5

HRTF Sound Localization

Martin Rothbucher, David Kronmüller, Marko Durkovic, Tim Habigt and Klaus Diepold
Institute for Data Processing, Technische Universität München
Germany

1. Introduction

In order to improve interactions between the human (operator) and the robot (teleoperator) in human centered robotic systems, e.g. Telepresence Systems as seen in Figure 1, it is important to equip the robotic platform with multimodal human-like sensing, e.g. vision, haptic and audition.

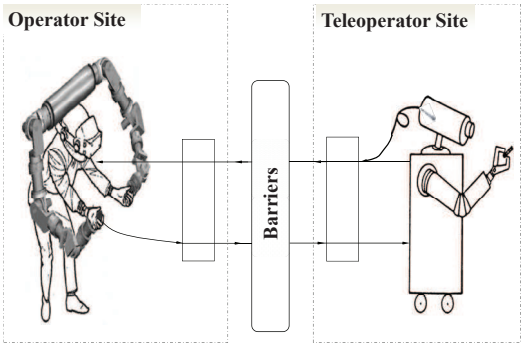


Fig. 1. Schematic view of the telepresence scenario.

Recently, robotic binaural hearing approaches based on Head-Related Transfer Functions (HRTFs) have become a promising technique to enable sound localization on mobile robotic platforms. Robotic platforms would benefit from this human like sound localization approach because of its noise-tolerance and the ability to localize sounds in a three-dimensional environment with only two microphones.
As seen in Figure 2, HRTFs describe spectral changes of sound waves when they enter the ear canal, due to diffraction and reflection of the human body, i.e. the head, shoulders, torso and ears. In far field applications, they can be considered as functions of two spatial variables (elevation and azimuth) and frequency. HRTFs can be regarded as direction dependent filters, as diffraction and reflexion properties of the human body are different for each direction. Since

the geometric features of the body differ from person to person, HRTFs are unique for each individual (Blauert, 1997).

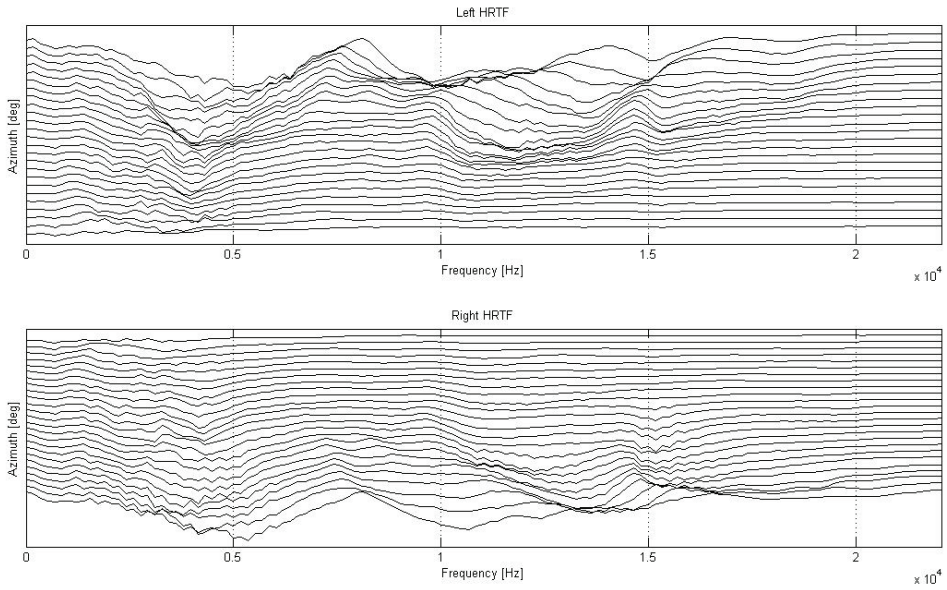


Fig. 2. HRTFs over varying azimuth and constant elevation

The problem of HRTF-based sound localization on mobile robotic platforms can be separated into three main parts, namely the HRTF-based localization algorithms, the HRTF data reduction and the application of predictors that improve the localization performance.
For robotic HRTF-based localization, an incoming sound signal is reflected, diffracted and scattered by the robot's torso, shoulders, head and pinnae, dependent on the direction of the sound source. Thus both left and right perceived signals have been altered through the robot's HRTF, which the robot has learned to associate with a specific direction. We have investigated several HRTF-based sound localization algorithms, which are compared in the first section.
Due to its high dimensionality, it is inefficient to utilize the robot's original HRTFs. Therefore, the second section will provide a comparison of HRTF reduction techniques. Once the HRTF dataset has been reduced and restored, it serves as the basis for localization.
HRTF localization is computational very expensive, therefore, it is advantageous to reduce the search region for sound sources to a region of interest (ROI). Given a HRTF dataset, it is necessary to check the presence of each HRTF in the perceived signal individually. Simply applying a brute force search will localize the sound source but may be inefficient. To improve upon this, a search region may be defined, determines which HRTF-subset is to be searched and in what order to evaluate the HRTFs.
The evaluation of the respective approaches is made by conducting comprehensive numerical experiments.

2. HRTF Localization Algorithms

In this section, we briefly describe four HRTF-based sound localization algorithms, namely the Matched Filtering Approach, the Source Cancellation Approach, the Reference Signal Approach and the Cross Convolution Approach. These algorithms return the position of the sound source using the recorded ear signals and a stored HRTF database. As illustrated in Figure 3, the unknown signal S emitted from a source is filtered by the corresponding left and right HRTFs, denoted by H_{L,i_0} and H_{R,i_0}, before being captured by a humanoid robot, i.e., the left and right microphone recordings X_L and X_R are constructed as

$$\begin{aligned} X_L &= H_{L,i_0} \cdot S, \\ X_R &= H_{R,i_0} \cdot S. \end{aligned} \tag{1}$$

The key idea of the HRTF-based localization algorithms is to identify a pair of HRTFs corresponding to the emitting position of the source, such that correlation between left and right microphone observations is maximized.

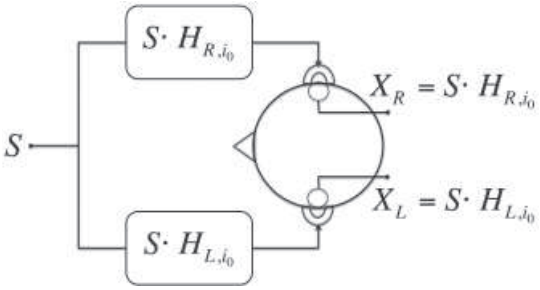


Fig. 3. Single-Source HRTF Model

2.1 Matched Filtering Approach

The Matched Filtering Approach seeks to reverse the H_{R,i_0} and H_{L,i_0}-filtering of the unknown sound source S as illustrated in Figure 3. A schematic view of the Matched Filtering Approach is given in Figure 4.

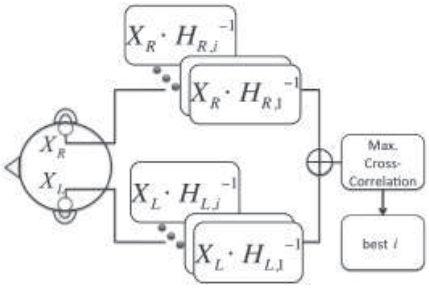


Fig. 4. Schematic view of the Matched Filtering Approach

The localization algorithm is based on the fact that filtering X_L and X_R with the inverse of the correct emitting HRTFs yields identical signals $\tilde{S}_{R,i}$ and $\tilde{S}_{L,i}$, i.e. the original mono sound signal S in an ideal case:

$$\begin{aligned}\tilde{S}_{L,i} &= H_{L,i}^{-1} \cdot X_L \\ &= H_{R,i}^{-1} \cdot X_R \\ &= \tilde{S}_{R,i} \iff i = i_0.\end{aligned} \tag{2}$$

In real case, the sound source can be localized by maximizing the cross-correlation between $\tilde{S}_{R,i}$ and $\tilde{S}_{L,i}$,

$$arg\max_i \left\{ \left(\tilde{S}_{R,i}\right) \oplus \left(\tilde{S}_{L,i}\right) \right\}, \tag{3}$$

where i is the index of HRTFs in the database and $\oplus$ denotes a cross-correlation operation. Unfortunately the inversion of HRTFs can be problematic due to instability. This is mainly due to the linear-phase component of HRTFs responsible for encoding ITDs. Hence a stable approximation must be made of the instable version, retaining all direction-dependent information. One method is to use outer-inner factorization, converting an unstable inverse into an anti-causal and bounded inverse (Keyrouz et al., 2006).

2.2 Source Cancellation Algorithm

The Source Cancellation Algorithm is an extension of the Matched Filtering Approach. Equivalently to cross-correlating all pairs $X_L \cdot H_{L,i}^{-1}$ and $X_R \cdot H_{R,i}^{-1}$, the problem can be restated as a cross-correlation between all pairs $\frac{X_L}{X_R}$ and $\frac{H_{L,i}}{H_{R,i}}$. The improvement is that the ratio of HRTFs does not need to be inverted and can be precomputed and stored in memory (Keyrouz & Diepold, 2006; Usman et al., 2008).

$$arg\max_i \left\{ \left(\frac{X_L}{X_R}\right) \oplus \left(\frac{H_{L,i}}{H_{R,i}}\right) \right\} \tag{4}$$

2.3 Reference Signal Approach

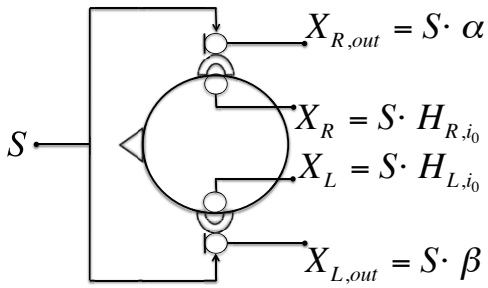


Fig. 5. Schematic view of the Reference Signal Approach setup

This approach uses four microphones as shown in Figure 5: two for the HRTF-filtered signals (X_L and X_R) and two outside the ear canal for original sound signals ($X_{L,out}$ and $X_{R,out}$). The previous algorithms used two microphones, each receiving the HRTF-filtered mono sound signals. The four signals now captured are:

$$X_L = S \cdot H_L \tag{5}$$

$$X_R = S \cdot H_R \tag{6}$$

$$X_{L,out} = S \cdot \alpha \tag{7}$$

$$X_{R,out} = S \cdot \beta \tag{8}$$

α and β represent time delay and attenuation elements that occur due to the heads shadowing. From these signals three ratios are calculated. $\frac{X_L}{X_{L,out}}$ and $\frac{X_R}{X_{R,out}}$ are the left and right HRTFs respectively and $\frac{X_L}{X_R}$ is the ratio between the left and right HRTFs. The three ratios are then cross correlated with the respective reference HRTFs (HRTF ratios in case of $\frac{X_L}{X_R}$). The cross-correlation coefficients are summed, and the HRTF pair yielding the maximum sum

$$arg\max_i \left\{ \left(\frac{X_L}{X_{L,out}} \oplus H_{L,i} \right) + \left(\frac{X_L}{X_R} \oplus \frac{H_{L,i}}{H_{R,i}} \right) + \left(\frac{X_R}{X_{R,out}} \oplus H_{R,i} \right) \right\} \tag{9}$$

defines the incident direction (Keyrouz & Abou Saleh, 2007). The advantage of this system is that HRTFs can be directly calculated yet retain the original undistorted sound signals $X_{L,out}$ and $X_{R,out}$. Thus the direction-dependent filter can alter the incident spectra without regard to the contained information, possibly allowing for better localization. However, the need for four microphones diverges from the concept of binaural localization, exhibiting more hardware and consequently higher costs.

2.4 Convolution Based Approach

To avoid the instability problem, this approach is to exploit the associative property of convolution operator (Usman et al., 2008). Figure 6 illustrates the single-source cross-convolution localization approach. Namely, left and right observations $\tilde{S}_{R,i}$ and $\tilde{S}_{L,i}$ are filtered with a pair of contralateral HRTFs. The filtered observations turn to be identical at the correct source position for the ideal case:

$$\begin{aligned} \tilde{S}_{L,i} =& H_{R,i} \cdot X_L \\ =& H_{R,i} \cdot H_{L,i_0} \cdot S \\ =& H_{L,i} \cdot H_{R,i_0} \cdot S \\ =& H_{L,i} \cdot X_R \\ =& \tilde{S}_{R,i} \Longleftrightarrow i = i_0. \end{aligned} \tag{10}$$

Similar to the matched filtering approach, the source can be localized in real case by solving the following problem:

$$arg\max_i \left\{ \left(\tilde{S}_{R,i} \right) \oplus \left(\tilde{S}_{L,i} \right) \right\}. \tag{11}$$

2.5 Numerical Comparison

In this section, the previously described localization algorithms are compared by numerical simulations. We use the CIPIC database (Algazi et al., 2001) for our HRTF-based localization experiments. The spatial resolution of the database is 1250 sampling points ($N_e = 50$ in elevation and $N_a = 25$ in azimut) and the length is 200 samples.

In each experiment, generic and real-world test signals are virtually synthesized to the 1250 directions of the database, using the corresponding HRTF. The algorithms are then used to localized the signals and a localization success rate is computed. Noise robustness of the algorithm is investigated by different signal-to-noise ratios (SNRs) of the test signals. It should be noted that testing of the localization performance is rigorous, meaning, that we

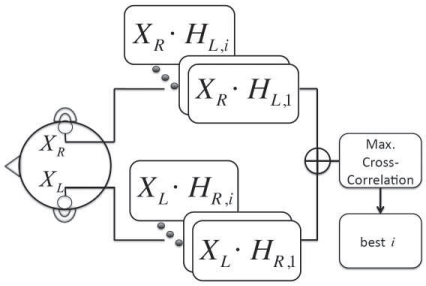


Fig. 6. Schematic view of the cross-convolution approach

do not apply any preprocessing to avoid e.g. instability of HRTF inversion. The localization algorithms are implemented as described above.
Figure 7 shows the achieved localization results of the simulation. The Convolution Based Algorithm, where no HRTF-inversion has to be computed, outperforms the other algorithms in terms of noise robustness and localization success. Furthermore, the best localization results are achieved with white Gaussian noise sources as these ideally cover the entire frequency spectrum. A more realistic sound source is music. It can be seen in Figure 7(d), that the localization performance is slightly degraded compared to the white Gaussian sound sources. The reason for this is that music generally does not inhabit the entire frequency spectrum equally. Speech signals are even more sparse than music resulting in localization success rates worse than for music signals.
Due to the results of the numerical comparison of the different HRTF-based localization algorithms, only the Convolution Based Approach will be utilized to evaluate HRTF data reduction techniques in Section 3 and predictors in Section 4.

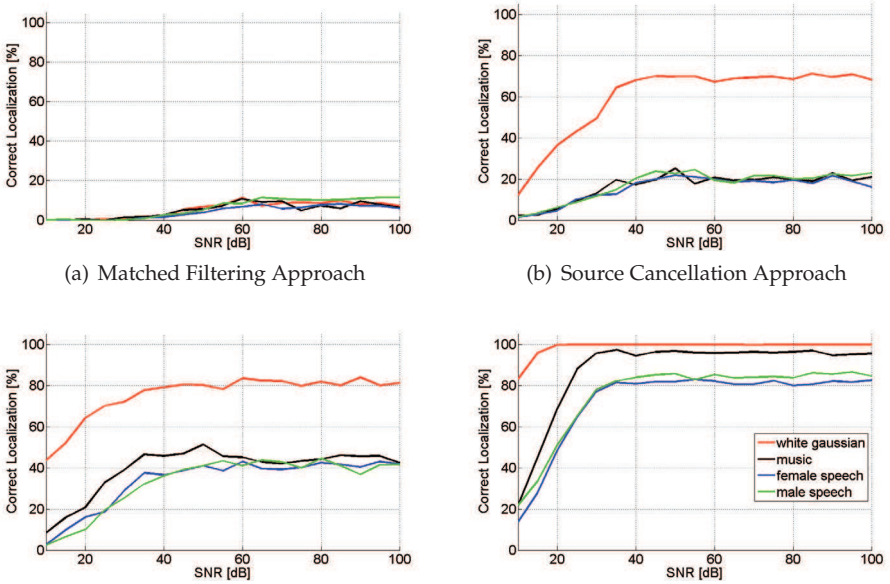
3. HRTF Data reduction techniques

In general, as illustrated in Figure 8, each HRTF dataset can be represented as a three-way array $\mathcal{H} \in \mathbb{R}^{N_a \times N_e \times N_t}$.
The dimensions N_a and N_e are the spatial resolutions of azimuth and elevation, respectively, and N_t the time sample size. By a Matlab-like notation, in this section we denote $\mathcal{H}(i,j,k) \in \mathbb{R}$ the (i,j,k)-th entry of $\mathcal{H}$, $\mathcal{H}(l,m,:) \in \mathbb{R}^{N_t}$ the vector with a fixed pair of (l,m) of $\mathcal{H}$ and $\mathcal{H}(l,:,:) \in \mathbb{R}^{N_e \times N_t}$ the l-th slide (matrix) of $\mathcal{H}$ along the azimuth direction.

3.1 Principal Component Analysis (PCA)

Principal Component Analysis expresses high-dimensional data in a lower dimension, thus removing information yet retaining the critical features. PCA uses statistics to extract the adequately named principal components from a signal (in essence being the information that defines the target signal).
The dimensionality reduction of HRIRs by using PCA is described as follows. First of all, we construct the matrix

$$H := [\mathrm{vec}(\mathcal{H}(:,:,1))^\top, \ldots, \mathrm{vec}(\mathcal{H}(:,:,N_t))^\top] \in \mathbb{R}^{N_t \times (N_a \cdot N_e)}, \tag{12}$$



(a) Matched Filtering Approach (b) Source Cancellation Approach

(c) Reference Signal Approach (d) Convolution Based Approach

Fig. 7. Comparison of HRTF-based sound localization algorithms.

where the operator vec$(\cdot)$ puts a matrix into a vector form. Let $H = [h_1, \ldots, h_{N_t}]$. The mean value of columns of H is then computed by

$$\mu = \frac{1}{N_t} \sum_{i=1}^{N_t} h_i. \tag{13}$$

After centering each row of H, i.e. computing $\widehat{H} = [\widehat{h}_1, \ldots, \widehat{h}_{N_t}] \in \mathbb{R}^{N_t \times (N_a \cdot N_e)}$ where $\widehat{h}_i = h_i - \mu$ for $i = 1, \ldots, N_t$, the covariance matrix of $\widehat{H}$ is computed as follows

$$C := \frac{1}{N_t} \widehat{H} \widehat{H}^\top. \tag{14}$$

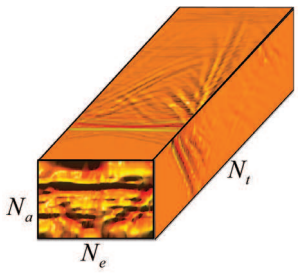


Fig. 8. HRIR dataset represented as a three-way array

Now we compute the eigenvalue decomposition of C and select q eigenvectors $\{x_1, \ldots, x_q\}$ corresponding to the q largest eigenvalues. Then by denoting $X = [x_1, \ldots, x_q] \in \mathbb{R}^{N_t \times q}$, the HRIR dataset can be reduced by the following

$$\widetilde{H} = X^\top \widehat{H} \in \mathbb{R}^{q \times (N_a \cdot N_e)}. \tag{15}$$

Note, that the storage space for the reduced HRIR dataset depends on the value of q. Finally to reconstruct the HRIR dataset one need to compute

$$H_r = X\widetilde{H} + \mu \in \mathbb{R}^{N_t \times (N_a \cdot N_e)}. \tag{16}$$

We refer to (Jolliffe, 2002) for further discussions on PCA.

3.2 Tensor-SVD of three-way array

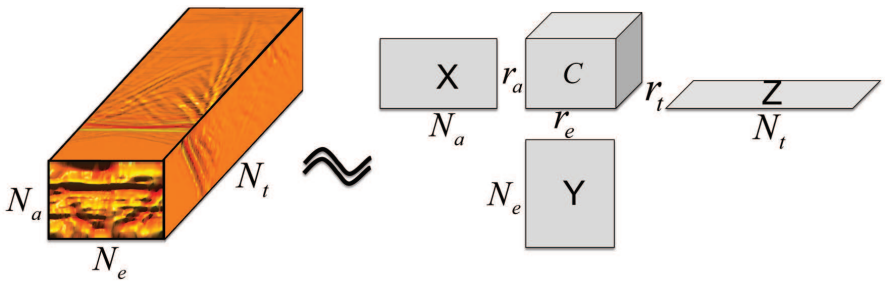


Fig. 9. Schematic view of the Tensor-SVD.

Unlike the PCA algorithm vectorizing the HRIR dataset, Tensor-SVD keeps the structure of the original 3D dataset intact. As shown in Figure 9, given a HRIR dataset $\mathcal{H} \in \mathbb{R}^{N_a \times N_e \times N_t}$, Tensor-SVD computes its best multilinear $rank-(r_a, r_e, r_t)$ approximation $\widehat{\mathcal{H}} \in \mathbb{R}^{N_a \times N_e \times N_t}$, where $N_a > r_a$, $N_e > r_e$ and $N_t > r_t$, by solving the following minimization problem

$$\min_{\widehat{\mathcal{H}} \in \mathbb{R}^{N_a \times N_e \times N_t}} \left\| \mathcal{H} - \widehat{\mathcal{H}} \right\|_{\mathrm{F}}, \tag{17}$$

where $\| \cdot \|_{\mathrm{F}}$ denotes the Frobenius norm of tensors. The $rank-(r_a, r_e, r_t)$ tensor $\widehat{\mathcal{H}}$ can be decomposed as a *trilinear* multiplication of a $rank-(r_a, r_e, r_t)$ core tensor $\mathcal{C} \in \mathbb{R}^{r_a \times r_e \times r_t}$ with three full-rank matrices $X \in \mathbb{R}^{N_a \times r_a}$, $Y \in \mathbb{R}^{N_e \times r_e}$ and $Z \in \mathbb{R}^{N_t \times r_t}$, which is defined by

$$\widehat{\mathcal{H}} = (X, Y, Z) \cdot \mathcal{C} \tag{18}$$

where the (i, j, k)-th entry of $\widehat{\mathcal{H}}$ is computed by

$$\widehat{\mathcal{H}}(i, j, k) = \sum_{\alpha=1}^{r_a} \sum_{\beta=1}^{r_e} \sum_{\gamma=1}^{r_t} x_{i\alpha} y_{j\beta} z_{k\gamma} \mathcal{C}(\alpha, \beta, \gamma). \tag{19}$$

Thus without loss of generality, the minimization problem as defined in (17) is equivalent to the following

$$\begin{aligned} &\min_{X,Y,Z,\mathcal{C}} \| \mathcal{H} - (X, Y, Z) \cdot \mathcal{C} \|_{\mathrm{F}}, \\ &\text{s.t.} \quad X^\top X = I_{r_a}, Y^\top Y = I_{r_e} \text{ and } Z^\top Z = I_{r_t}. \end{aligned} \tag{20}$$

We refer to (Savas & Lim, 2008) for Tensor-SVD algorithms and further discussions.

3.3 Generalized Low Rank Approximations of Matrices

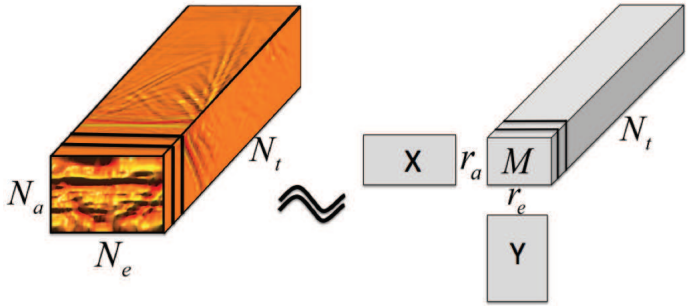


Fig. 10. Schematic view of the Generalized Low Rank Approximations of Matrices

Similar to Tensor-SVD, GLRAM methods, shown in Figure 10 do not require destruction of a 3D tensor. Instead of compressing along all three directions as Tensor-SVD, GLRAM methods work with two pre-selected directions of a 3D data array.
Given a HRIR dataset $\mathcal{H} \in \mathbb{R}^{N_a \times N_e \times N_t}$, we assume to compress $\mathcal{H}$ in the first two directions. Then the task of GLRAM is to approximate slides (matrices) $\mathcal{H}(:,:,i)$, for $i = 1, \ldots, N_t$, of $\mathcal{H}$ along the third direction by a set of low rank matrices $\{XM_iY^\top\} \subset \mathbb{R}^{N_a \times N_e}$, for $i = 1, \ldots, N_t$, where the matrices $X \in \mathbb{R}^{N_a \times r_a}$ and $Y \in \mathbb{R}^{N_e \times r_e}$ are of full rank, and the set of matrices $\{M_i\} \subset \mathbb{R}^{r_a \times r_e}$ with $N_a > r_a$ and $N_e > r_e$. This can be formulated as the following optimization problem

$$\begin{aligned} \min_{X,Y,\{M_i\}_{i=1}^{N_t}} & \sum_{i=1}^{N_t} \left\| (\mathcal{H}(:,:,i) - XM_iY^\top) \right\|_{\mathrm{F}}, \\ \text{s.t.} \quad & X^\top X = I_{r_a} \text{ and } Y^\top Y = I_{r_e}. \end{aligned} \tag{21}$$

Here, by abuse of notations, $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of matrices. Let us construct a 3D array $\mathcal{M} \in \mathbb{R}^{r_a \times r_e \times N_t}$ by assigning $\mathcal{M}(:,:,i) = M_i$ for $i = 1, \ldots, N_t$. The minimization problem as defined in (21) can be reformulated in a Tensor-SVD style, i.e.

$$\begin{aligned} \min_{X,Y,\mathcal{M}} & \left\| \mathcal{H} - (X, Y, I_{N_t}) \cdot \mathcal{M} \right\|_{\mathrm{F}}, \\ \text{s.t.} \quad & X^\top X = I_{r_a} \text{ and } Y^\top Y = I_{r_e}. \end{aligned} \tag{22}$$

We refer to (Ye, 2005) for more details on GLRAM algorithms.
GLRAM methods work on two pre-selected directions out of three. There are then in total three different combinations of directions to implement GLRAM on an HRIR dataset. Performance of GLRAM in different directions might vary significantly. This issue will be investigated and discussed in section 3.5.

3.4 Diffuse Field Equalization (DFE)

A technique that provides good compression performance is diffuse field equalization. The technique reduces the number of samples per HRIR, yet retains the original characteristics.
We define the matrix H containing the HRTFs as

$$H := [\mathrm{vec}(\mathcal{H}(:,:,1)), \ldots, \mathrm{vec}(\mathcal{H}(:,:,N_t))] \in R^{(N_a \cdot N_e) \times N_t}, \tag{23}$$

where the operator $vec(\cdot)$ puts a matrix into a vector form. Let $H = [h_1, \ldots, h_{(N_a \cdot N_e)}]$.
DFE removes the time delay at the beginning of each HRTF and then calculates the average power spectrum from all HRTFs, which then is deconvolved from each HRTF, thus removing direction-independent information. The average power $\widetilde{h}$ is computed by

$$\widetilde{h} = \mathcal{F}^{-1}\{\frac{1}{(N_a \cdot N_e)} \sum_{i=1}^{(N_a \cdot N_e)} |\mathcal{F}\{h_i\}|^2\}, \tag{24}$$

where $\mathcal{F}\{\cdot\}$ denotes the fourier transform. Then, $\widetilde{h}$ is shifted circularily by half the kernel length:

$$\widetilde{h}_1 = [\widetilde{h}(\frac{N_t}{2} + 1 \ldots N_t)\ \ \widetilde{h}(1 \ldots \frac{N_t}{2})]. \tag{25}$$

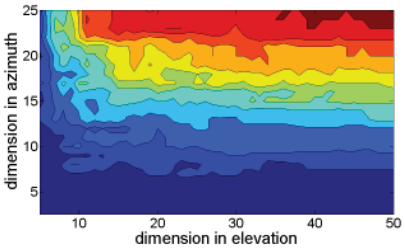
The filter kernel $\widetilde{h}_1$ is inverted and minimum phase reconstruction is applied, yielding $\widetilde{h}_1^{-1}$. The diffused field equalized dataset is retrieved by

$$h_{\mathrm{DFE}} = [(h_1 * \widetilde{h}_1^{-1}), \ldots, (h_{(N_a \cdot N_e)} * \widetilde{h}_1^{-1})]. \tag{26}$$

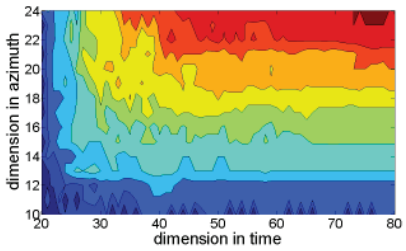
After retrieving the dataset h_{DFE} the time delay samples at the beginning of each HRIR can be removed. To achieve higher compression of the dataset, also samples at the end of each HRTFs, which do not contain crucial direction dependent information, can be removed. For further information on DFE see (Moeller, 1992).

3.5 Numerical Comparison

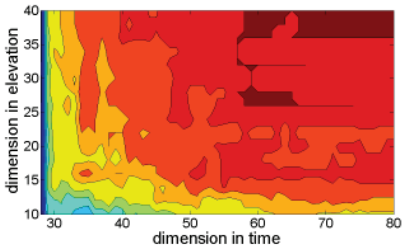
In this section, PCA, GLRAM, Tensor-SVD and Diffused Field Equalization are applied to a HRTF-based sound localization problem, in order to evaluate performance of these methods for data reduction. In each experiment, left and right ear KEMAR HRTF are reduced with one of the introduced reduction methods. A test signal, which is white noise is virtually synthesized using the corresponding original HRTF. The convolution based sound localization algorithm as descirbed in Section 2.4, is fed with the restored databases and used to localize the signals. Finally, the localization success rate is computed.
As already mentioned, GLRAM works on two preselected directions out of three. Therefore, we conduct localization experiments for a subset of directions (35 randomly chosen locations) to detect a combination of well working parameters for GLRAM. After finding a suitable combination of the variables, localization experiments for all 1250 directions are conducted. Firstly, the dataset is reduced for the first two directions, i.e. elevation and azimuth. The contour plot given in Figure 11(a) shows the localization success rate for a fixed pair of values (N_{r_a}, N_{r_e}). Similar results with respect to the pairs (N_{r_a}, N_{r_t}) and (N_{r_e}, N_{r_t}) are ploted in Figure 11(b) and Figure 11(c), respectively. Clearly, applying GLRAM on the pair of (N_{r_e}, N_{r_t}) outperforms the other two combinations.
The application of GLRAM in the directions of elevation and time performs best, therefore, we compare this optimal GLRAM with the standard PCA and Tensor-SVD. As mentioned in section 3.3, GLRAM is a simple form of Tensor-SVD with leaving one direction out. Thus, we investigate the effect of additionally reducing the third direction, whereas the dimensions in elevation and time are fixed to the parameters of the optimal GLRAM. Figure 13 shows that additionally decreasing the dimension in azimuth leads to a huge loss of localization accuracy. After determining the optimal parameters for GLRAM, the simulations are conducted for all 1250 directions of the CIPIC dataset. Figure 12 shows the localization success rate in dependency of the compression rate for GLRAM and PCA. It can be seen that an optimized GLRAM outperforms the standard PCA in terms of compression.



(a) GLRAM on $(azimuth, elevation)$



(b) GLRAM on $(azimuth, time)$



(c) GLRAM on $(elevation, time)$

Fig. 11. Contour plots of localization success rate of using GLRAM in different settings.

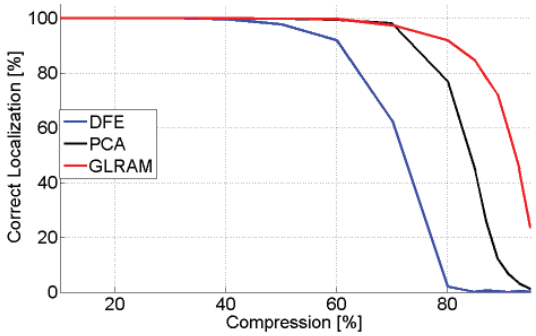


Fig. 12. Comparison between DFE, PCA and GLRAM

4. Predictors for HRTF sound localization

To reduce the computational costs of HRTF-based sound localization, especially for moving sound sources, it is advantageous to determine a region of interest (ROI) as illustrated in Figure 15. A ROI constricts the 3D search space around the robotic platform leading to a reduced set of eligible HRTFs.

Various tracking models have been implemented in microphone sound localization. Primarily they predict the path of a sound source as it is traveling and thus acquiring faster and more accurate non-ambiguous localization results (Belcher et al., 2003; Ward et al., 2003). Most of these filters are updated periodically in scans. In this section, three predictors, namely Time

Delay of Arrival, Kalman filter and Particle filter, are briefly introduced to determine a ROI to reduce the set of eligible HRTFs to be processed to localize moving sound sources.

4.1 Time Delay of Arrival

The time delay between the two signals $x_i[n]$ and $x_j[n]$ is found when the cross-correlation value $R_{ij}(\tau)$ is maximal. Given that τ has been determined, the time delay is calculated by

$$\Delta T = \frac{\tau}{f_s}, \tag{27}$$

where f_s is the sampling rate. Knowing the geometry (distance between the robot's ears) of the microphones and the delays between microphone pairs, a number of locations for the sound source can be disregarded (Brandstein & Ward, 2001; Kwok et al., 2005; Potamitis et al., 2004; Valin et al., 2003). Then, an HRTF-based localization algorithm only evaluates the remaining possible locations of the source.

4.2 Kalman Filter

The Kalman filter is a frequently used predictor (usage for microphone array localization described in (Belcher et al., 2003)). The discrete version exhibits two main states: time update (prediction) and measurement update (correction). The Kalman filter predicts the state of x_k at time k given the linear stochastic difference equation

$$\mathbf{x}_k = A\mathbf{x}_{k-1} + Bu_{k-1} + w_{k-1} \tag{28}$$

and measurement

$$z_k = H\mathbf{x}_k + v_k. \tag{29}$$

Matrices A, B and H provide relation from discrete time $k-1$ to k for their respective variables x (the state) and u (optional control input). w and v add noise to the model. A set of time and measurement update equations are used to predict the next state (Kalman, 1960). The state vector is defined by current location coordinates x and y and the velocity components v_x and v_y (Potamitis et al., 2004; Usman et al., 2008). Note that here the predictor is applied to two dimensional space.

$$\mathbf{x} = [x, v_x, y, v_y]^T \tag{30}$$

An unreliable location estimate during initialization of the the Kalman filter may be a source of error. To improve upon this, particle filters have been implemented in (Chen & Rui, 2004).

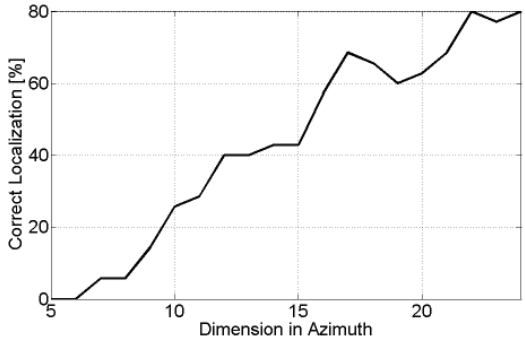


Fig. 13. Localization success rate by Tensor-SVD

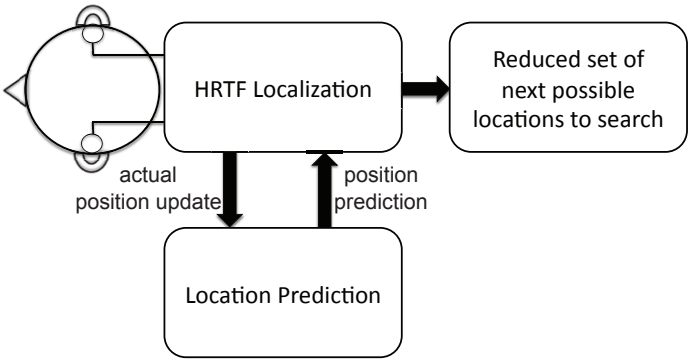


Fig. 14. Schematic view of the application of predictors in HRTF-based localization.

4.3 Particle Filter

The particle filter bases itself on the idea of randomly generating samples from a distribution and assigning weights to each to define their reliability. The particles and their associated weights define an averaged center which is the predicted value for the next step. Each weight w_k^i is associated to a particle x^i in iteration k. A set of N particles is initially drawn from a distribution $q(x_i|x_{k-1}^i, z_k)$ with z_k being the current observed value. For each particle the weight is calculated by

$$w_k^i = w_{k-1}^i \frac{p(z_k|x_k^i)p(x_k^i|x_{k-1}^i)}{q(x_k^i|x_{0:k-1}^i, z_{1:k})}. \tag{31}$$

Once all weights are calculated, their sum is normalized. To determine the predicted value, the weighted average of the particles is taken:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} w_k^i \cdot x_i \tag{32}$$

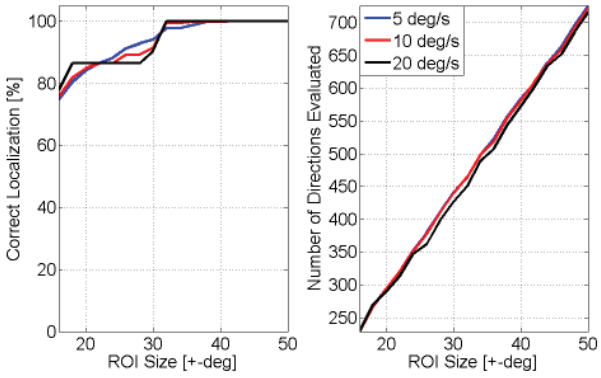
Over time it may occur that very few particles possess most of the weight. This case requires resampling to protect from particle degeneration. The variance of the weights is used as a measure to check for this case and if required, the set of weights is exchanged with a better approximation (Gordon et al., 1993).

Many particle filter variations exist, such as the Monte Carlo approximations and Sampling Importance Resampling. However a particle filter may find only a local optimum and thus never reaching the global optimum. Evolutionary estimation is proposed in (Kwok et al., 2005) to overcome such problems. Initially a set of potential speaker locations are estimated and then a heuristic search is performed. The speaker locations are called chromosomes and can only move within a defined region. After the initialization, the Time Delay of Arrival (TDOA) is evaluated for each potential location as well as each microphone. The difference v_i between expected and actual TDOAs is used to define a fitness function for each chromosome i together with error variance σ_τ^2:

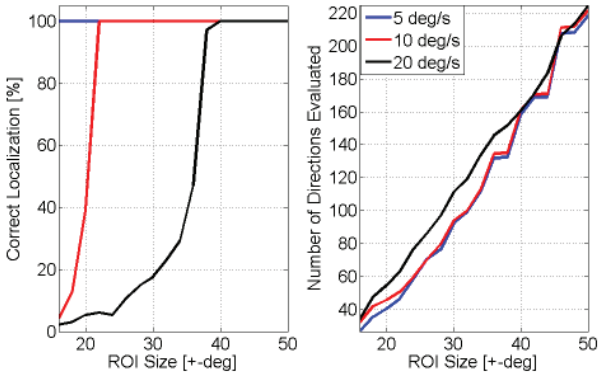
$$\omega_i = e^{-0.5\frac{v_i^2}{\sigma_\tau^2}} \tag{33}$$

ω_i is then scaled such that $\sum_{i=1}^{n} \omega_i = 1 \rightarrow \widetilde{\omega}_i$ The new estimate of source location is given by

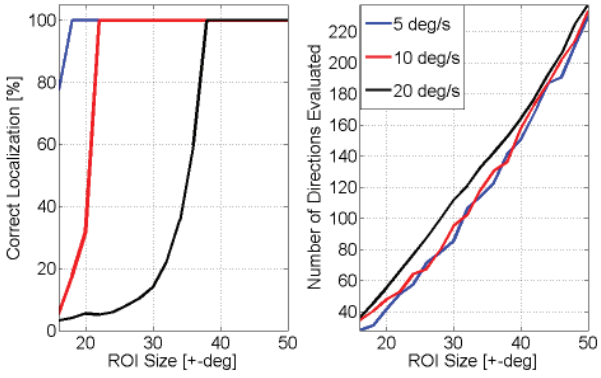
$$s_x = \sum_{i=1}^{n} \widetilde{\omega}_i s_{xi}. \tag{34}$$



(a) Time Delay of Arrival



(b) Particle Filter



(c) Kalman Filter

Fig. 15. Comparison of predictors for HRTF Sound Localization.

Chromosomes are then selected according to a linearly spaced pointer spanning the fitness magnitude scale, with higher fitness chromosomes being selected more often. The latter chromosomes receive less mutation as compared to weaker chromosomes depending on r_g, the zero mean Gaussian random number variance, and d_m, the distance for mutation (Kwok et al., 2005).

$$s_{xi+1} = s_{xi} + r_g d_m \tag{35}$$

4.4 Numerical comparison

This section gives a performance overview of the applied predictors in a HRTF-based sound localization scenario. We simulate moving sound by virtually synthesizing a sound source, which is white noise, using different pairs of HRTFs. This way, a random path of 500 different source positions is generated, simulating a moving sound source. Then, Time Delay of Arrival, the Kalman filter and the Particle filter seek to reduce the search region for the HRTF-based sound localization to a region of interest. The Convolution Based Algorithm is utilized to localize the moving sound source. The experiments were conducted three times with different speed of the sound source.

Figure 15 summarizes the results of applying predictors to HRTF-based sound localization. The left plots show the localization success rates in dependency of the size of the region of interest. In the right plots the number of directions that have to be evaluated within the localization algorithms are shown. The bigger the region of interest, the more HRTF-pairs have to be utilized to maximize the cross correlation (11) resulting in a higher processing time. On the other hand, the smaller the region of interest, the higher the danger of excluding the HRTF pair that is maximizing the cross correlation (11), leading to false localization results. Our simulation results show that the number of HRTFs to be evaluated for the Convolution Based Algorithm can be significantly reduced to speed up HRTF-based localization for moving sources. Time Delay of Arrival is reducing the search region to 500 directions while reaching hundred percent correct localization of the path, meaning all 500 source positions are detected correctly for the different speeds of the sources. Particle- and Kalman filter are able to reduce the search region to 130 directions in case of sound sources with a speed of $20 deg/s$. For slower sources, only 60 directions need to be taken into account.

Acknowledgements

This work was fully supported by the German Research Foundation (DFG) within the collaborative research center SFB-453 "High Fidelity Telepresence and Teleaction".

5. References

Algazi, V. R., Duda, R. O., Thompson, D. M. & Avendano, C. (2001). The CIPIC HRTF database, *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, pp. 21–24.

Belcher, D., Grimm, M. & Kroschel, K. (2003). Speaker tracking with a microphone array using a kalman filter, *Advances in Radio Science* 1: 113–117.

Blauert, J. (1997). An introduction to binaural technology, *Binaural and Spatial Hearing*, R. Gilkey, T. Anderson, Eds., Lawrence Erlbaum, Hilldale, NJ, USA, pp. 593–609.

Brandstein, M. & Ward, D. (2001). *Microphone arrays - signal processing techniques and applications*, Springer.

Chen, Y. & Rui, Y. (2004). Real-time speaker tracking using particle filter sensor fusion, *Proceedings of the IEEE* 92(3): 485–494.

Gordon, N., Salmond, D. & Smith, A. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *Radar and Signal Processing, IEE Proceedings F* 140(2): 107 –113.

Jolliffe, I. T. (2002). *Principal Component Analysis*, second edn, Springer.

Kalman, R. (1960). A new approach to linear filtering and prediction problems, *Transactions of the ASME - Journal of Basic Engineering* 82(Series D): 35–45.

Keyrouz, F. & Abou Saleh, A. (2007). Intelligent sound source localization based on head-related transfer functions, *IEEE International Conference on Intelligent Computer Communication and Processing*, pp. 97–104.

Keyrouz, F. & Diepold, K. (2006). An enhanced binaural 3D sound localization algorithm, *2006 IEEE International Symposium on Signal Processing and Information Technology*, pp. 662–665.

Keyrouz, F., Diepold, K. & Dewilde, P. (2006). Robust 3D Robotic Sound Localization Using State-Space HRTF Inversion, *IEEE International Conference on Robotics and Biomimetics, 2006. ROBIO'06*, pp. 245–250.

Kwok, N., Buchholz, J., Fang, G. & Gal, J. (2005). Sound source localization: microphone array design and evolutionary estimation, *IEEE International Conference on Industrial Technology*, pp. 281–286.

Moeller, H. (1992). Fundamentals of binaural technology, *Applied Acoustics* 36(3-4): 171–218.

Potamitis, I., Chen, H. & Tremoulis, G. (2004). Tracking of multiple moving speakers with multiple microphone arrays, *IEEE Transactions on Speech and Audio Processing* 12(5): 520–529.

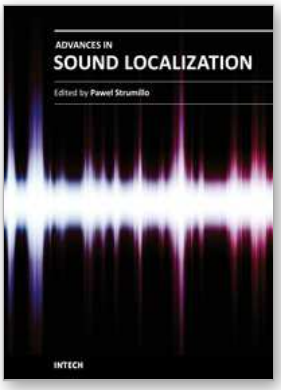
Savas, B. & Lim, L. (2008). Best multilinear rank approximation of tensors with quasi-Newton methods on Grassmannians, *Technical Report LITH-MAT-R-2008-01-SE*, Department of Mathematics, Linkpings University.

Usman, M., Keyrouz, F. & Diepold, K. (2008). Real time humanoid sound source localization and tracking in a highly reverberant environment, *Proceedings of 9th International Conference on Signal Processing*, Beijing, China, pp. 2661–2664.

Valin, J., Michaud, F., Rouat, J. & Letourneau, D. (2003). Robust sound source localization using a microphone array on a mobile robot, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 2.

Ward, D. B., Lehmann, E. A. & Williamson, R. C. (2003). Particle filtering algorithms for tracking an acoustic source in a reverberant environment, *IEEE Transactions on Speech and Audio Processing* 11(6): 826–836.

Ye, J. (2005). Generalized low rank approximations of matrices, *Machine Learning* 61(1-3): 167–191.



Advances in Sound Localization

Edited by Dr. Pawel Strumillo

ISBN 978-953-307-224-1
Hard cover, 590 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

Sound source localization is an important research field that has attracted researchers' efforts from many technical and biomedical sciences. Sound source localization (SSL) is defined as the determination of the direction from a receiver, but also includes the distance from it. Because of the wave nature of sound propagation, phenomena such as refraction, diffraction, diffusion, reflection, reverberation and interference occur. The wide spectrum of sound frequencies that range from infrasounds through acoustic sounds to ultrasounds, also introduces difficulties, as different spectrum components have different penetration properties through the medium. Consequently, SSL is a complex computation problem and development of robust sound localization techniques calls for different approaches, including multisensor schemes, null-steering beamforming and time-difference arrival techniques. The book offers a rich source of valuable material on advances on SSL techniques and their applications that should appeal to researches representing diverse engineering and scientific disciplines.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821