and no liability for any resulting personal damage or material damage will be accepted if it is used for other purposes than originally intended, if it is not correctly connected or not repaired in an expert way.

If the adapter is to be put out of operation definitively, take it to a local recycling plant for a disposal which is not harmful to the environment.

3 Specifications

Input:	mini XLR plug, unbalanced
Output:	XLR plug, balanced
Ambient temperature:	0 – 40 °C
Dimensions, weight:	∅ 19 mm × 92 mm, 80 g
Power supply:	phantom power 9 – 48 V⎓

Subject to technical modification.

F B CH Adaptateur d’alimentation fantôme

Veuillez lire la présente notice avec attention avant le fonctionnement et conservez-la pour pouvoir vous y reporter ultérieurement.

1 Possibilités d’utilisation

L’adaptateur EMA-300P est spécialement conçu pour le fonctionnement des microphones électret ECM-300B et ECM-300L de la gamme “img Stage Line”. Il permet de brancher ces microphones à un appareil audio (par exemple table de mixage, amplificateur), disposant d’une alimentation fantôme 9 – 48 V⎓.

2 Conseils de sécurité

L’adaptateur répond à toutes les directives nécessaires de l’Union Européenne et porte donc le symbole CE.

- L’adaptateur n’est conçu que pour une utilisation en intérieur. Protégez-le de tout type de projections d’eau, des éclaboussures, d’une humidité élevée et de la chaleur (plage de température de fonctionnement autorisée : 0 – 40 °C).
- Pour le nettoyer, utilisez uniquement un chiffon sec et doux, en aucun cas de produits chimiques ou d’eau.
- Nous déclinons toute responsabilité en cas de dommages matériels ou corporels résultants si l’adaptateur est utilisé dans un but autre que celui pour lequel il a été conçu, s’il n’est pas correctement branché ou s’il n’est pas réparé par une personne habilitée ; en outre, la garantie deviendrait caduque.

Lorsque l’adaptateur est définitivement retiré du service, vous devez le déposer dans une usine de recyclage adaptée pour contribuer à son élimination non polluante.

3 Caractéristiques techniques

Entrée :	fiche mini XLR, asymétrique
Sortie :	fiche XLR, symétrique
Température d’utilisation :	0 – 40°C
Dimensions, poids :	∅ 19 mm × 92 mm, 80 g
Alimentation :	alimentation fantôme 9 – 48 V⎓

Tout droit de modification réservé.

I Adattatore di alimentazione phantom

Vi preghiamo di leggere attentamente le presenti istruzioni prima della messa in funzione e di conservarle per un uso futuro.

1 Possibilità d’impiego

L’adattatore EMA-300P è stato realizzato per il funzionamento dei microfoni all’elettrete ECM-300B e ECM-300L del programma “img Stage Line”. Permette il collegamento di questi microfoni con un apparecchio audio (p. es. mixer, amplificatore) equipaggiato con alimentazione phantom (9 – 48 V⎓).

2 Avvertenze di sicurezza

L’adattatore è conforme a tutte le direttive rilevanti dell’UE e pertanto porta la sigla CE.

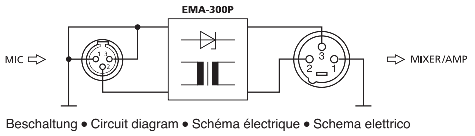
- Usare l’adattator solo all’interno di locali e proteggerlo dall’acqua gocciolante e dagli spruzzi d’acqua, da alta umidità dell’aria e dal calore (temperatura d’impiego ammessa fra 0 e 40 °C).
- Per la pulizia usare solo un panno morbido, asciutto; non impiegare in nessun caso prodotti chimici o acqua.
- Nel caso d’uso improprio, di collegamenti sbagliati o di riparazione non a regola d’arte dell’ adattatore, non si assume nessuna responsabilità per eventuali danni consequenziali a persone o a cose e non si assume nessuna garanzia per il microfono.

Se si desidera eliminare l’adattatore definitivamente, consegnarlo per lo smaltimento ad un’istituzione locale per il riciclaggio.

3 Dati tecnici

Ingresso:	connettore mini-XLR, asimmetrico
Uscita:	connettore XLR, simmetrico
Temperatura d’impiego:	0 – 40°C
Dimensioni, peso:	∅ 19 mm × 92 mm, 80 g
Alimentazione:	alimentazione phantom 9 – 48 V⎓

Con riserva di modifiche tecniche.

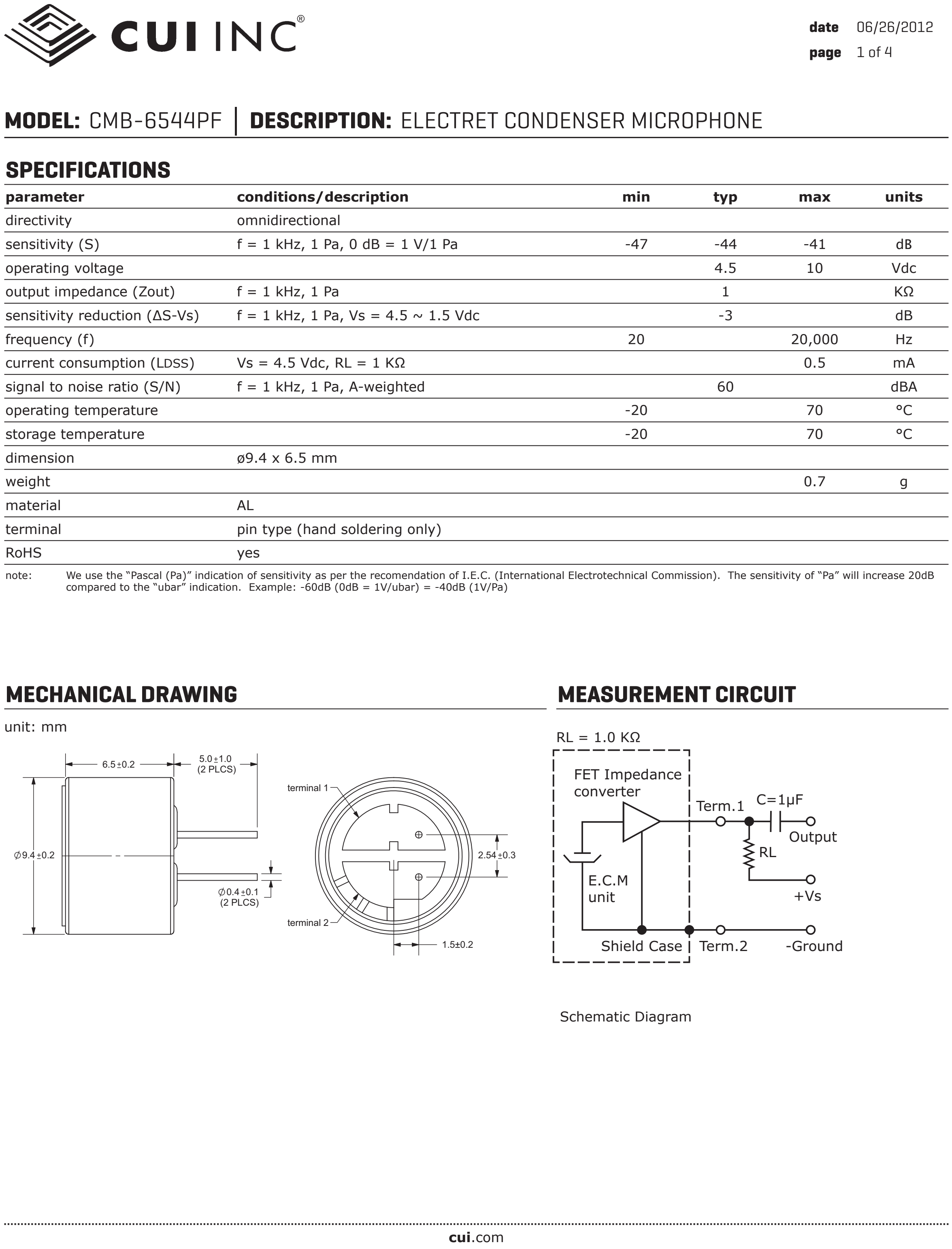


Beschaltung • Circuit diagram • Schéma électrique • Schema elettrico

www.imgstageline.com

MONACOR® INTERNATIONAL
MONACOR INTERNATIONAL GmbH & Co. KG • Zum Falsch 36 • 28307 Bremen • Germany

A-0979.99.02.06.2011



CUI INC®

date 06/26/2012
page 1 of 4

MODEL: CMB-6544PF | DESCRIPTION: ELECTRET CONDENSER MICROPHONE

SPECIFICATIONS

parameter	conditions/description	min	typ	max	units
directivity	omnidirectional				
sensitivity (S)	f = 1 kHz, 1 Pa, 0 dB = 1 V/1 Pa	-47	-44	-41	dB
operating voltage			4.5	10	Vdc
output impedance (Zout)	f = 1 kHz, 1 Pa		1		KΩ
sensitivity reduction (ΔS-Vs)	f = 1 kHz, 1 Pa, Vs = 4.5 ~ 1.5 Vdc		-3		dB
frequency (f)		20		20,000	Hz
current consumption (LDSS)	Vs = 4.5 Vdc, RL = 1 KΩ			0.5	mA
signal to noise ratio (S/N)	f = 1 kHz, 1 Pa, A-weighted		60		dBA
operating temperature		-20		70	°C
storage temperature		-20		70	°C
dimension	ø9.4 x 6.5 mm				
weight				0.7	g
material	AL				
terminal	pin type (hand soldering only)				
RoHS	yes				

note: We use the "Pascal (Pa)" indication of sensitivity as per the recomendation of I.E.C. (International Electrotechnical Commission). The sensitivity of "Pa" will increase 20dB compared to the "ubar" indication. Example: -60dB (0dB = 1V/ubar) = -40dB (1V/Pa)

MECHANICAL DRAWING

unit: mm

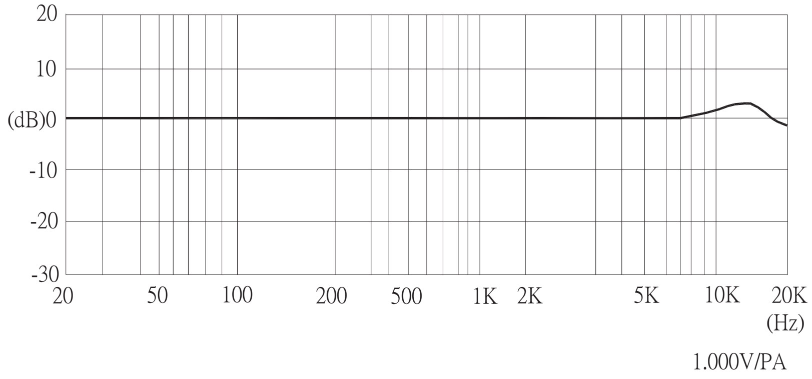
MEASUREMENT CIRCUIT

RL = 1.0 KΩ

Schematic Diagram

cui.com

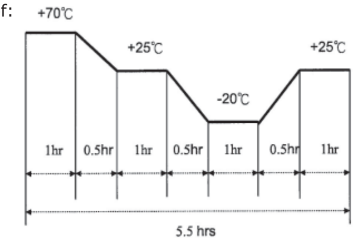
FREQUENCY RESPONSE CURVE



MECHANICAL CHARACTERISTICS

item	test condition	evaluation standard
soldering heat resistance	Soldering iron of +270 ±5°C should be placed on the terminal for 2 ±0.5 seconds.	No interference in operation.
PCB wire pull strength	The pull force should be applid to double lead wire: Horizontal 4.9 N (0.5 kg) for 30 seconds	No damage or cutting off.
vibration test	The part should be measured after a vibration amplitude of 1.5 mm with 10~55 Hz band of vibration frequency to each of the 3 perpendicular directions for 2 hours.	After any tests, the sensitivity should be within ±3 dB of the initial sensitivity.
drop test	The part without packaging is subjected to 3 drops on each axis from the height of 1 m onto a 20 mm thick wooden board.	

ENVIRONMENT TEST

item	test condition	evaluation standard
high temperature test	After being placed in a chamber at +70°C for 72 hours.	After any tests and 6 hours of conditioning at +25°C, the sensitivity should be within ±3 dB of the initial sensitivity.
low temperature test	After being placed in a chamber at -20°C for 72 hours.	
thermal shock	After being placed in a chamber at +40°C and 90 ±5% RH for 240 hours.	
temperature cycle test	The part will be subjected to 10 cycles. One cycle will consist of: 	

TEST CONDITIONS

standard test conditions	a) Temperature: +5 ~ +35°C	b) Humidity: 45 ~ 85%	c) Pressure: 860 ~ 1060 mbar
judgement test conditions	a) Temperature: +25 ±2°C	b) Humidity: 60 ~ 70%	c) Pressure: 860 ~ 1060 mbar

List of Acronyms

BSS	Blind Source Separation	8
CCL	Cross Convolution Localization	17
COMPaSS	loCalization Of MultiPle Sound Sources	18
DFT	Discrete-Fourier-Transformation	15
DUET	Degenerate Unmixing Estimation Technique	19
GSS	Geometric Source Separation	21
HRTFs	Head Related Transfer Functions	7
ICA	Independent Component Analysis	5
IR	Impulse Response	5
IVA	Independent Vector Analysis	5
MAE	Mean Angular Error	30
MLS	Maximum Length Sequence	31
PCA	Principle Component Analysis	6
SAR	Sources to Artifacts Ratio	30
SDR	Source to Distortion Ratio	29
SIR	Source to Interference Ratio	29
SRP-PHAT	Steered Response Power - Phase Transform	20
STFT	Short-Time-Fourier-Transformation	15
TF	Transfer Function	5

Bibliography

[1] NewTec (Design : Audio. Visited on March 12, 2013.

[2] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174 – 188, 2002.

[3] J. Benesty, M.M. Sondhi, and Y. Huang. *Springer Handbook of Speech Processing*. Springer-Verlag Berlin Heidelberg, 2008.

[4] C. Denk and M. Rothbucher. Robotic sound source separation using independent vector analysis, 2011. Project thesis at the *Institute for Data Processing, Technische Universität München*.

[5] Marko Durkovic. Localization, tracking, and separation of sound sources for cognitive robots, 2012. Ph.D. thesis at the *Institute for Data Processing, Technische Universität München*.

[6] Marko Durkovic, Tim Habigt, Martin Rothbucher, and Klaus Diepold. Low latency localization of multiple sound sources in reverberant environments. *Journal of the Acoustical Society of America Express Letters*, 130(6):EL392–EL398, 2011.

[7] Johannes Feldmaier. Sound localization and separation for teleconferencing systems, 2011. Diploma thesis at the *Institute for Data Processing, Technische Universität München*.

[8] J. Hao, I. Lee, T.W. Lee, and T.J. Sejnowski. Independent vector analysis for source separation using a mixture of gaussians prior. *Neural Computation*, 22(6):1646 –1673, 2010.

[9] Shunichiro Hirayanagi and Nozomu Hamada. A solution for the permutation proplem of overdetermined source separation using subspace method. In *International Workshop on Acoustic Echo and Noise Control*, pages 101–104, 2005.

[10] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, Inc. New York, 2001.

[11] Christoph Dominik Kozielski. Online speaker recognition for teleconferencing systems, 2011. Diploma thesis at the *Institute for Data Processing, Technische Universität München*.

[12] E. Lopez-Poveda and R. Meddis. A physical model of sound diffraction and reflections in the human concha. *Journal of the Acoustical Society of America*, 100(5):3248 – 3259, 1996.

[13] L.C. Parra and C.V. Alvino. Geometric source separation: Merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, 10(6):352 – 362, 2002.

[14] Tobias Plutka. Extension of a binaural localization and tracking algorithm, 2012. Bachelor thesis at the *Institute for Data Processing, Technische Universität München*.

[15] Scott Richard. *Blind Speech Separation*, chapter The DUET blind source separation algorithm, pages 217–241. Springer Netherlands, 2007.

[16] Scott Rickard and Maurice Fallon. The gini index of speech. *Annual Conference on Information Sciences and Systems*, page 5, 2004.

[17] Martin Rothbucher, Matthias Kaufmann, Johannes Feldmaier, Tim Habigt, Marko Durkovic, Christoph Kozielski, and Klaus Diepold. 3d audio conference system with backward compatible conference server using hrtf synthesis. *Journal of Multimedia Processing and Technologies*, 2(4):159–175, 2011.

[18] Michael Unvervdorben. Blind source separation for speaker recognition systems, 2012. Diploma thesis at the *Institute for Data Processing, Technische Universität München*.

[19] Muhammad Usman, Fakheredine Keyrouz, and Klaus Diepold. Real time humanoid sound source localization and tracking in a highly reverberant environment. *International Conference on Signal Processing*, pages 2661–2664, 2008.

[20] J.M. Valin, F. Michaud, and J. Rouat. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems*, 55(3):216 – 228, 2007.

[21] J.M. Valin, J. Rouat, and F. Michaud. Enhanced robot audition based on microphone array source separation with post-filter. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pages 2123 – 2128, 2004.

[22] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462 –1469, 2006.

[23] Emmanuel Vincent. BSS Eval A toolbox for performance measurement in (blind) source separation. Online. Last visit: 16.02.2013.