On the Impact of Localization Errors on HRTF-based Robust Least-Squares Beamforming

Hendrik Barfuss and Walter Kellermann*
Multimedia Communications and Signal Processing, Friedrich-Alexander University Erlangen-Nürnberg
{barfuss, wk}@f-m.de

Introduction
In a typical human-machine dialogue scenario, the target source and additional interfering sources, located at different positions from the target source, may be active at the same time. Clearly, these interfering sources have to be suppressed in order to establish a successful human-machine interaction. A common strategy is to apply spatial filtering techniques which are usually based on the free-field assumption of acoustic wave propagation. However, for scenarios where the microphones are mounted on a scatterer, the free-field assumption is not optimum, since the influence of the scatterer on the sound field is neglected. One example of such a scenario is a microphone array mounted on a robot head used for robot audition, which is also the focus of this article.

In order to design a beamformer which accounts for the influence of the scatterer, i.e., the robot head, on the sound field, the free-field steering vectors have to be replaced by Head-Related Transfer Functions (HRTFs) see, e.g., [1].

In [2], we proposed an HRTF-based Robust Least-Squares Frequency-Invariant (RLSFI) beamformer design and verified experimentally that employing HRTFs instead of free-field steering vectors leads to a significantly improved beamforming performance and correspondingly better Automatic Speech Recognition (ASR), in a robot audition scenario. Since the proposed beamformer design depends on a set of HRTFs, the question arises how the beamformer performs if these HRTFs do not correspond to the true position of the target source, e.g., due to localization errors. Therefore, in this contribution, we investigate the impact of localization errors on the performance of the HRTF-based RLSFI beamformer.

The remainder of this article is organized as follows: In the next section, the HRTF-based beamformer design from [2] is briefly reviewed. After this, the results of our investigation of the HRTF robustness are presented, followed by a conclusion and an outlook to future work in the last section.

HRTF-based robust beamforming

Fig. 1 illustrates the block diagram of a Filter-and-Sum Beamformer (FSB), consisting of N microphones at positions \( p_n \), where \( p_n \) represents the position of the \( n \)-th microphone in Cartesian coordinates. In this article, vectors and matrices are denoted by lower- and upper-case boldface letters, respectively. The output signal \( y[k] \) at time instant \( k \) is obtained by convolving the microphone signals \( x_n[k] \) with Finite Impulse Response (FIR) filters \( w_n = [w_{n,0}, \ldots, w_{n,L-1}]^T \) of length \( L \) and a subsequent summation over all \( N \) channels. The beamformer response of an FSB is given as [3] [4]:

\[
B(\omega, \phi, \theta) = \sum_{n=0}^{N-1} W_n(\omega) g_n(\omega, \phi, \theta),
\]

where \( W_n(\omega) = \sum_{l=0}^{L-1} w_{n,l} e^{-j\omega l} \) is the Discrete-Time Fourier Transform (DTFT) representation of \( w_n \). Moreover, \( g_n(\omega, \phi, \theta) \) is the response of the \( n \)-th microphone to a plane wave with frequency \( \omega \) traveling in the direction \((\phi, \theta)\), where \( \phi \) and \( \theta \) denote azimuth and elevation angle, respectively, and are defined as in [3].

In [4], the design of an RLSFI FSB was proposed, where a desired beamformer response \( B(\omega, \phi, \theta) \) is approximated in the Least-Squares (LS) sense at each frequency \( \omega \) subject to a distortionless response constraint in the desired look direction and a constraint on the White Noise Gain (WNG). The LS approximation is performed for a discrete set of \( P \) frequencies \( \omega_p \) and \( M \) look directions \((\phi_m, \theta_m)\), and can be formulated in matrix notation as [4]:

\[
\arg\min_{w_l(\omega_p)} \| G(\omega_p) w_l(\omega_p) - b \|^2_2 \quad (2)
\]

subject to constraints on the WNG and the response in desired look direction, respectively:

\[
\frac{|w_l(\omega_p) d(\omega_p)|^2}{w_l^H(\omega_p) w_l(\omega_p)} \geq \gamma > 0, \quad w_l^T(\omega_p) d(\omega_p) = 1, \quad (3)
\]

*The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 600465.

1Note that in the context of this work, HRTFs only model the direct propagation path between a source and a microphone mounted on a robot head, but no reverberation components.
where \( \mathbf{w}_t(\omega_p) = [W_0(\omega_p), \ldots, W_{N-1}(\omega_p)]^T \), 
\( [\mathbf{G}(\omega_p)]_{mn} = g_n(\omega_p, \phi_m, \theta_m) \), vector \( \mathbf{b} = 
[\mathbf{B}(\phi_0, \theta_0), \ldots, \mathbf{B}(\phi_{M-1}, \theta_{M-1})]^T \) contains the
desired responses for all \( M \) discrete look directions, and 
\( \mathbf{d}(\omega_p) = [g_0(\omega_p, \phi_1, \theta_1), \ldots, g_{N-1}(\omega_p, \phi_1, \theta_1)]^T \) is
the steering vector corresponding to the desired look
direction \((\phi_1, \theta_1)\). Operators \( \| \cdot \|^2 \), \( (\cdot)^T \), and \((\cdot)\) denote
the Euclidean norm, and the transpose and conjugate
transpose of vectors or matrices, respectively. Note that
the same desired response is chosen for all frequencies,
as can be seen from the frequency-independent entries of \( \mathbf{b} \).

Equations (2) and (3) can be interpreted as follows: The
first part of (3) represents the WNG constraint, with the lower bound \( \gamma \) on
the WNG, which has to be defined by the user. The second part of (3)
describes the distortionless response constraint which
ensures that the target signal, coming from the desired
look direction, passes the beamformer undistorted.
The time-domain FIR filters \( \mathbf{w}_n \) are obtained by solving (2), (3) for each frequency \( \omega_p \), separately, followed by an FIR
approximation of the optimum filter coefficients.

Assuming the microphones are located in the free field,
the sensor response is given as
\[
g_n(\omega_p, \phi_m, \theta_m) = e^{-j \mathbf{k}(\omega_p, \phi_m, \theta_m) \cdot \mathbf{p}_m},
\]
where \( \mathbf{k}(\omega_p, \phi_m, \theta_m) \) denotes the wave vector which depends
on the current frequency and look direction, and the speed of sound \( \mathbf{c} \). Thus, matrix \( \mathbf{G}(\omega_p) \) in (2)
contains the well-known free-field-based steering vectors
with respect to the \( M \) look directions and the \( N \) microphones, and vector \( \mathbf{d}(\omega_p) \) in (3) is the free-field-based
steering vector corresponding to the desired look direction.

The HRTF-based RLSFI beamformer design, as proposed in (2), is obtained by including measured or simu-
lated HRTFs in (2) and (3) instead of free-field-based
steering vectors. In this case, the sensor response is given as
\[
g_n(\omega_p, \phi_m, \theta_m) = h_{mn}(\omega_p),
\]
where \( h_{mn}(\omega_p) \) is the HRTF modeling the propagation
between the \( m \)-th source position and \( n \)-th sensor at
frequency \( \omega_p \). Consequently, \( \mathbf{G}(\omega_p) \) now consists of all
HRTFs between the \( M \) look directions and the \( N \) micro-
phones, and \( \mathbf{d}(\omega_p) \) contains the HRTFs corresponding to
the desired look direction. Note that in contrast to the
free-field-based design (1), the HRTFs-based design
implicitly depends on the robot-source distance for which
the HRTFs have been measured (see, e.g., [5]).

In Fig. 2 an example of the HRTF-based RLSFI beam-
former according to (2), (3), and (5) is illustrated for a
frequency range of \( 300 \text{ Hz} \leq f \leq 5000 \text{ Hz} \). The
design was carried out for the 5-microphone robot head
array illustrated in Fig 3(b). Beampatterns for two
different WNG constraint values \( \gamma_{dB} = 10 \log_{10}(\gamma) \in
\{-10, -20\} \text{ dB} \) are shown to demonstrate the impact of
the WNG constraint on the beamformer. It is important
to note that the beampatterns were computed by evalu-
ating (1) with (5). Thus, they effectively show the transfer
function between source position and beamformer output
with HRTFs modeling the acoustic system. We used a
filter length of \( L = 1024 \) for the FIR approximation,
and the HRTFs which were incorporated in the beam-
former design were measured for a robot-source distance
of 1.1m. The main beam was steered towards broadside.
Figs. 2(a) and 2(b) illustrate the resulting beampatterns
for \( \gamma_{dB} = -10 \text{ dB} \) and \( \gamma_{dB} = -20 \text{ dB} \). Subfigure (c)
shows the resulting WNG.

**Experimental results**

In the following, we analyze the relative robustness of
the HRTF-based beamformer design by comparing the
impact of localization errors on the performance of the
HRTF- and free-field-based RLSFI beamformer. More
specifically, we investigate the impact of localization
errors with respect to Direction-Of-Arrival (DOA) and
robot-source distance. At first, the experimental setup
and performance measures are introduced, followed by a
presentation of the experimental results.

**Setup and parameters**

We use Word Error Rates (WERs) of an automatic
speech recognizer to evaluate the overall quality of the
enhanced signals at the beamformer outputs, since a high speech recognition accuracy is the main goal in robot audition. As ASR engine, we employed PocketSphinx [6] with a Hidden Markov Model (HMM)-Gaussian Mixture Model (GMM)-based acoustic model trained on clean speech from the GRID corpus [7], using MFCC+$\Delta+\Delta\Delta$ features and cepstral mean normalization. For the computation of the WER scores, only the letter and the number in the utterance were evaluated, as in the CHiME challenge [8]. Our test signal contained 200 utterances. In addition, the frequency-weighted segmental Signal-to-Noise Ratio (fwSegSNR) as defined in [9] was evaluated, calculating the fwSegSNR at the input and output of the beamformer. As ASR engine, we employed PocketSphinx [6] for speech recognition accuracy is the main goal in robot audition.

(a) Source positions. (b) Microphone positions.

Figure 3: Illustration of the source positions of the two-speaker scenario and the employed microphone positions at the robot head.

Figure 4: Illustration of WERs in % and fwSegSNR levels in dB, obtained at the output of the HRTF-based beamformer for Scenario 1) $\phi_d = 90^\circ$, $\phi_{int} = 70^\circ$ and 2) $\phi_d = 90^\circ$, $\phi_{int} = 170^\circ$ for DOA estimation errors of $\pm 5^\circ$ and $\pm 10^\circ$. Measures at input: Scenario 1) WER$_{in}$ = 49.0% and fwSegSNR$_{in}$ = 5.2dB and Scenario 2) WER$_{in}$ = 44.3% and fwSegSNR$_{in}$ = 5.8dB.

(b) fwSegSNR, $\phi_{int} = 70^\circ$.

(c) WER, $\phi_{int} = 170^\circ$.

(d) fwSegSNR, $\phi_{int} = 170^\circ$.

In Fig. 3 the results for the two scenarios are summarized. The subfigures on the left- and right-hand side show the WERs in % and fwSegSNR levels in dB obtained at the HRTF-based beamformer output, respectively. Each horizontal bar represents the results for one specific localization error of $\pm 5^\circ$ or $\pm 10^\circ$. From Figs. 4(a) and 4(b) it can be seen that when the interferer is very close to the target source, localization errors have a strong impact on the beamforming performance. When the beamformer is accidentally steered closer towards the interfering source (localization errors of $-5^\circ$ and $-10^\circ$), the beamforming performance decreases. This is because the beamformer’s main beam is steered towards the interfering source, leading to a lower attenuation of the latter. If the localization error leads to the beamformer being steered away from the interfering source (localization errors of $5^\circ$ and $10^\circ$), an increasing beamforming performance can be observed. This can be explained by the fact that in this particular scenario, a spatial null of the beam pattern is getting closer to the interferer’s direction the larger the localization error is. If the interferer is far away from the target source, as in Scenario 2), localization errors do not have a strong impact on the beam-
In Fig. 5, the average WERs in % and average fwSegSNR levels in dB, obtained at the output of the HRTF- and free-field-based beamformer for $\phi_{d} = 90^\circ$, $\phi_{int} \in \{10^\circ : (20^\circ) : 70^\circ, 110^\circ : (20^\circ) : 170^\circ\}$, and localization errors of $\pm 5^\circ$ and $\pm 10^\circ$. Average input measures: WER$_{in} = 47.1\%$ and fwSegSNR$_{in} = 5.5$ dB.

### Impact of localization errors with respect to source distance

In a second experiment, we evaluated the impact of localization errors with respect to the robot-source distance. In this work, we investigated the impact of localization errors on the performance of a recently proposed HRTF-based RLSFI beamformer. Localization errors with respect to the DOA of the target signal as well as the robot-source distance were evaluated. The results confirmed that both, erroneous DOA and robot-source distance estimates lead to a significant decrease in beamforming performance. Thus, it is of vital importance to use a set of HRTFs for the design of the HRTF-based RLSFI beamformer, which matches the position of the target source. Future work includes evaluation of the effect of localization errors on the behaviour of the HRTF-based beamformer for sources in the near-field, and an extension of the RLSFI beamformer design to two-dimensional beam steering.

### Conclusion

In this work, we investigated the impact of localization errors on the performance of a recently proposed HRTF-based RLSFI beamformer. Localization errors with respect to the DOA of the target signal as well as the robot-source distance were evaluated. The results confirmed that both, erroneous DOA and robot-source distance estimates lead to a significant decrease in beamforming performance. Thus, it is of vital importance to use a set of HRTFs for the design of the HRTF-based RLSFI beamformer, which matches the position of the target source. Future work includes evaluation of the effect of localization errors on the behaviour of the HRTF-based beamformer for sources in the near-field, and an extension of the RLSFI beamformer design to two-dimensional beam steering.

### References

[1] Maazaoui, M., Abed-Merai, K., and Grenier, Y.: Blind source separation for robot audition using fixed HRTF beamforming, EURASIP J. Advances Signal Processing, 2012 (58), Mar. 2012.

[2] Barfuss, H., Huemer, C., Lamani, G., Schwarz, A., and Kellermann, W.: HRTF-based robust least-squares frequency-invariant beamforming, IEEE Workshop Applications Signal Processing Audio Acoustics (WASPAA), Oct. 2015.

[3] Van Trees, H.L.: Detection, estimation, and modulation theory: Optimum array processing, Wiley, 2004.

[4] Mabande, E., Schad, A., and Kellermann, W.: Design of robust sup-predicate beamformers as a convex optimization problem, IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), Apr. 2009.

[5] Blauert, J.: Spatial hearing: The psychophysics of human sound localization, The MIT Press, 1997.

[6] Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M., and Rudnick, A.I.: PocketSphinx: A free, real-time continuous speech recognition system for hand-held devices, IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), May 2006.

[7] Cooke, M., Barker, J., Cunningham, S., and Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition, J. Acoustical Society America, 120 (5), Nov. 2006.

[8] Christensen, H., Barker, J., Ma, N., and Green, P.D.: The CHIME Corpus: A resource and a challenge for computational hearing in multisource environments, INTERSPEECH, Sept. 2010.

[9] Hu, Y. and Loizou, P.C.: Evaluation of Objective Quality Measures for Speech Enhancement, Audio, Speech, Language Proc., IEEE Trans., 16 (1), Jan. 2008.