Concept and Perceptual Validation of Listener-Position Adaptive Superdirective Crosstalk Cancellation Using a Linear Loudspeaker Array

Concept and Perceptual Validation of Listener-Position Adaptive Superdirective Crosstalk Cancelation Using a Linear Loudspeaker Array

Xiaohui Ma[1]*, Christoph Hohnerlein[2], Jens Ahrens[3]†

[1]Dynaudio A/S, Sverigesvej 15, 8660 Skanderborg, Denmark
[2]Berlin Institute of Technology, Ernst-Reuter-Platz 7, 10587 Berlin, Germany
[3]Chalmers University of Technology, 412 96 Gothenburg, Sweden

May 17, 2024

Abstract

This paper presents a multiband approach for crosstalk cancelation (CTC) based on superdirective near-field beamforming (SDB) that adapts dynamically to a change in the listener position. SDB requires the computation of a separate set of beamformer weights for each listener position. Our beamformer uses weights that exhibit a smooth evolution for listening positions along a linear trajectory parallel to the array. The beamformer weights can therefore be parameterized by using only a few parameters for each frequency. Upon real-time execution, the beamformer weights are determined efficiently for any position from the parameters with negligible error. Simulations and measurements show that the proposed method provides high channel separation and is robust with respect to small uncertainties of the listener position. A user study with 20 subjects and binaural signals shows consistent auditory localization accuracy across the different tested listening positions that is comparable to the localization accuracy of headphone rendering. The study also confirms the previously informal observation that fewer front-back confusions are be observed when the listeners face away from the loudspeaker array compared to the listeners facing towards the array.

*Also at Department of Engineering, Aarhus University, Finlandsgade 22, 8200 Aarhus N, Denmark.
†To whom correspondence should be addressed, Email: jens.ahrens@chalmers.se

1 INTRODUCTION

Binaural audio reproduction provides immerse 3D audio to users when signals encoded with head-related transfer functions (HRTFs) are delivered to the two ears of users independently. Headphone rendering is commonly used due to its natural channel separation between the two ears, and is widely applied in virtual reality and augmented reality applications. Despite the fact that the implementation is straightforward, headphone-based reproduction causes problems like social isolation and head internalization of sound [1], which makes loudspeaker rendering a good alternative. One also speaks of *transaural* [2] reproduction in this case. A fundamental challenge for transaural audio is to eliminate the crosstalk between the two ears [3], ideally, the signal intended for the ipsilateral ear will not be received by the contralateral ear.

Extensive transaural studies use a two-loudspeaker setup, and crosstalk cancelation (CTC) is achieved by system inversion [1, 4], which can easily break down in the presence of even small deviations from the assumptions. The obtained inverse filters are also subject to large errors around ill-conditioned frequencies [5], therefore strong coloration on the reproduced signals exists. Many studies have been focusing on improving the system robustness, among others, using frequency-dependent regularization [1], optimizing the loudspeaker positions [6], and involving the feedback control strategy [7]. Alternatively, recursive ambiophonic crosstalk elimination (RACE) proposed by Glasgal [8] provides simple means of CTC for two loudspeaker symmetric setup that is surprisingly robust with respect to head movement. However, the performance can strongly depend on loudspeaker position and even loudspeaker model.

Though the robustness with respect to the aforementioned limitations can be improved, the listeners are constrained in a narrow sweet spot, and CTC can easily break down outside the optimized position. It is therefore favourable to achieve CTC for a larger area. Recent research using multiple (more than two) loudspeakers has shown promising results, which falls into two general categories, i.e. inverse filter based methods and beamforming based methods. Based on the inverse filter design, Bauck [9] uses a multi-way loudspeaker array system to enlarge the sweet spot; Takeuchi and Nelson [5] propose the optimal source distribution, which greatly improves the robustness in terms of room reflections and misalignment of the system; Bai *et al.* [10] optimizes the array configuration and control points in the illuminated and shadow zones around the head. Alternatively, Hohnerlein and Ahrens [11] employ least-squares frequency-invariant beamforming to achieve CTC that is robust with respect to small head movements. In all cases, the listener is constrained to a fixed listening position, allowing for small deviations, up to $\sim \pm 5$ cm off the optimizing position.

In order to account for moving listeners, the CTC filters need to be adaptive and updated according to the listeners' instantaneous position, which needs to be measured with a suitable tracking system. Cecchi *et al.* [12] propose an extension of RACE applied to asymmetric two loudspeaker setups, where the delays and attenuation for each channel are updated based on the real-time position. Gálvez *et al.* [13] implement dynamic CTC by combining a fixed CTC filter with a delay-and-sum beamformer that steers towards the listener. However, the listener can only move along a circular trajectory, and the achievable crosstalk cancelation is limited. As to our awareness, no perceptual evaluation of the approach is available.

This paper presents a hybrid-full-band adaptive CTC based on a linear loudspeaker array. We employ the adaptive superdirective beamformer from [11] to achieve high amounts crosstalk cancelation in the frequency range from 1 kHz to 8 kHz, and a modified RACE algorithm in the frequency range from 250 Hz to 1 kHz. We showed in [14] that the beamformer weights can be parameterized without considerable loss of accuracy. The present paper complements the work with a perceptual validation of the approach.

This paper is organized as follows. In Sec. 2, a short introduction to the least-squares frequency-invariant beamforming is given, followed by the a description of RACE and the proposed dynamic crosstalk cancelation. In Sec. 3, the simulation setup and results are given. The experimental work and results are presented in Sec. 4, which is followed by a user study in Sec. 5 and discussion of the results in Sec. 6. The paper is concluded in Sec. 7.

2 METHOD

2.1 Least-squares frequency-invariant beamforming

For a linear loudspeaker array with N equispaced drivers, the directional response of a filter-and-sum beamformer at frequency ω is [15]

$$B(\omega, \vec{r}) = \sum_{n=0}^{N-1} W_n(\omega) \frac{1}{r_n} \mathrm{e}^{-\mathrm{j}\omega \frac{r_n}{c}}, \quad r_n = ||\vec{r} - \vec{x}_n||, \tag{1}$$

where $\vec{x}_n$ is the position of the n-th driver; $\vec{r}$ denotes the prescribed position for the directional response $B(\omega, \vec{r})$; $W_n(\omega)$ is the frequency response of the beamformer filter for the n-th driver, and c is the sound speed in air.

Least-squares (LS) beamforming approximates a target response $\hat{B}(\omega, \vec{r})$ by $B(\omega, \vec{r})$ in the LS sense. If the target directional response is frequency independent, i.e. $\hat{B}(\omega, \vec{r}) = \hat{B}(\vec{r})$, the beamformer is named least-squares frequency-invariant beamforming (LSFIB) [16].

Combining all prescribed $\vec{r}_m$ $(m = 1, \cdots, M)$, the directional response in Eq. (1) is reformed as

$$\boldsymbol{b}(\omega) = \boldsymbol{G}(\omega)\boldsymbol{w}_f(\omega), \tag{2}$$

where $[\boldsymbol{G}(\omega)]_{mn} = \mathrm{e}^{-\mathrm{j}\omega \frac{r_{mn}}{c}}/r_{mn}$, $r_{mn} = ||\vec{r}_m - \vec{x}_n||$, and $[\boldsymbol{w}_f(\omega)]_n = W_n(\omega)$.

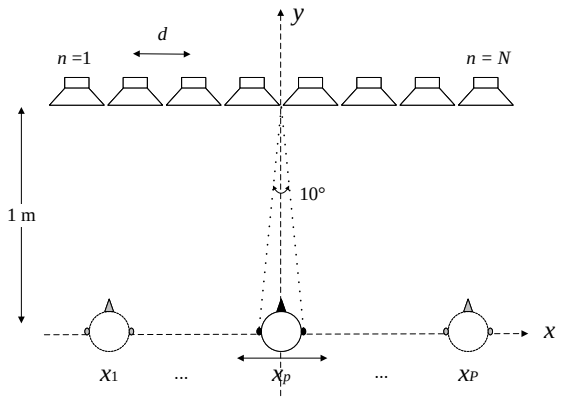

Figure 1: System geometry. The listener moves along a straight line 1 m from the array. The angle between left and right ear to the array center is 10° and assumed to be constant along the moving trajectory. The central listening position is at the coordinate origin.

The filter responses $\boldsymbol{w}_f(\omega)$ are determined by minimizing the squared error between the predicted and desired directional responses.

$$\min_{\boldsymbol{w}_f(\omega)} ||\boldsymbol{G}(\omega)\boldsymbol{w}_f(\omega) - \hat{\boldsymbol{b}}||_2^2, \tag{3}$$

where $\hat{\boldsymbol{b}} = [\hat{B}(\vec{r}_1), \cdots, \hat{B}(\vec{r}_M)]^T$.

CTC is achieved by setting the beamformer's mainlobe in the direction of the illuminated ear, in the meanwhile minimizing the sound energy in the shadowed ear direction by adding an energy constraint in the shadowed ear direction [11],

$$||\boldsymbol{G}_s(\omega)\boldsymbol{w}_f(\omega)||_2^2 \leq a, \tag{4}$$

where $\boldsymbol{G}_s(\omega)$ is a subset of $\boldsymbol{G}(\omega)$ containing the directions around the exact shadowed ear to allow for a smooth pressure transition; a determines the energy attenuation. The separation between the illuminated and shadowed ears with respect to the array center is set to be 10° at a distance of 1 m. This optimization problem can be solved using the CVX toolbox [17].

2.2 Adaptive crosstalk cancelation

To obtain CTC for moving listeners, the beamformer needs to be updated in real-time according to the listener's position. We assume for convenience in the remainder of this paper that the listener moves along a straight line parallel to the linear loudspeaker array in use as depicted in Fig. 1.

As LSFIB is computationally expensive, we propose adaptive CTC that employs an off-line modelling phase together with an on-line updating phase. The algorithm is illustrated in Fig. 2. In the off-line modelling phase, the designated contour of possible listener positions is discretized into P positions with a resolution of 5 mm. For each frequency ω_k, $k = 1, \cdots, K$, the beamformer weights are calculated for all positions $W_n(\omega_k, x_p)$, $p = 1, \cdots, P$. A typical evolution of the

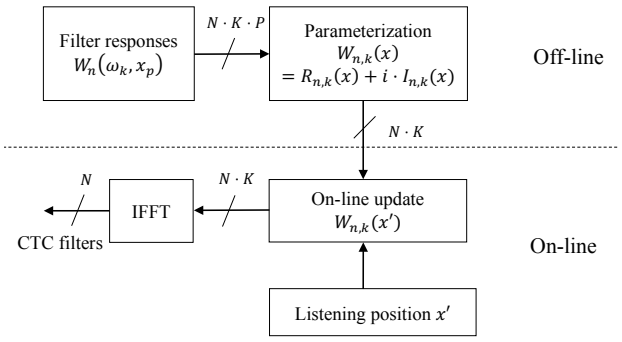

Figure 2: Signal flow of the proposed approach.

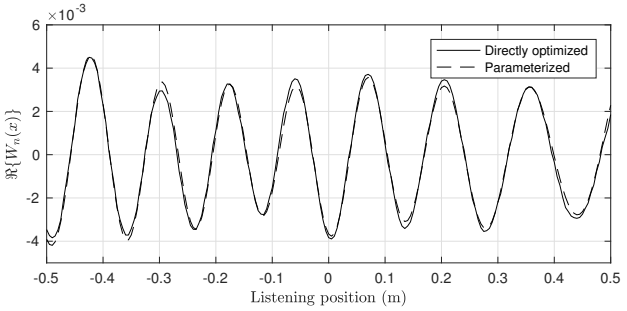

Figure 3: Sine parameterization of the beamformer weights. The real part of the beamformer weights of loudspeaker 1 at 5.7 kHz is shown by the solid curve; the parameterized beamformer weights are given by the dashed curve.

beamformer weights as a function of listening position is depicted in Fig. 3. As a smooth and periodic evolution is observed, a sum of sine functions is employed to parameterize the beamformer weights. We perform this separately for the real and imaginary parts as functions of the listener position

$$\begin{aligned} R(\omega_k, x) &= \Re\{W_n(\omega_k, x)\} \\ &= \sum_{q=1}^{Q} A_q(\omega_k) \sin[B_q(\omega_k)x + C_q(\omega_k)]. \end{aligned} \tag{5}$$

$$\begin{aligned} I(\omega_k, x) &= \Im\{W_n(\omega_k, x)\} \\ &= \sum_{q=1}^{Q} D_q(\omega_k) \sin[E_q(\omega_k)x + F_q(\omega_k)]. \end{aligned} \tag{6}$$

The coefficient set $\{A_q(\omega_k), \cdots, F_q(\omega_k)\}$, $q = 1, \cdots, Q$, is then used to calculate the actual CTC filters in real-time. In total there are $6Q \cdot K \cdot N$ coefficients need to be stored.

2.3 Recursive Ambiophonic Crosstalk Elimination

RACE is designed for symmetric two-loudspeaker setups. The crosstalk path from one loudspeaker channel

to the contralateral ear is estimated as a delayed attenuator, and is subsequently cancelled out by an attenuated, delayed and inverted signal from the other channel. The cancelation signal, in turn, causes crosstalk to the original ipsilateral ear, which should be further compensated, thus leading to a recursive process.

RACE is implemented as an IIR filter with two parameters: time delay τ and attenuation α. Fig. 4 shows the filter structure of RACE. To make RACE work for moving listeners, a new pair of loudspeakers is chosen when there is an update on the listening position; these two loudspeakers are symmetric with respect to the listener. In this paper, RACE is applied for the frequency range from 250 Hz to 1 kHz.

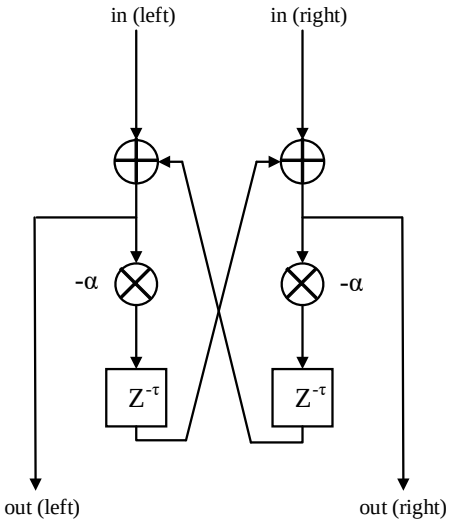

Figure 4: RACE filter structure.

3 SIMULATIONS

The proposed adaptive CTC is simulated by a linear equispaced loudspeaker array with eight drivers as depicted in Fig. 1, the spacing between two adjacent drivers is 15.2 cm. The distance between the mid point of the two out-most loudspeakers is therefore 1.06 m. Each loudspeaker is modelled by a point source. The listener moves along a straight trajectory 1 m from the array. The separation angle between the ears with respect to the array center is assumed to be constant 10° at all positions along the trajectory.

The applicable frequency bandwidth is investigated for the central listening position (the origin of the coordinate system). The main lobe of the beamformer is steered towards the left ear, while a null is steered towards the right ear. To incorporate the physical uncertainties in reality, e.g. mismatch between the loudspeakers, variations in the loudspeaker placement, Gaussian noise

$$\Delta A \sim \mathcal{N}(0, 0.3) \quad \text{and} \quad \Delta\Phi \sim \mathcal{N}(0, 0.001\omega/\mathrm{c}), \quad (7)$$

is added to the beamformer weights at each frequency for each loudspeaker, i.e.

$$W_n(\omega_n, x) = (|W_n| + \Delta A_n)\mathrm{e}^{\mathrm{j}(\angle W_n + \Delta\Phi_n)}. \quad (8)$$

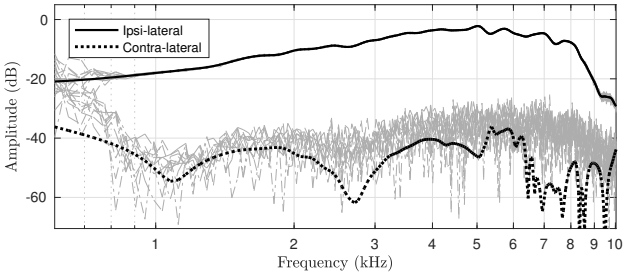

Figure 5: Frequency responses at the user's ears. Thick black lines represent ideal results; thin lines give the results for 10 simulations with random noise applied to the beamformer weights.

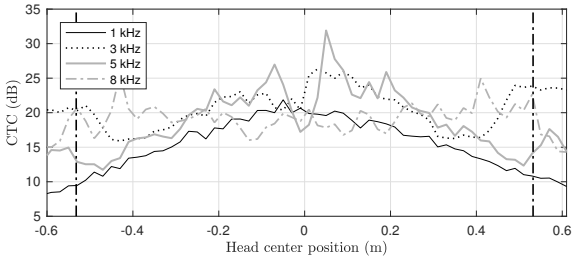

Figure 6: Crosstalk cancelation at different listening positions 1 m from the loudspeaker array. Results are obtained by averaging 20 random positions around the exact ear position; the head is modelled as a rigid sphere. Four frequencies within the applicable frequency range are shown. The vertical dash-dotted lines indicate the usable array aperture.

Simulated transfer functions from the array input to the ears of the user are presented in Fig. 5. Though thick lines show a wide applicable frequency band for the ideal setup, the mismatch reduces the usable frequency band to from 900 Hz to 8 kHz, where the channel separation is larger than 15 dB, which constitutes the lower boundary for binaural audio systems [18].

The achievable performance of the position-adaptive CTC using LSFIB is simulated at discrete positions along the listening trajectory. The listener's head is modelled as a rigid sphere with a radius of 9 cm to take the scattering into consideration. Fig. 6 shows the results for the obtained CTC. At each listening position, the simulation averages the sound pressure level for 20 random positions around the exact ear position with distances up to 4 cm. It is noticeable the listener should be located at $|x_P| \leq 0.35$ m, so that a channel separation of more than 15 dB is maintained over the entire beamforming frequency range.

Above presented results hold for the case that the beamformer weights $W_n(\omega_k, x)$ were obtained directly from the optimization. Fig. 3 depicts exemplarily the real part of $W_n(\omega_k, x)$ for a given loudspeaker at a given frequency as a function of the listener location. The deviation of the curves sum-of-sines parameterization represented by Eq. (5) and Eq. (6) from the optimal data is small and is similar across loudspeakers and across different frequencies.

The performance of the proposed parameterized

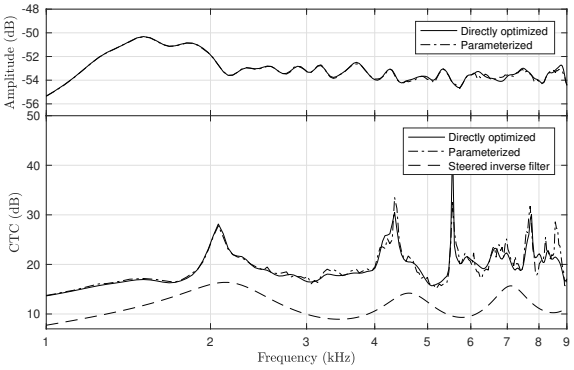

Figure 7: Frequency response at the ipsilateral ear (top) and crosstalk cancelation (bottom) at the listener position at 0.3 m from the center, solid lines are results from direct optimization dash-dot lines are results from parameterization. As a comparison, CTC obtained from the steered inverse filter from [13] is shown by the dashed line.

adaptive CTC is compared to CTC through direct optimization in Fig. 7, which shows the frequency response and obtained CTC for the listening position at 30 cm from the center. It can be observed that the proposed CTC gives approximately the same performance as the directly optimized CTC. Similar behaviors are found for other listening positions.

Fig. 7 also depicts a comparison of the performance of the proposed CTC to the approach presented in [13]. The obtained CTC for the depicted listening position at 30 cm from the center is ∼8dB in average higher for the proposed approach. This can be attributed to the fact that the presented approach applies superdirective beamforming contrary to [13]. Further simulations show that the approach from [13] has an asymmetric CTC performance about the y-axis, i.e., CTC is significantly higher if the ipsilateral ear is facing the array center. Our parameterized superdirective beamformer achieves a more symmetric CTC without increasing the computational cost significantly compare to the approach from [13].

System robustness with respect to accuracy of the positioning of the listener's head is investigated based on the listening position at 30 cm from the center. The head is modelled as a rigid sphere with radius of 9 cm; the two ears are on the ends the diameter parallel to the array. CTC for 20 random positions with deviations conforming to $\mathcal{N}(0, 0.015)$ around the exact ear position are depicted in Fig. 8. It can be seen that the loss in channel suppression is moderate so that we can conclude that the proposed CTC is robust in terms of inaccuracies in the head-tracking system that is employed.

Always assuming an ear separation angle of 10° might not be a good estimate when the head approaches the array ends. When analysis CTC at random positions around the exact ear position, some ear position pairs might match the 10° separation angle better than the exact ones and therefore give better channel separation. This can explain the observation that CTC are sometimes higher for the random positions than that for the assumed ear positions.

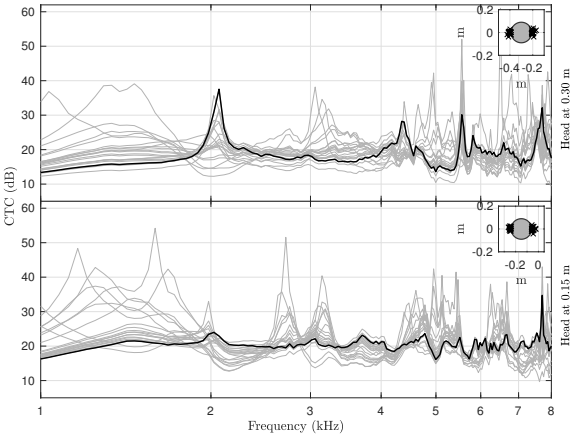

Figure 8: Crosstalk cancelation at random ear positions up to 4 cm away from the exact ear positions. The thick black line shows the results at the ears; the thin gray lines show the results at 20 random position. The inset figure shows the head and the random evaluation position. Top: Listening position 30 cm from the central one; bottom: 15 cm from the central one.

4 INSTRUMENTAL VALIDATION

Performance of the proposed adaptive CTC was evaluated in an anechoic chamber. Eight Genelec 8020 studio monitors were closely arranged in a linear array, driven by an Antelope Orion 32 audio interface. As in Sec. 3, the loudspeaker spacing is 15.2 cm. Sound pressure was measured by a KEMAR manikin placed 1 m in front of the array, at the array center and 30 cm left to the array center, as depicted in Fig. 9.

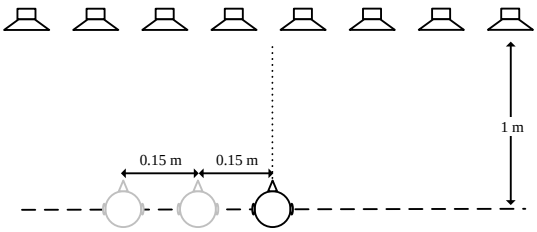

Figure 9: Measurement/test positions. For transfer function measurements using KEMAR, two positions 1 m from the loudspeaker array are measured: array center and 30 cm left to the array center. For user study, the subject is tested at three positions: array center, 15 cm off center, and 30 cm off center. At each position, two look directions are tested: facing the array and away from the array.

Since the proposed beamforming CTC works for the frequency band from 1 kHz to 8 kHz, to obtain a full band system, a hybrid/multiband approach is employed:

- f <250 Hz: Single sub-woofer
- 250 Hz< f <1 kHz: RACE
- 1 kHz< f <8 kHz: Beamforming
- f >8 kHz: Stereo through the two out-most loudspeakers to maximize natural head shadowing

A fourth-order four-band crossover filter with cut-off frequencies of 250 Hz, 1 kHz, and 8 kHz is therefore designed to perform the bandsplitting.

At the listening position 30 cm left to the array center, when the left ear is illuminated, the measured transfer functions from the sound source to the ear canals are given in Fig. 10(a). In the beamforming range, a general channel separation larger than 15 dB is obtained, and 10 dB in the RACE range. Amplitude dips between 1–2 kHz are observed, which could be due to small reflections from the chair on which KEMAR was sitting. Fig. 10(b) shows the transfer functions when the right ear is illuminated, despite small variations, the responses present comparable channel separation as the case when the left ear is illuminated. Similar observations are also found when KEMAR is at the array center, and Fig. 10(c) shows the result when the left ear is illuminated.

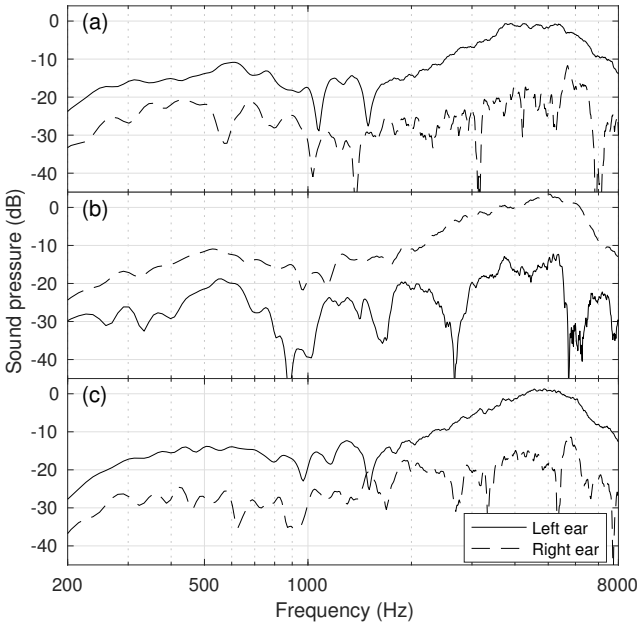

Figure 10: Measured transfer functions at the ear positions. (a,b) depict results when KEMAR is positioned at a distance of 1 m to the array and 30 cm left to the array center. (a) left ear is the illuminated ear; (b) right ear is the illuminated ear. (c) depicts the transfer function when KEMAR is at the array center, and the left ear is the illuminated one.

In the frequency range from 1 kHz to 8 kHz, the measured channel separation is slightly lower compared to the simulations which predicts around 20 dB. The overall gains of the loudspeakers were calibrated so that the uncertainty is the loudspeaker directivity, which might depart from the free-field point source model that is used by the beamformer in Eq. (1). Exemplary data are depicted in Fig. 11. We found that the loudspeaker directivity indeed changes considerably as a function of the listener position and, more importantly, that the amplitude changes differently over distance than that of a spherical wave. This aspect requires a closer look in future work as it provides significant potential for reducing the discrepancy between the simulated and the measured CTC.

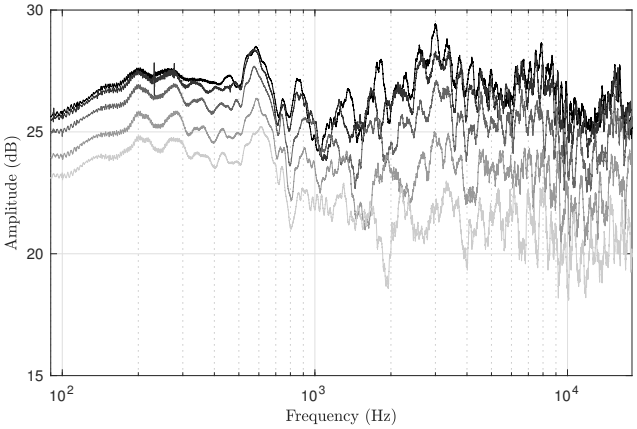

Figure 11: Measured transfer functions of one of the array loudspeaker to an omnidirectional microphone at a distance of 1 m from the array for different displacements of the microphone parallel to the array in steps of 10 cm starting on-axis

It is also evident from Fig. 10 that the transfer functions to the ipsilateral ear are not flat and vary slightly between the listening positions. We applied a gentle global equalization to account for this. No considerable changes of the timbre were apparent between different listening positions.

5 PERCEPTUAL VALIDATION

5.1 Setup

To assess the performance of the proposed dynamic CTC, a localization user study involving 20 subjects with self-reported unimpaired hearing was conducted. The experiment was conducted in a lab room with very little reverberation. A Polhemus PATRIOT$^{\text{TM}}$ headtracker was used to obtain the user's real-time position, see Fig. 12, the sensor was attached to a hair band and fixed on the subject's head. The signal processing was performed by a modified version of the SoundScapeRenderer [19]. The subject was positioned inside a booth of size 2 m $\times$ 2 m formed of acoustically transparent cloth to prevent the subjects from identifying the locations of the loudspeakers. Five dummy loudspeakers were arranged around the booth to prevent the subjects from making assumptions about the limitations of the setup.

Three listening positions 1 m from the array were tested: array center; 15 cm off center, and 30 cm off center, as seen in Fig. 9. At each position, two look directions were investigated: facing towards the array

and away from the array. The study was executed and processed on an Apple iMac Pro and Max/MSP.

Figure 12: Experimental setup for the user study. A Polhemus PATRIOT^{TM} head-tracker is used to yield the real-time positions. Five dummy loudspeaker are arranged around the booth. The loudspeaker array is located behind and above the computer screen.

The experiment paradigm was identical to the one in [11]. The subjects were asked to identify the directions of a virtual sources oscillating around the angles [0°, ±15°, ±35°, ±60°, ±90°, ±120°, ±145°, ±165°, 180°] by selecting the direction segment in the graphical user interface depicted in Fig. 13. The graphical interface was always positioned on the screen such that it appeared straight in the front of the subject to prevent paralax effects.

Each of the 16 potential virtual source locations was presented twice for each combination of the 3 listening positions and the 2 listener orientations resulting in 6 x 32 responses per subject. The order of the source positions are well as the listening position and orientation were randomized.

The subjects were discourage of performing rotations of the head as these are not tracked, and the current system cannot account for them.

The virtual source locations were achieved by imposing HRTFs onto the same infinite rock music loop that was used in [11]. To ensure the best localization performance, each user was subject to a HRTF selection session before the test, where 16 HRTFs set from the repository [20] were presented. The subjects first selected four sets of HRTFs out of 16, and then went through a A/B comparison test to find the best HRTF, whereby the same rock music loop was used like in the actual experiment. The virtual source emitting this signal was continuously rotating around the head in the horizontal plane; the selection criteria were the same as in [11]:

- Constant planar height on ear level
- Constant perceived distance
- Constant loudness
- No change in timbre
- Smooth movement

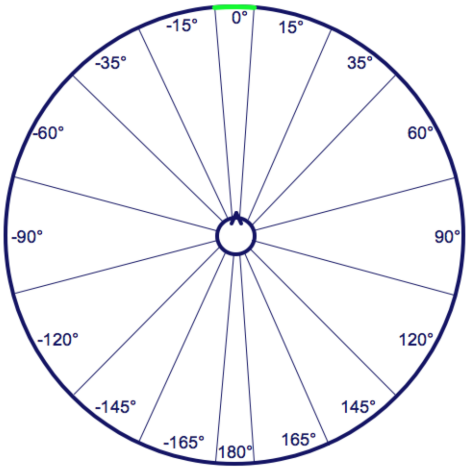

Figure 13: GUI of the localization experiment, the green mark at the top indicates the direction the listener is currently hovering over.

Summing up, the experiment procedure was composed of written instructions, HRTF selection, training of the subject on the experimental task, and finally the experiment. The session was wrapped up with an informal interview of the subject regarding his/her experience during the experiment. The duration of the entire session varied between 30 and 45 min.

5.2 Results

The localization test results are presented in the density plots in Fig. 14. Bubbles represent the answer distribution, and bubble size indicates the answer frequency. Fig. 14(a–c) show the results for the case when the subjects were facing the array at the central position, 15 cm off-center, and 30 cm off-center positions, respectively. It can be observed that the answer distribution patterns are similar for the three positions, and the identified sound sources have a tendency to be within [-90°, 90°], i.e. the front half plane; hardly any virtual sources was localized from the back.

This is contrary to the case where the subjects were facing away from the array. The localization results for this situations for the three listening positions are given in Fig. 14(d–f). A clear diagonal answer distribution is observed for all three positions indicating sounds were identified from both front and back and without a significant amount of front-back confusions. The orthogonal branches show the existence of front-back confusion, the amount of which is comparable to headphone rendering, the data for which is shown in Fig. 15. The depicted data are from the experiment

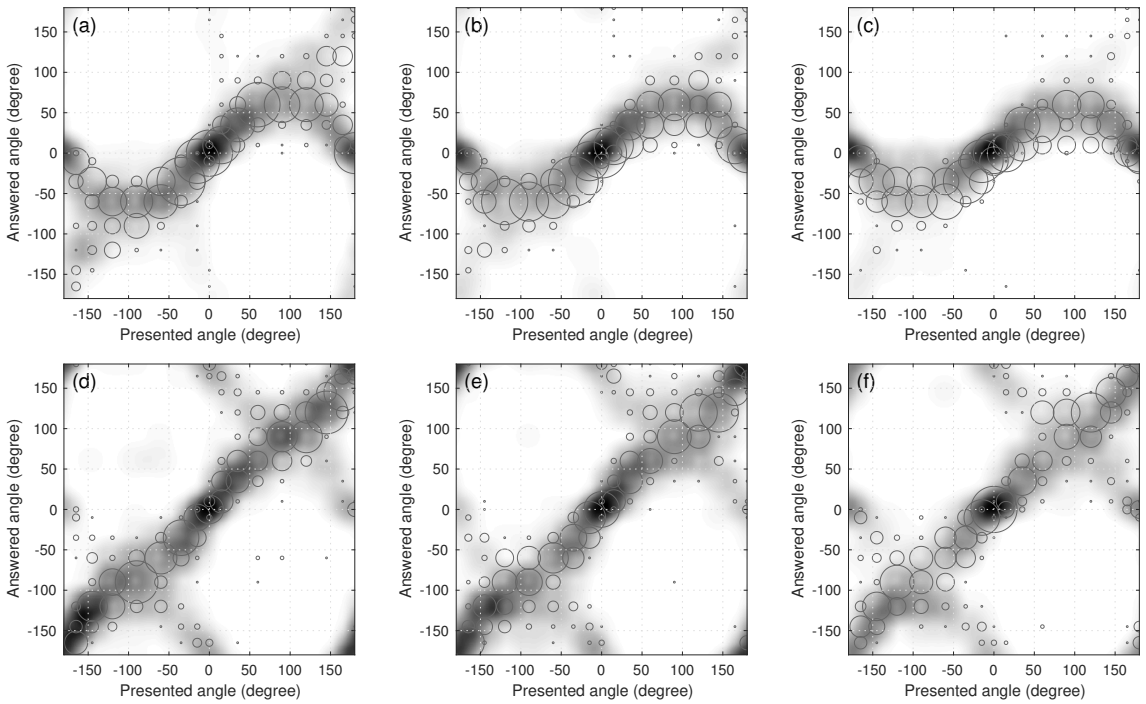

Figure 14: Localization experiments. The answers are presented in the density plot, bubble size corresponds to answer frequency. (a, b, c) give the results when the listener facing the array; (d, e, f) give the results when the listener facing away from the array. (a, d) gives the results when the listener in the array center; (b, e) are 15 cm off center; (c, f) are 30 cm off center.

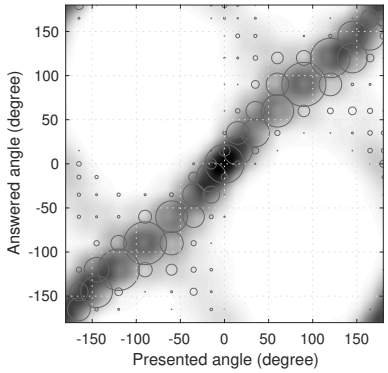

Figure 15: Localization experiment using headphone.

presented in [11], which used the exact same HRTF sets, signals, and experimental procedure.

Since no substantial differences are observed for the three listening positions, it can be concluded that the proposed adaptive CTC shows stable performance with respect to listener movements. The observations related to front-back confusions are discussed in Sec. 6.2.

To investigate the localization accuracy, the localization errors were calculated. Due to the similarities of the results at positions of the same look direction, only results for the edge position, 30 cm from array center, are given. Fig. 16(a) depicts the localization error of the headphone reference data from Fig. 15; Fig. 16(b) depicts the results of the array with the subjects facing towards the array; Fig. 16(c) depicts the results of the array with the subjects facing away from the array. It can be observed that when facing away from the array, the localization errors present a similar distribution as the errors in the headphone data: large errors are found when the virtual sources approach the media plane. For the case when subjects are facing towards the array, small errors are found when the virtual sources locating in the front half-plane; large errors are found when the virtual sources are in the back half-plane because of the front-back confusions.

6 DISCUSSION

6.1 General

The measured channel separation that is achieved by the prototype array is in the range of 10 and 20 dB for the evaluated listening positions depending on the frequency. The simulated channel separation is significantly higher than the measured one, which suggests some departure of the properties of the hardware from the assumptions in the beamformer design. A deviation of the loudspeaker directivity from the employed spherical wave model is apparent the effect of which is subject to future research.

The achieved auditory localization accuracy is comparable to the accuracy achieved with headphone playback where the channel separation may be assumed to be perfect. Localization at the out-most tested listening position is somewhat squashed towards the median plane. In other words, lateralization is slightly reduced.

As common with CTC, our subjects reported that externalization was good. They also reported that locatedness of the source was high, but the virtual source

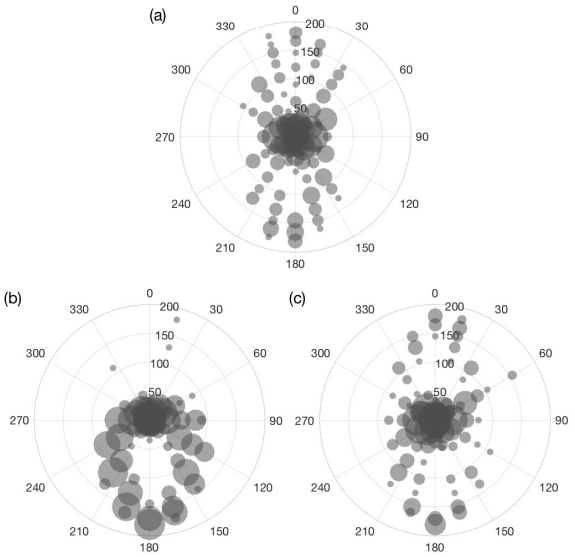

Figure 16: Absolute localization error using headphone (a) and loudspeaker array at position 30 cm from array center (b,c); bubble size corresponds to the error frequency. (a) using headphone (b) facing towards the array; (c) facing away from the array.

had a somewhat artificial quality. Some subjects even reported a minor spatial irritation immediately after the experiment such as affected balance. The likely cause for this is the considerable residual crosstalk. The channel separation seems high enough to achieve full lateralization of the virtual source but seems to be affecting the perception of timbre.

The effect of the reproduction room is unclear at this stage and requires further research. The channel separation is difficult to measure as a function of frequency in this case as it is not possible to segregate the direct sound of the loudspeakers from the room. The channel separation is definitely reduced by the floor reflection, which perceptually merges with the direct sound and which is not modeled in the beamformer design. The remaining room response may be interpreted as some sort of reverberation as it will arrive significantly later than the combination of direct sound and floor reflection. To what extent the reverberation may be assumed to further reduce the channel separation is unclear.

6.2 Front-Back Confusions

The observation that fewer front-back confusions occur when the listener faces away from the array compared to the listener facing the array has been reported based on informal observations in [21, Sec. 5.4.4]. Up to date, we have deployed the present approach with 4 different linear loudspeaker array prototypes that were all employing comparable parameters such as number of loudspeakers and loudspeaker spacing but were using different loudspeaker models (all of comparable size). We have made the observation that fewer front-back confusions occur when the subjects face away from the array with all our prototypes, although the magnitude of this effect varied with the prototype.

It has been reported by the author of [2] that the subjects' awareness of the locations of the loudspeakers can affect the front-back discrimination. The subjects' simply refuse to localize virtual sources at directions where there are no loudspeakers. The fact that we added dummy loudspeakers as well as non-see-through cloth to the experimental setup suggests that we can exclude the expectation of the subjects as influence. The subjects were simply not aware of neither the location nor the amount of loudspeakers. The causes for the named differences therefore have to be acoustic.

Assuming stationary conditions, the only difference between the two listener orientations is the slightly different filtering of the signals by the outer ear. Analysis of the employed HRTF sets suggests that the beams impinging from the rear experience an attenuation of a few dB in the range of, say, 2-6 kHz. In a separate experiment, we applied a similar attenuation to the signals while the listeners were facing the array so that the signal that arise at the listeners' ears are similar to those that arise with no such attenuation and the listener facing away from the array. We found that this attenuation did not affect the observed amount of front-back confusions.

Recall that we were only tracking translations of the listener but no head rotations. Small and possibly subconscious head rotations of the listener alter primarily the interaural time difference (ITD), assuming that the head rotations are so small that no considerable changes in the interaural level difference occur. This alteration is roughly consistent with the changes in ITD that occur with head rotations in natural hearing for those case where the virtual sound source is located in the same hemisphere like the loudspeaker array. I.e., when both the loudspeaker array and the virtual source are located in front of the listener, or equivalently when both the loudspeaker array and the virtual source are located behind the listener, then the ITD alterations due to small head movements should not cause considerable irritation. However, when the virtual source and the loudspeaker array are located on different sides, then head rotations will cause ITDs that represent virtual source locations that are different from the intended one so that the virtual source can appear to be moving. None of the two fundamental listener orientations therefore shows an advantage in these terms over the other one.

Summing up, the considerations presented above suggest than neither listener awareness nor head rotations are a likely cause for the difference in the frequency of front-back confusions. We have no final conclusion on the causes for the observation.

7 CONCLUSIONS

Adaptive crosstalk cancelation based on sine parameterization of the beamformer weights is proposed, which involves an off-line modelling phase and an on-

line updating phase. Sum-of-sines is used to model the beamformer weights as a function the listening position, which is assumed to be a straight line parallel to the array. Simulations were conducted for the frequency range from 1 kHz to 8 kHz, and show that the system is robust with respect to user movement (along the straight line) and outperforms non-superdirective solutions. A real-time rendering system was implemented involving a head-tracker to provide the position updates, and a crossover filter splitting the audio content into four frequency bands: content below 250 Hz played through a sub-woofer; content between 250 Hz and 1 kHz rendered through RACE; content between 1 kHz and 8 kHz rendered by beamforming; content above 8 kHz rendered through a stereo setup.

Measurements performed with a KEMAR manikin show that the channel separation is larger than 15 dB in the beamforming range and 10 dB in the RACE range. A user study with 20 subjects shows consistent localization performance at all tested listening positions with slight reduction of the lateralization for listing locations far away from the array center. This proofs the effectiveness of the proposed adaptive CTC method.

The test results also reveal that better localization is obtained when the array is located behind the listeners in the form of significantly fewer front-back confusions. The causes to this phenomenon are not completely clear and will be investigated in the future.

8 ACKNOWLEDGMENT

Innovation Fund Denmark is kindly acknowledged for co-funding X.M.s Industrial Ph.D. scholarship. We also thank Georgios Zachos for building the loudspeaker array.

References

[1] E. Choueiri, "Optimal Crosstalk Cancellation for Binaural Audio with Two Loudspeakers," presented at the *Princeton University* (2010).

[2] W. G. Gardner, "3-D Audio Using Loudspeakers," PhD thesis, MIT Media Lab (1997).

[3] B. B. Bauer, "Stereophonic Earphones and Binaural Loudspeakers," *J. Audio Eng. Soc*, vol. 9, no. 2, pp. 148–151 (1961).

[4] O. Kirkeby, P. A. Nelson, "Digital Filter Design for Inversion Problems in Sound Reproduction," *J. Audio Eng. Soc*, vol. 47, no. 7/8, pp. 583–595 (1999).

[5] T. Takeuchi, P. A. Nelson, "Optimal source distribution for binaural synthesis over loudspeakers," *J. Acoust. Soc. Am.*, vol. 112, no. 6, pp. 2786–2797 (2002).

[6] D. B. Ward, G. W. Elko, "Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation," *IEEE Signal Process. Lett.*, vol. 6, no. 5, pp. 106–108 (1999).

[7] T. Samejima, Y. Sasaki, I. Taniguchi, H. Kitajima, "Robust transaural sound reproduction system based on feedback control," *Acoustical Science and Technology*, vol. 31, no. 4, pp. 251–259 (2010), [Online]. Available: 10.1250/ast.31.251.

[8] R. Glasgal, "360 Localization via 4.x RACE Processing," presented at the *Audio Engineering Society Convention 123* (2007 Oct).

[9] J. Bauck, "A Simple Loudspeaker Array and Associated Crosstalk Canceler for Improved 3D Audio," *J. Audio Eng. Soc*, vol. 49, no. 1/2, pp. 3–13 (2001).

[10] M. R. Bai, C.-W. Tung, C.-C. Lee, "Optimal design of loudspeaker arrays for robust cross-talk cancellation using the Taguchi method and the genetic algorithm," *J. Acoust. Soc. Am.*, vol. 117, no. 5, pp. 2802–2813 (2005), [Online]. Available: 10.1121/1.1880852.

[11] C. Hohnerlein, J. Ahrens, "Perceptual evaluation of a multiband acoustic crosstalk canceler using a linear loudspeaker array," presented at the *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 96–100 (2017 March).

[12] S. Cecchi, A. Primavera, M. Virgulti, F. Bettarelli, J. Li, F. Piazza, "An efficient implementation of acoustic crosstalk cancellation for 3D audio rendering," presented at the *2014 IEEE China Summit International Conference on Signal and Information Processing (ChinaSIP)*, pp. 212–216 (2014 July), [Online]. Available: 10.1109/ChinaSIP.2014.6889234.

[13] M. F. Simón Gálvez, T. Takeuchi, F. M. Fazi, "Low-Complexity, Listener's Position-Adaptive Binaural Reproduction Over a Loudspeaker Array," *Acta Acust. united Ac.*, vol. 103, no. 5, pp. 847–857 (2017).

[14] X. Ma, C. Hohnerlein, J. Ahrens, "Listener-Position Adaptive Crosstalk Cancelation Using A Parameterized Superdirective Beamformer," presented at the *Proc. of IEEE SAM* (2018 July).

[15] H. L. V. Trees, *Optimum array processing: Part IV of detection, estimation, and modulation* (Wiley, New York) (2002).

[16] E. Mabande, M. Buerger, W. Kellermann, "Design of robust polynomial beamformers for symmetric arrays," presented at the *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–4 (2012 March).

[17] M. Grant, S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," Website (2014), `http://cvxr.com/cv`.

[18] Y. L. Parodi, P. Rubak, "A Subjective Evaluation of the Minimum Channel Separation for Reproducing Binaural Signals over Loudspeakers," *J. Audio Eng. Soc*, vol. 59, no. 7/8, pp. 487–497 (2011).

[19] M. Geier, S. Spors, J. Ahrens, "The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods," presented at the *124th Convention of the AES* (2008 May).

[20] F. Brinkmann, A. Lindau, S. Weinzierl, S. v. d. Par, M. Müller-Trapet, R. Opdam, M. Vorländer, "A High Resolution and Full-Spherical Head-Related Transfer Function Database for Different Head-Above-Torso Orientations," *J. Audio Eng. Soc*, vol. 65, no. 10, pp. 841–848 (2017).

[21] C. Hohnerlein, "Beamforming for Acoustic Crosstalk Cancellation," MSc thesis, University of Technology Berlin (2016).