HRTF-BASED ROBUST LEAST-SQUARES FREQUENCY-INVARIANT POLYNOMIAL BEAMFORMING

Hendrik Barfuss, Marcel Mueglich, and Walter Kellermann

Multimedia Communications and Signal Processing,
Friedrich-Alexander University Erlangen-Nürnberg
Cauerstr. 7, 91058 Erlangen, Germany
{hendrik.barfuss,walter.kellermann}@fau.de, marcel.mueglich@studium.fau.de

ABSTRACT

In this work, we propose a robust Head-Related Transfer Function (HRTF)-based polynomial beamformer design which accounts for the influence of a humanoid robot's head on the sound field. In addition, it allows for a flexible steering of our previously proposed robust HRTF-based beamformer design. We evaluate the HRTF-based polynomial beamformer design and compare it to the original HRTF-based beamformer design by means of signal-independent measures as well as word error rates of an off-the-shelf speech recognition system. Our results confirm the effectiveness of the polynomial beamformer design, which makes it a promising approach to robust beamforming for robot audition.

Index Terms— Spatial filtering, robust superdirective beamforming, polynomial beamforming, white noise gain, signal enhancement, robot audition, head-related transfer functions

1. INTRODUCTION

Spatial filtering techniques are a widely used means to spatially focus on a target source by exploiting spatial information of a wave field which is sampled by several sensors at different positions in space.

When spatial filtering techniques are applied to a robot audition scenario, i.e., when the microphones are mounted on a humanoid robot's head, the influence of the head on the sound field has to be taken into account by the beamformer design in order to obtain a satisfying spatial filtering performance. To this end, Head-Related Transfer Functions (HRTFs)[1] can be incorporated into the beamformer design as steering vectors, see, e.g., [1, 2, 3]. In [4], Mabande et al. proposed a Robust Least-Squares Frequency-Invariant (RLSFI) beamformer design which allows the user to directly control the tradeoff between the beamformer's spatial selectivity and its robustness. Recently, we extended this design to an HRTF-based RLSFI beamformer design by following the approach described above [5]. Despite all advantages of the beamformer designs in [4, 5], a clear disadvantage is that whenever the beamformer is steered to another direction, a new optimization problem has to be solved which makes it unattractive for real-time processing. To overcome this limitation, Mabande et al. proposed a Robust Least-Squares Frequency-Invariant Polynomial (RLSFIP) beamformer design [6] as extension of [4], which allows for a flexible steering of the beamformer.

In this work, we extend the HRTF-based RLSFI beamformer design [5] to the concept of polynomial beamforming in order to allow for a flexible steering of the HRTF-based beamformer in a robot audition scenario.

The remainder of this article is structured as follows: In Section 2, the HRTF-based RLSFIP beamformer design is introduced. Then, an evaluation of the new HRTF-based polynomial beamformer design is presented in Section 3. Finally, conclusions and an outlook to future work are given in Section 4.

2. HRTF-BASED ROBUST POLYNOMIAL BEAMFORMING

2.1. Concept of polynomial beamforming

In Fig. 1, the block diagram of a Polynomial Filter-and-Sum Beamformer (PFSB), as presented in [6, 7, 8], is illustrated. It consists of a beamforming stage containing $P+1$ Filter-and-Sum Units (FSUs), followed by a Polynomial Postfilter (PPF). The output signal $y_p[k]$ of the p-th FSU at time instant k is obtained by convolving the microphone signals $x_n[k]$, $n \in \{0,\dots,N-1\}$ with the FSU's Finite Impulse Response (FIR) filters $\mathbf{w}_{n,p} = [w_{np,0},\dots,w_{np,L-1}]^{\mathrm{T}}$ of length L, followed by a summation over all N channels. Operator $(\cdot)^{\mathrm{T}}$ represents the transpose of vectors or matrices, respectively. In the PPF, the output $y_D[k]$ of the PFSB is obtained by weighting the output of each FSU by a factor D^p and summing them up:

$$y_D[k] = y_0[k] + Dy_1[k] + D^2 y_2[k] + \ldots + D^P y_P[k]. \qquad (1)$$

Hence, the output signal of each FSU can be seen as one coefficient of a polynomial of order P with variable D. The advantage of a PFSB is that the steering of the main beam is accomplished by simply changing the scalar value D, whereas the FIR filters of the FSUs can be designed beforehand and remain fixed during runtime. A more detailed explanation of how the steering is controlled by D is given in Section 2.2.

The beamformer response of the PFSB is given as [6]:

$$B_D(\omega,\phi,\theta) = \sum_{p=0}^{P} D^p \sum_{n=0}^{N-1} W_{n,p}(\omega) g_n(\omega,\phi,\theta), \qquad (2)$$

where $W_{n,p}(\omega) = \sum_{l=0}^{L-1} w_{np,l} e^{-j\omega l}$ is the Discrete-Time Fourier Transform (DTFT) representation of $\mathbf{w}_{n,p}$, and $g_n(\omega,\phi,\theta)$ is the sensor response of the n-th microphone to a plane wave with frequency ω traveling in the direction (ϕ,θ). Variables ϕ and θ denote

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 609465.

[1] In the context of this work, HRTFs only model the direct propagation path between a source and a microphone mounted on a humanoid robot's head, but no reverberation components.

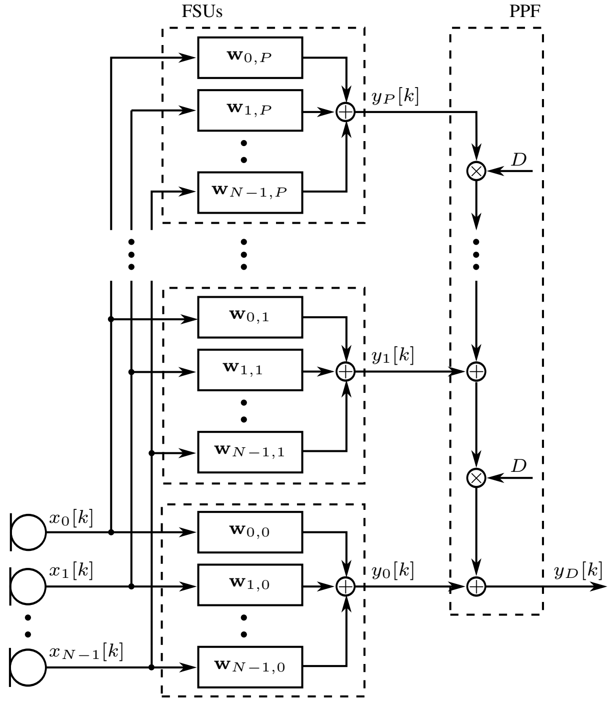


Fig. 1. Illustration of a polynomial filter-and-sum beamformer after [6].

azimuth and elevation angle, and are measured with respect to the positive x-axis and the positive z-axis, respectively, as in [9].

2.2. HRTF-based robust least-squares frequency-invariant polynomial beamforming

The main goal of the proposed HRTF-based RLSFIP beamformer design is to jointly approximate I desired beamformer responses $\hat{B}_{D_i}(\omega,\phi,\theta)$, each with a different Prototype Look Direction (PLD) (ϕ_i,θ_i), $i = 0,\dots,I-1$, by the actual beamformer response $B_{D_i}(\omega,\phi,\theta)$, where $D_i = (\phi_i - 90)/90$, in the Least-Squares (LS) sense. Hence, D_i lies in the interval $-1 \leq D_i \leq 1$, where, for example, $D = 0$ and $D = -1$ steer the main beam towards $\phi = 90°$ and $\phi = 0°$, respectively. For values of D which do not correspond to one of the PLDs, the PPF will interpolate between them, as expressed in (1). In this work, we apply polynomial beamforming only in the horizontal dimension. Thus, D_i only depends on the azimuth angle, whereas θ_i is constant for all PLDs. The extension to two-dimensional beam steering is an aspect of future work. In addition to the LS approximation, a distortionless response constraint and a constraint on the White Noise Gain (WNG) is imposed on each of the I PLDs. The approximation is carried out for a discrete set of Q frequencies ω_q, $q \in \{0,\dots,Q-1\}$ and M look directions (ϕ_m,θ_m), $m \in \{0,\dots,M-1\}$ (where, in this work, θ_m remains fixed) in order to obtain a numerical solution. Hence, the optimization problem of the HRTF-based RLSFIP beamformer design can be expressed as:

$$\underset{\mathbf{w}_{\mathrm{f}}(\omega_q)}{\operatorname{argmin}} \sum_{i=0}^{I-1} \|\mathbf{G}(\omega_q)\mathbf{D}_i\mathbf{w}_{\mathrm{f}}(\omega_q) - \hat{\mathbf{b}}_i\|_2^2, \tag{3}$$

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

$$\frac{|\mathbf{a}_i^{\mathrm{T}}(\omega_q)\mathbf{D}_i\mathbf{w}_{\mathrm{f}}(\omega_q)|^2}{\|\mathbf{D}_i\mathbf{w}_{\mathrm{f}}(\omega_q)\|_2^2} \geq \gamma > 0, \quad \mathbf{a}_i^{\mathrm{T}}(\omega_q)\mathbf{D}_i\mathbf{w}_{\mathrm{f}}(\omega_q) = 1, \quad \forall i = 0,\dots,I-1. \tag{4}$$

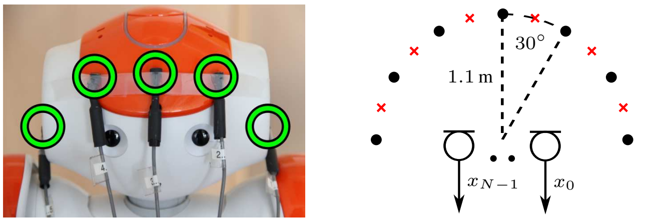
where $\hat{\mathbf{b}}_i = [\hat{B}_{D_i}(\phi_0,\theta_0),\dots,\hat{B}_{D_i}(\phi_{M-1},\theta_{M-1})]^{\mathrm{T}}$ is a vector of dimension $M \times 1$ containing the i-th desired response for all M angles, matrix $[\mathbf{G}(\omega_q)]_{mn} = g_n(\omega_q,\phi_m,\theta_m)$, vector $\mathbf{a}_i(\omega_q) = [g_0(\omega_q,\phi_i,\theta_i),\dots,g_{N-1}(\omega_q,\phi_i,\theta_i)]^{\mathrm{T}}$ is the steering vector which contains the sensor responses for the i-th PLD (ϕ_i,θ_i), and vector $\mathbf{w}_{\mathrm{f}}(\omega_q) = [W_{0,0}(\omega_q),\dots,\ W_{N-1,P}(\omega_q)]^{\mathrm{T}}$ of dimension $N(P+1) \times 1$ contains all filter coefficients. Furthermore, $\mathbf{D}_i = \mathbf{I}_N \otimes [D_i^0,\dots,D_i^P]$ is an $N \times N(P+1)$ matrix, where $\mathbf{I}_N$ is an $N \times N$ identity matrix and $\otimes$ denotes the Kronecker product. Operator $\|\cdot\|_2$ denotes the Euclidean norm of a vector. The optimization problem in (3), (4) can be interpreted as follows: Equation (3) describes the LS approximation of the I desired responses $\hat{B}_{D_i}(\omega_q,\phi_m,\theta_m)$ by the actual beamformer response. The first part of (4) represents the WNG constraint which is imposed on each of the I PLDs. γ is the lower bound on the WNG and has to be defined by the user. Hence, the user has the possibility to directly control the beamformer's robustness against small random errors like sensor mismatch or position errors of the microphones. The second part of (4) ensures a distortionless beamformer response for each of the I PLDs.

As in [5], we include measured HRTFs in (3) and (4) instead of the free-field-based steering vectors (which are only based on the microphone positions and the look directions). By doing this, the beamformer design can account for the influence of the humanoid robot's head on the sound field which would not be the case if we used free-field-based steering vectors as in [6]. The sensor responses are given as $g_n(\omega_q,\phi_m,\theta_m) = h_{mn}(\omega_q)$, where $h_{mn}(\omega_q)$ is the HRTF modeling the propagation between the m-th source position and the n-th microphone, mounted at the humanoid robot's head, at frequency ω_q. Matrix $\mathbf{G}(\omega_q)$ consists of all HRTFs between the M look directions and the N microphones, and $\mathbf{a}_i(\omega_q)$ contains the HRTFs corresponding to the i-th PLD.

The optimization problem has to be solved for each frequency ω_q separately. We use the same desired response for all frequencies for the design of the polynomial beamformer, which is indicated by the frequency-independent entries of $\hat{\mathbf{b}}_i$ [4, 5, 6]. The optimization problem in (3), (4) is formulated as a convex optimization problem [6] and we use CVX, a package for specifying and solving convex programs in Matlab [10], to solve it. After the optimum filter weights at each frequency ω_q have been found, FIR filters of length L are obtained by FIR approximation, see, e.g., [11], of the optimum filter weights using the fir2 method provided by Matlab [12].

3. EVALUATION

In the following, we evaluate the proposed HRTF-based RLSFIP beamformer design and compare it to the HRTF-based RLSFI beamformer design proposed in [5]. At first, the experimental setup is introduced. Then, the two beamformer designs are compared with respect to their approximation errors of the desired beamformer response. Eventually, the signal enhancement performance is evaluated in terms of Word Error Rates (WERs) of an Automatic Speech Recognition (ASR) system.



(a) Microphone positions.

(b) Source positions.

Fig. 2. Illustration of the employed microphone positions (green circles) at the humanoid robot's head and the source positions of the two-speaker scenario.

3.1. Setup and parameters

The evaluated beamformers were designed for the five-microphone robot head array in Fig. 2(a), using a WNG constraint of $\gamma_{\mathrm{dB}} = -20\mathrm{dB}$ and a filter length of $L = 1024$. For the design of the polynomial beamformer, we used $I = 5$ PLDs $\phi_i \in \{0°, 45°, 90°, 135°, 180°\}$ and a PPF of order $P = 4$. The set of HRTFs which is required for the HRTF-based beamformer design was measured in a low-reverberation chamber ($T_{60} \approx 50\mathrm{ms}$) using maximum-length sequences, see, e.g., [13, 14]. The HRTFs were measured for the same five-microphone array shown in Fig. 2(a) for a robot-loudspeaker distance of 1.1m. The loudspeaker was at an elevation angle of $\theta = 56.4°$ with respect to the robot. We chose this setup to simulate a taller human interacting with the NAO robot which is of height 0.57 m. The measurements were carried out for the robot looking towards broadside $(\phi, \theta) = (90°, 90°)$.

3.2. Evaluation of HRTF-based polynomial beamformer design

In this section, we investigate how well the desired beamformer response $\hat{B}_{D_i}(\phi, \theta)$ is approximated by the beamformer response of either the HRTF-based RLSFI or the HRTF-based RLSFIP beamformer. Ideally, the polynomial beamformer should be as good as the RLSFI beamformer in the best case, because it approximates the latter, i.e., the performance of both beamformers should be similar when steered towards one of the I PLDs.

Fig. 3 shows the beampatterns of the HRTF-based RLSFI beamformer and of the HRTF-based RLSFIP beamformer in Figs 3(a) and 3(b), respectively, steered towards $(\phi_{\mathrm{ld}}, \theta_{\mathrm{ld}}) = (135°, 56.4°)$. The resulting WNG of both beamformer designs is shown in Fig. 3(c). Please note that the beampatterns were computed with HRTFs modeling the acoustic system. Thus, they effectively show the transfer function between source position and beamformer output. A comparison of the beampatterns of the HRTF- and free-field-based RLSFI beamformer can be found in [5], illustrating the effect of the humanoid robot's head as scatterer on the sound field. From Fig. 3 it can be seen that the beampatterns of both beamformers look almost identical. This is because the actual look direction of the beamformers is equal to one of the five PLDs of the polynomial beamformer design. One can also see that the WNG is successfully constrained for both beamformer designs across the entire frequency range (with some slight deviations due to the FIR approximation with finite filter length). Comparison of Figs 3(a) and 3(b) with Fig 3(c) reveals that the beamformer's main beam broadens when the WNG reaches its lower bound. This points to the tradeoff between robustness and spatial selectivity which the user can control via γ in (4).

In Fig. 4 the beampatterns of the HRTF-based RLSFI and RLSFIP beamformers are shown for the look direction $(\phi_{\mathrm{ld}}, \theta_{\mathrm{ld}}) = (110°, 56.4°)$, which lies roughly half-way between two PLDs and can be expected to exhibit a large deviation from the desired response. The two beampatterns now look different, which is due to the interpolation between the PLDs by the polynomial beamformer. While for the lower frequencies the two main beams still look similar, the main beam of the polynomial beamformer is degraded for higher frequencies. Moreover, it can be observed that the polynomial beamformer cannot maintain a distortionless response in the desired look direction across the entire frequency range. The mismatch between RLSFI and RLSFIP beamformer also becomes obvious when looking at the WNG in Fig. 4(c). The WNG of the RLSFIP beamformer is generally lower than that of the RLSFI beamformer. In addition, the polynomial beamformer also exhibits a stronger violation of the WNG constraint than the RLSFI beamformer for $f < 500\mathrm{Hz}$.

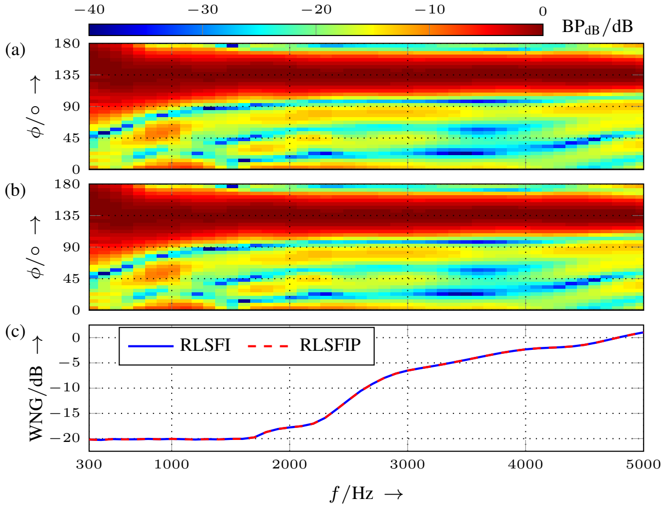


Fig. 3. Illustration of beampatterns of (a) the HRTF-based RLSFI beamformer and (b) the HRTF-based RLSFIP beamformer when the polynomial beamformer's look direction coincides with a PLD. The beamformers were designed for the five-microphone robot head array in Fig. 2(a) with look direction $(\phi_{\mathrm{ld}}, \theta_{\mathrm{ld}}) = (135°, 56.4°)$ and WNG constraint $\gamma_{\mathrm{dB}} = -20\,\mathrm{dB}$. The resulting WNG is illustrated in Subfigure (c).

In the following, we measure the approximation error of the desired response $\hat{B}_{D_{\mathrm{ld}}}(\phi, \theta)$ for a certain look direction ϕ_{ld} by the actual beamformer response $B_{D_{\mathrm{ld}}}(\omega, \phi, \theta)$, where $D_{\mathrm{ld}} = (\phi_{\mathrm{ld}} - 90)/90$, of either the RLSFI or RLSFIP beamformer by calculating the Mean Squared Error (MSE) which is defined as [8]:

$$\mathrm{MSE}(\phi_{\mathrm{ld}}) = \sum_{q=0}^{Q-1} \sum_{m=0}^{M-1} \frac{\left(|B_{D_{\mathrm{ld}}}(\omega_q, \phi_m, \theta_m)| - |\hat{B}_{D_{\mathrm{ld}}}(\phi_m, \theta_m)|\right)^2}{Q \cdot M}. \tag{5}$$

Fig. 5 depicts the MSE of the HRTF-based RLSFI and RLSFIP beamformer designs, calculated in steps of five degrees over the entire steering range $0° \leq \phi_{\mathrm{ld}} \leq 180°$. When steered to one of the five PLDs, i.e., when $\phi_{\mathrm{ld}} = \phi_i$, the RLSFIP beamformer design yields a similar MSE as the RLSFI beamformer design. In between those PLDs, the MSE of the polynomial beamformer design is usually larger than that of the RLSFI beamformer design. The lower MSE of the polynomial beamformer for $\phi_{\mathrm{ld}} \in \{5°, 175°\}$ may be explained by side lobes of the polynomial beamformer which are less pronounced at higher frequencies than those of the RLSFI beamformer for these two particular look directions.

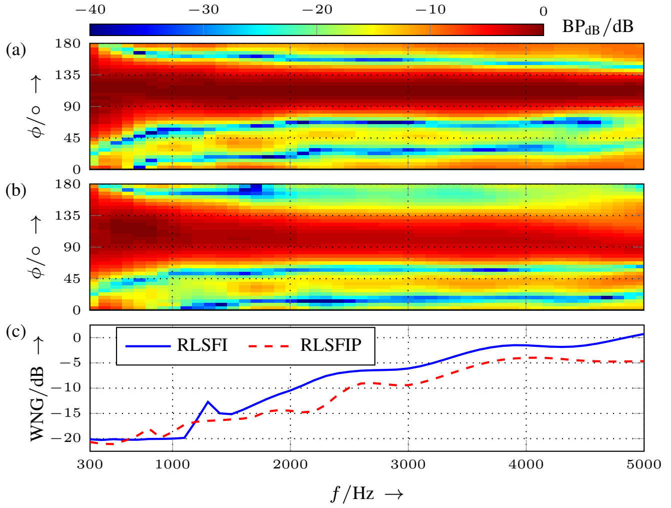


Fig. 4. Illustration of beampatterns of (a) the HRTF-based RLSFI beamformer and (b) the HRTF-based RLSFIP beamformer when the polynomial beamformer's look direction does not coincide with one of the PLDs. The beamformers were designed for the five-microphone robot head array in Fig. 2(a) with look direction $(\phi_{\text{ld}}, \theta_{\text{ld}}) = (110^\circ, 56.4^\circ)$ and WNG constraint $\gamma_{\text{dB}} = -20\,\text{dB}$. Subfigure (c) shows the resulting WNG.

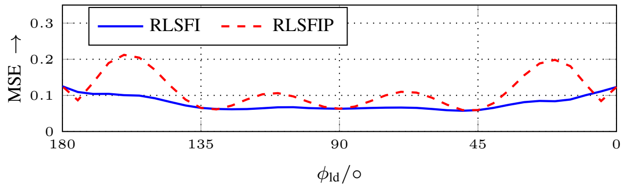


Fig. 5. Illustration of the MSE (5) of the HRTF-based RLSFI (blue curve) and HRTF-based RLSFIP (red curve) beamformer designs, calculated in steps of five degrees over the entire steering range.

3.3. Evaluation of signal enhancement performance

In this section, we evaluate the overall quality of the enhanced signals at the outputs of the HRTF-based RLSFI and RLSFIP beamformers. In addition, we also evaluate the original free-field-based RLSFIP beamformer [6] which assumes free-field propagation of sound waves and, therefore, cannot account for the influence of robot's head on the sound field. To this end, we use 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 [15] with a Hidden Markov Model (HMM)-Gaussian Mixture Model (GMM)-based acoustic model which was trained on clean speech from the GRID corpus [16], using MFCC+Δ+ΔΔ 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 [17]. Our test signal contained 200 utterances. Note that since the ASR system was trained on clean speech, we implicitly measure the amount of target signal distortion and interferer suppression.

We evaluated the signal enhancement in a two-speaker scenario, where the target signal was located at positions between $\phi_{\text{ld}} = 0^\circ$ and $\phi_{\text{ld}} = 180^\circ$ in steps of 30°. The Direction Of Arrival (DOA) of the target signal was assumed to be known for the experiments, i.e., no localization algorithm was applied. An investigation of the HRTF-based beamformer's robustness against DOA estimation errors can be found in [18]. For each target position, seven interfering speaker positions between $\phi_{\text{int}} = 15^\circ$ and $\phi_{\text{int}} = 165^\circ$ in steps of 30° were evaluated. An overview over all source positions is given in Fig. 2(b), where target and interfering sources are represented by black circles and red crosses, respectively. We created the microphone signals by convolving clean speech signals with Room Impulse Responses (RIRs) which we measured in a lab room with a reverberation time of $T_{60} \approx$ 190ms and a critical distance [19] of approximately 1.2m. The RIRs were measured with the same configuration as was used for the HRTF measurements described above. The WERs were calculated for each combination of target and interfering source position and averaged over the WERs obtained for the different interferer positions. The resulting average target source position-specific WERs are depicted in Fig. 6. The obtained WERs show that both HRTF-based beamformers significantly improve the speech recognition accuracy of the input signal. Moreover, they also outperform the free-field-based RLSFIP beamformer significantly, which emphasizes the necessity to incorporate the effect of the robot's head on the sound field into the beamformer design. It is interesting to see that the HRTF-based RLSFIP beamformer performs as well as the HRTF-based RLSFI beamformer whenever the target source is located in one of the PLDs which were used for designing the polynomial beamformer. When this is not the case, only a slightly higher average WER is obtained. This confirms that the polynomial interpolation of the HRTF-based RLSFI beamformer design works reasonably well such that it can be used in a robot audition scenario.

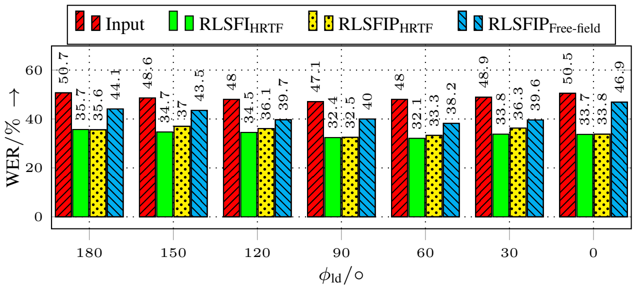


Fig. 6. Illustration of average target source position-specific WERs in %, obtained at the input (red bars) and at the output of the HRTF-based RLSFI (green bars), HRTF-based RLSFIP (yellow bars), and free-field-based RLSFIP (cyan bars) beamformers.

4. CONCLUSION

In this work, we proposed an HRTF-based RLSFIP beamformer design which allows for a flexible steering of a previously proposed robust HRTF-based RLSFI beamformer. We evaluated both beamformer designs with respect to their corresponding approximation error of the desired beamformer response and with respect to their signal enhancement performance which was evaluated by means of WERs of an ASR system. The results showed that the polynomial beamformer design provides a good approximation of the RLSFI beamformer design and, therefore, can be used successfully in a robot audition scenario instead of the computationally much more complex RLSFI beamformer design. Future work includes an investigation of the proposed HRTF-based polynomial beamformer design for more irregular sensor arrangements as well as an evaluation with a state-of-the-art Deep Neural Network (DNN)-based ASR system. An extension of the RLSFIP beamformer design to allow for a flexible steering of the main beam in two dimensions is another aspect of future work.

5. REFERENCES

[1] M.S. Pedersen, U. Kjems, K.B. Rasmussen, and L.K. Hansen, "Semi-blind source separation using head-related transfer functions," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, May 2004, pp. V–713–716.

[2] M. Maazaoui, K. Abed-Meraim, and Y. Grenier, "Blind source separation for robot audition using fixed HRTF beamforming," *EURASIP J. Advances Signal Proc. (JASP)*, vol. 2012, pp. 58, 2012.

[3] M. Maazaoui, Y. Grenier, and K. Abed-Meraim, "From binaural to multimicrophone blind source separation using fixed beamforming with HRTFs," in *Proc. IEEE Int. Conf. Systems, Signals, Image Process. (IWSSIP)*, Apr. 2012, pp. 480–483.

[4] E. Mabande, A. Schad, and W. Kellermann, "Design of robust superdirective beamformers as a convex optimization problem," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, Apr. 2009, pp. 77–80.

[5] H. Barfuss, C. Huemmer, G. Lamani, A. Schwarz, and W. Kellermann, "HRTF-based robust least-squares frequency-invariant beamforming," in *IEEE Workshop Applicat. Signal Process. Audio Acoustics (WASPAA)*, Oct. 2015, pp. 1–5.

[6] E. Mabande and W. Kellermann, "Design of robust polynomial beamformers as a convex optimization problem," in *Proc. IEEE Int. Workshop Acoustic Echo, Noise Control (IWAENC)*, Aug. 2010, pp. 1–4.

[7] M. Kajala and M. Hamalainen, "Filter-and-sum beamformer with adjustable filter characteristics," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, May 2001, pp. 2917–2920.

[8] E. Mabande, *Robust Time-Invariant Broadband Beamforming as a Convex Optimization Problem*, Ph.D. thesis, Friedrich-Alexander University Erlangen-Nürnberg, Apr. 2014, https://opus4.kobv.de/opus4-fau/frontdoor/index/index/year/2015/docId/6138.

[9] H.L. Van Trees, *Detection, Estimation, and Modulation Theory, Optimum Array Processing*, Detection, Estimation, and Modulation Theory. Wiley, 2004.

[10] Inc. CVX Research, "CVX: Matlab software for disciplined convex programming, version 2.1," June 2015, http://cvxr.com/cvx.

[11] A.V. Oppenheim, R.W. Schafer, and J.R. Buck, *Discrete-time Signal Processing (2nd Ed.)*, Prentice-Hall, Inc., 1999.

[12] MathWorks, "Fir2: Frequency sampling-based FIR filter design," Apr. 2016.

[13] M. R. Schroeder, "Integrated-impulse method measuring sound decay without using impulses," *J. Acoust. Soc. Am. (JASA)*, vol. 66, no. 2, 1979.

[14] M. Holters, T. Corbach, and U. Zoelzer, "Impulse response measurement techniques and their applicability in the real world," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Sept. 2009, pp. 1–5.

[15] D. Huggins-Daines, M. Kumar, A. Chan, A.W. Black, M. Ravishankar, and A.I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, May 2006, pp. I–183–188.

[16] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am. (JASA)*, vol. 120, no. 5, pp. 2421–2424, Nov. 2006.

[17] H. Christensen, J. Barker, N. Ma, and P.D. Green, "The CHiME corpus: A resource and a challenge for computational hearing in multisource environments.," in *Proc. INTERSPEECH*, Sept. 2010, pp. 1918–1921.

[18] H. Barfuss and W. Kellermann, "On the impact of localization errors on HRTF-based robust least-squares beamforming," in *Jahrestagung für Akustik (DAGA)*, Aachen, Germany, Mar. 2016, pp. 1072–1075.

[19] H. Kuttruff, *Room Acoustics, Fourth Edition*, E-Libro. Taylor & Francis, 2000.