PERCEPTUAL EVALUATION OF A MULTIBAND ACOUSTIC CROSSTALK CANCELER USING A LINEAR LOUDSPEAKER ARRAY

Christoph Hohnerlein[*]

Quality & Usability Lab
Berlin Institute of Technology
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
christoph.hohnerlein@qu.tu-berlin.de

Jens Ahrens[†]

Division of Applied Acoustics
Chalmers University of Technology
412 96 Gothenburg, Sweden
jens.ahrens@chalmers.se

ABSTRACT

We propose and evaluate an acoustic crosstalk canceler for binaural sound presentation based on an 8-channel linear loudspeaker array. The system uses traditional LSFI beamforming and path estimation in the mid and low frequency range, respectively. Simulations show a channel separation of around $40\,\mathrm{dB}$ over the vast part of the considered frequency range.

The channel separation in a real-world implementation was measured to be $10\,\mathrm{dB} - 25\,\mathrm{dB}$ over a broadband spectrum of $250\,\mathrm{Hz} - 9000\,\mathrm{Hz}$. When comparing localization precision of virtual binaural signals to ground-truth presentation with headphones, a 21 subject study revealed significantly higher localization precision ($F(1) = 23.56, p < 0.001$) when discounting front-back confusions, which in turn occurred significantly more often for most participants. Perceptually, externalization was rated significantly higher ($t(20) = 3.983, p > 0.001$); task difficulty, timbre and spaciousness where rated the same accross both presentation methods. Robustness of the method was high and no fixation of the subjects' heads was applied.

Index Terms— spatial sound, transaural sound, crosstalk cancellation, beamforming, RACE

1. INTRODUCTION

Spatial audio presentation using binaural synthesis requires tight control over the signals at the listener's ears, as any alteration of the received signals may change the perception of localization cues or timbre in unpredictable ways. Binaural audio is already readily available using traditional headphones, enabling the placement of sound sources in virtual 3D space around the user. Untethering this experience would allow binaural synthesis in a larger range of situations where the use of headphones might be unfavorable or impossible such as in social interactions, work settings or in traffic.

Two goals have to be met: Firstly, the separation between the channels reaching the left and the right ear have to be maximized. Secondly, the transmission to the intended ear should match the original content. Previously published approaches have the tendency to be sensitive in regards to inaccuracies in listener positioning or system control and are often only applicable over a small frequency band. We propose a multiband, array-based speaker system, where traditional Least-Squares-Frequency-Invariant (LSFI) beamforming and path estimation are simultaneously employed in their respective comfortable frequency ranges. Such a system was shown to be robust against inaccuracies in target position and reproduction setup in simulation, measurement and user testing.

[*]The presented work was partly carried out at CCRMA (Stanford University), supported by the DAAD PROMOS stipend as well as the Quality & Usability Lab (TU Berlin), supported by DFG Grant AH 269/2-1.

[†]Jens Ahrens was partly supported by the DAAD at the CCRMA (Stanford University).

2. PRIOR WORK

Already in 1961, Bauer [1] investigated recording and playback mismatch of stereo and binaural signals and 2 years later, compensation filters were used to present binaural signals to a listener inside an anechoic chamber by Schroeder [2]. But as he noted himself in [3], approaches based on the inversion of the crosstalk path break down easily outside a very small sweet spot or even due to a non-average head shape or even slight turning of the head. These filter inversion schemes have been refined by [4, 5, 6, 7, 8], and [9]. Gardner [10] designed and tested a real-time system that dynamically tracks the subject and adjusts the filter inversion accordingly.

Other approaches, such as optimization of speaker position was explored by [11, 12, 13, 14] and [15]. Virtual sources [16] and the beamforming [17], control the radiation pattern to optimally transmit binaural signals to the listener's ear, as do the specifically designed speaker systems of [18]. Several transaural approaches have been evaluated perceptually in [19, 20, 21].

Beamforming, just like the signal processing for spatially filtering, has a long history of application both for input (sensor) and output (speaker) arrays. The authors suggest [22] for an in-depth review as well as [23] and [24] for an even broader analysis of array signal processing.

3. PROPOSED METHOD

The beamforming based cross-talk cancellation approach presented here follows closely the work of Mabande et. al. [25], the extension at the low end is an implementation of Recursive Ambiophonic Crosstalk Elimination (RACE) by Glasgal [26].

3.1. Least Squares Frequency Invariant Beamforming (LSFI)

For an linear array with N equidistant sensors, the array response $\mathbf{b}(\omega)$ may be written in matrix notation as:

$$\mathbf{b}(\omega) = \mathbf{G}(\omega)\mathbf{w}_f(\omega)\,, \tag{1}$$

where $\mathbf{G}(\omega)$ is a matrix of $e^{-j\omega\tau_n(\vartheta)}$ over all propagation delays τ_n in the columns and all spatial angles ϑ in the rows and $\mathbf{w}_f(\omega)$

holds all N filters $W_n(\omega)$ in the Fourier domain. The array's steering vector $\mathbf{d}$ towards a target angle ϑ_{target} is defined as:

$$\mathbf{d}(\omega) = \begin{bmatrix} e^{-j\omega\tau_0(\vartheta_{target})} \\ \vdots \\ e^{-j\omega\tau_{N-1}(\vartheta_{target})} \end{bmatrix} \quad (2)$$

Least-Squares beamformers optimally approximate a desired response $\hat{\mathbf{b}}(\omega_p)$:

$$\hat{\mathbf{b}}(\omega_p) \stackrel{!}{=} \mathbf{G}(\omega_p)\mathbf{w}_f(\omega_p) \quad (3)$$

A set of filters $\mathbf{w}_f$ is to be derived by minimizing the second norm of the difference to the array response. Because this problem space is overdetermined for $M > N$ (number of angles larger than number of sensors), convex optimization can be used to solve:

$$\min_{\mathbf{w}_f(\omega_p)} ||\mathbf{G}(\omega_p)\mathbf{w}_f(\omega_p) - \hat{\mathbf{b}}(\omega_p)||_2^2 \quad (4)$$

3.2. Path estimation (RACE)

Recursive Ambiophonic Crosstalk Elimination (RACE) [26] is a heuristic approach for symmetric two-channel loudspeaker setups. The crosstalk at a given contralateral ear is actively canceled by a delayed and attenuated copy of the signal that caused the crosstalk. The cancellation signal has opposite sign and is emitted by the considered ear's ipsilateral loudspeaker.

The two parameters delay Δt and attenuation Δa account for the longer path to the contralateral ear and head shadowing, respectively. RACE is typically applied in the frequency range of 250 Hz ... 5000 Hz. Remarkably, frequency independent delay and attenuation seem to be sufficiently accurate, which makes the implementation of RACE straightforward. Of course, the now delayed and attenuated cancelation signal also needs to be cancelled, resulting in a recursive suppression scheme. RACE requires carefully tuned hardware and software parameters to optimize its performance.

4. IMPLEMENTATION

The core method to achieve the crosstalk cancelation is the LSFI beamforming. It turned out that its robustness for the chosen array parameters (8 speakers, 14.5 cm inter-speaker distance) dramatically decreases outside the frequency band of $f_{BF} \approx 1\,\text{kHz} \ldots 8\,\text{kHz}$, as shown in Sec. 5.1. RACE was used to extend the lower working frequency range ($f_{RACE} \approx 250\,\text{Hz} \ldots 1000\,\text{Hz}$).

4.1. Beamformer

To achieve crosstalk cancelation, the beamformer has to exhibit a main lobe in direction ϑ_{target} of the illuminated ear and ideally a zero in direction ϑ_{stop} of the shadowed ear. It was decided to not employ the distortionless constraint of $\mathbf{wd} \stackrel{!}{=} 1$ as this allows for harder constraints at ϑ_{stop} while the frequency response at ϑ_{target} can be equalized at a later stage. To increase robustness, the constraints for ϑ_{stop} are extended to neighboring angles (null width), leading to the following constraint:

$$||\mathbf{G}_{stop}\mathbf{w}||_2^2 <= 0.01 (\approx -40\,\text{dB}) \quad (5)$$

The desired response $\hat{\mathbf{b}}(\omega_p)$ was all 0 except a pulse at the target direction ϑ_{target} of the shape $[\ldots, 0, 0.2360, 0.4719, 0.7079, 0.9900, 1.0000, 0.9900, 0.7079, 0.4719, 0.2360, 0, \ldots]$. The optimization statement as implemented in Matlab using the CVX toolbox [27] is shown in Listing 1.

```
for f=1:P
  cvx_begin quiet
    variable wf(N) complex
    minimize(norm(G(:,:,f) * wf - b, 2))
    subject to
      norm( Gstop(:,:,f) * wf ) <= 0.01
  cvx_end
end
```

List. 1. Optimization statement to derive the weights $\mathbf{w}$ that minimize Eq. 4 under the null constraint of Eq. 5

The parameters $\vartheta_{target} = 6^\circ$ and $\vartheta_{stop} = -6^\circ$ with a null width of 9° were found favorable for a subject with a head diameter of 20 cm located broadside at a distance of 1 m form the array center. The desired FIR filter can then be simply obtained as the appropriately windowed inverse Fourier transforms of $\mathbf{w}$. Their length L can be adjusted by adjusting $\mathbf{G}$ accordingly, balancing precision versus optimization performance, and was set to 1024.

The array response at the reference radius 1 m shown in Fig. 1 clearly shows the low amplitude in the stop direction and high amplitude in the target direction.

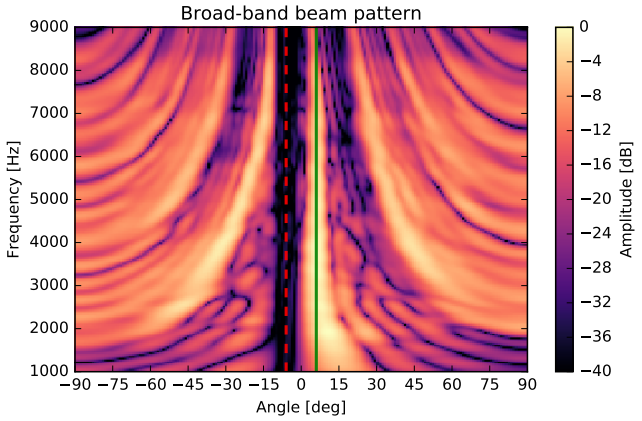


Fig. 1. Broadband beam-pattern of the sample 8-speaker array with 14.5 cm spacing. $\vartheta_{target} = 6^\circ$ (solid line), $\vartheta_{stop} = -6^\circ$ (dashed line), null width 9°.

4.2. RACE

The RACE structure was implemented in Max MSP using the *gen~* module and cosine interpolation between samples, as shown in Fig. 2. Optimal settings for parameters attenuation G and delay D were found manually based on the ear signals of a manikin.

4.3. Band splitting

4^{th} order Linkwitz-Riley crossover filters were used for their symmetrical radiation response and 0 dB amplitude on-axis, implemented using the Jamoma Toolbox [28]. A monophonic subwoofer is proposed for frequencies below $f_{RACE,min} = 250\,\text{Hz}$, as contributions to localization are marginal. No cancelation was applied in the frequency band above $f_{LSFI,max} = 8\,\text{kHz}$ and was played back by the two speaker at the ends of the array so as to maximize natural head shadowing.

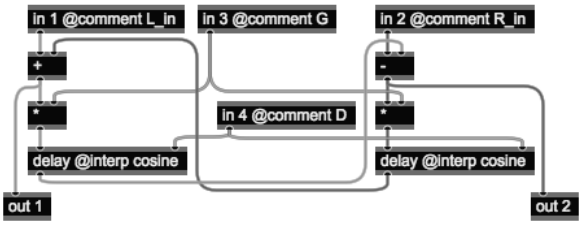


Fig. 2. Implementation of RACE filter structure.

5. EVALUATION

The performance of the array is evaluated by the robustness against inter-speaker equalization disturbance as well as against variance of the subject's position, which is first simulated and then confirmed by a user study.

5.1. Simulation

Calculating the sound pressure level around a rigid sphere placed at $1\,\mathrm{m}$ in front of the linear array allows to predict the resulting crosstalk suppression and robustness.

Fig. 3(a) shows the system's transfer function to the subject's ears in the presence of transducer mismatch, which was simulated by applying random gain and phase noise to the obtained speaker weights $\mathbf{w}$. In Fig. 3(b), the position of the ears around the rigid sphere was randomized, which corresponds to variance of the subject's physiology, position, and orientation. The comparison between both figures suggests that precise speaker placement and equalization have a larger detrimental effect and should therefore be prioritized over precise subject location.

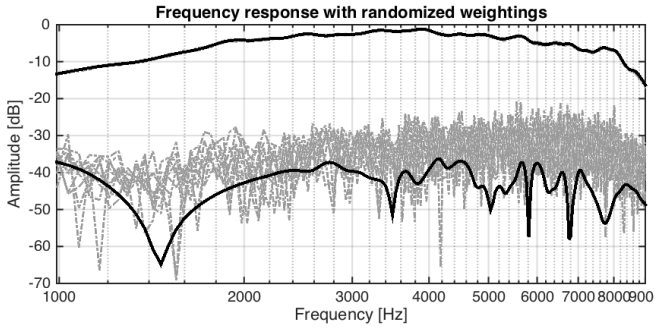
In summary, fig. 3 suggests robust channel separation of $30\,\mathrm{dB} \ldots 60\,\mathrm{dB}$, exceeding the necessary $15\,\mathrm{dB}$ to convey binaural signals according to [29]. A significant drop in crosstalk suppression can be observered outside of $f_{LSFI} = 1\,\mathrm{kHz} \ldots 8\,\mathrm{kHz}$.

5.2. Study setup

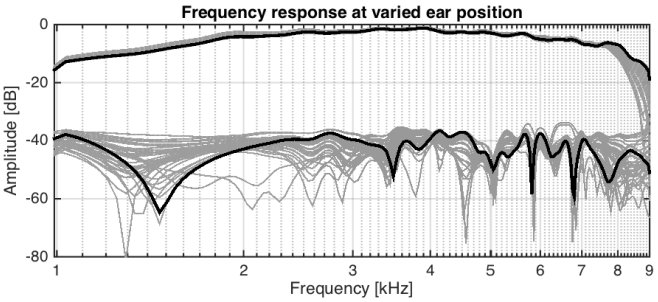
Eight *Fostex PM0.4* speakers were arranged in a line array configuration, driven by a *RME Fireface UCX* soundcard and equalized using a *Beyerdynamic MM-1* microphone. For the study, the array was placed behind an acoustically transparent curtain while three more speakers were positioned in the room as dummy sources to prevent subject bias with respect to possible virtual source location [10]. The reference condition was presented using equalized *Sennheiser HD-25* headphones. All interaction was done using a Max/MSP GUI in front of the subject.

The stimulus used was $15\,\mathrm{s}$ loop of a dry guitar and drum set. It was chosen for its natural sound, its decent coverage of the frequency range of interest, the mix of transients and tonal content as well as its unobtrusiveness, making it tolerable to listen to it on repeat for the full run time of the study.

A visual marker was mounted in front of the subject marking the 0° viewing direction, the subject's seat was fixed over a similar marker on the floor in a $2\,\mathrm{m} \times 2\,\mathrm{m}$ booth. No other sort of head restraint was applied, but the subjects were asked to position themselves according to the two markers. Rotations of the head were discouraged.



(a) Array transfer function with random gain and phase noise on the beamforming weights with zero mean and variances of $\sigma^2_{\text{gain}} = 0.3$ dB and $\sigma^2_{\text{phase}} = 0.001\frac{\omega}{c}$, respectively.



(b) Array transfer function at 40 randomly distributed points around the assumed ear positions at a distance of up to 5 cm

Fig. 3. Transfer functions around a rigid sphere at a distance of $1\,\mathrm{m}$ from the array center. The top curve ($0\,\mathrm{dB} \ldots -10\,\mathrm{dB}$) represents the ipsilateral ear, the bottom curve ($-40\,\mathrm{dB} \ldots -60\,\mathrm{dB}$) represents contralateral ear; bold black lines represent ideal conditions.

5.3. HRTF selection

To select a best-fit HRTF for each subject, a mix of the approaches presented in [30] and [31] was used. An initial set of 16 HRTFs from [32] were presented to the subjects, out of which the best four were compared in an A/B fashion. The participants were always able to seamlessly switch between all stimuli and were asked to grade an virtual source slowly moving around their head inside the horizontal plane according to the following criteria:

- Constant height on ear level
- Constant distance
- Constant loudness
- Equal and constant timbre (i.e. no coloration)
- Consistent movement around head

5.4. Localization task

Front-back ambiguity is a big issue when localizing source positions in the horizontal plane, as each position has a corresponding position with identical ITD and ILD cues in the other halfplane. Due to the missing dynamic cues of the static binaural synthesis, we instead move the sources tangentially around given positions in the horizontal plane, which has shown to help resolving the front-back ambiguity [33] and [34].

The subject's task was to identify the circular segment of a virtual source slowly oscillating around one of the angles $\vartheta_{test} \in [0^\circ, \pm 15^\circ, \pm 35^\circ, \pm 60^\circ, \pm 90^\circ, \pm 120^\circ, \pm 155^\circ, \pm 165^\circ, 180^\circ]$.

The binaural stimulus was presented in one session via headphones, in another session via the array, the order was fully randomized.

The results are presented in Fig. 4-6. It can be seen that localization accuracy is generally high with partly considerable deviations for fully lateralized virtual source positions. Front/back confusions occurred more frequently in array presentation. Remarkably, the array exhibits higher localization accuracy than headphone presentation when the confusions are corrected before the analysis.

The *best case* scenario reflects the data from only those 6 subjects that showed the lowest localization error.

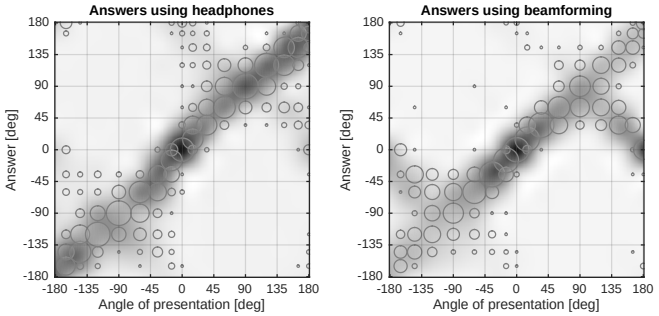


Fig. 4. Scatter plot of answers over density plot, bubble size corresponds to answer frequency.

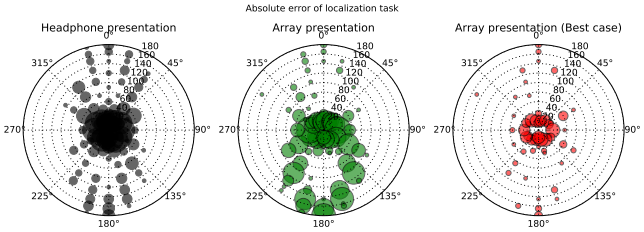


Fig. 5. Polar scatter plot of absolute error at each angle, bubble size corresponds to error frequency

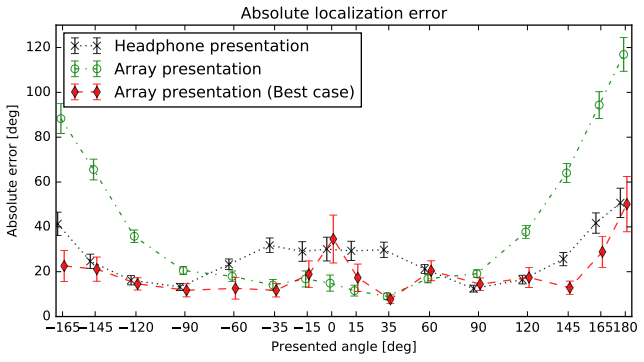


Fig. 6. Absolute localization error for headphones presentation (black), array presentation (red) and best-case array presentation (green), error bars mark 95% confidence interval.

5.5. Survey

Lastly, all subjects were asked to fill out a short survey on demographics (gender, age, technical background, knowledge on binaural technique, experience with listening to binaural material, understanding of the system before the experiment) and subjective rating on a 5-point scale with qualitative descriptions between conditions "WITH headphones" and "WITHOUT headphones" concerning the dimensions shown in table 1.

Differences in the perception between conditions were tested on significance using a paired-sample T-Test, where the null hypothesis is that the pairwise difference between the answers distribution has a mean equal to zero. All distributions pass the Shapiro-Wilk test of normality.

dimension	scale anchors	t	p
Difficulty	hard ⇔ easy	.513	.614
Timbre	not ⇔ very (natural)	−1	.329
Extern.	in ⇔ far outside (head)	**3.983**	**< .001**
Spaciousness	not ⇔ very (spatial)	.677	.506

Table 1. T-Test between the answer distributions for headphone and beamforming presentation. Bold dimensions are statistically significant with 95% confidence level, df = 20.

As shown in Table 1, **Externalization** was rated significantly ($p < 0.001$) higher for the array-based reproduction ($M = 4$, $SD = 0.7746$) compared to the headphones ($M = 2.857$, $SD = 1.014$). This might be explained by the additional externalization cues added by the room. To increase the comparability, an appropriate room impulse response could be added to the headphone condition. No other dimensions were rated significantly different.

6. CONCLUSIONS AND FUTURE WORK

Two crosstalk cancellation approaches were successfully used simultaneously to deliver virtual sound sources to a seated subject using binaural synthesis. An increase in localization precision was noted at the cost of more front-back confusions. Externalization was rated significantly higher.

[10] confirms a large reduction in front-back reduction when using dynamic binaural synthesis, which is an obvious extension to the present system - ideally this would be accomplished by optical tracking to leave the user completely untethered. Furthermore, this would enable real-time adjustment of the beamforming and RACE parameters, allowing for presenting binaural audio to a moving target. Additionally, it was noted that front-back ambiguity decreases substantially when placing the array behind the user. This is probably due to the additional localization cues of the room and should be investigated.

7. REFERENCES

[1] Benjamin B. Bauer, "Stereophonic earphones and binaural loudspeakers," *Journal of the Audio Engineering Society*, vol. 9, no. 2, pp. 148–151, 1961.

[2] M.R. Schroeder and B.S. Atal, "Computer simulation of sound transmission in rooms," *Proceedings of the IEEE*, vol. 51, no. 3, pp. 536–537, Mar. 1963.

[3] Manfred R. Schroeder, "Progress in Architectural Acoustics and Artificial Reverberation: Concert Hall Acoustics and Number Theory," *J. Audio Eng. Soc*, vol. 32, no. 4, pp. 194–203, 1984.

[4] P. Damaske, "Head-related two-channel stereophony with loudspeaker reproduction," *The Journal of the Acoustical Society of America*, vol. 50, no. 4B, pp. 1109–1115, 1971.

[5] T. Mori, G. Fujiki, N. Takahashi, and F. Maruyama, "Precision sound-image-localization technique utilizing multitrack tape masters," *J. Audio Eng. Soc*, vol. 27, no. 1/2, pp. 32–38, 1979.

[6] Duane H. Cooper and Jerald L. Bauck, "Prospects for transaural recording," *Journal of the Audio Engineering Society*, vol. 37, no. 1/2, pp. 3–19, Feb. 1989.

[7] Jerry Bauck and Duane H. Cooper, "Generalized transaural stereo and applications," *Journal of the Audio Engineering Society*, vol. 44, no. 9, pp. 683–705, 1996.

[8] Mark Poletti, "Improved virtual acoustics using a line array," in *XIX-th Biennial Conference of the New Zealand Acoustical Society*, 2008, pp. 1–6.

[9] Edgar Y. Choueiri, "Optimal crosstalk cancellation for binaural audio with two loudspeakers," *Princeton University*, p. 28, 2008.

[10] William G. Gardner, *3-D audio using loudspeakers*, Springer Science & Business Media, 1998.

[11] David Griesinger, "Equalization and spatial equalization of dummy-head recordings for loudspeaker reproduction," *J. Audio Eng. Soc*, vol. 37, no. 1/2, pp. 20–29, 1989.

[12] Darren B. Ward and G.W. Elko, "Optimum loudspeaker spacing for robust crosstalk cancellation," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1998, vol. 6, pp. 3541–3544 vol.6.

[13] Darren B. Ward and Gary W. Elko, "Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation," *Signal Processing Letters, IEEE*, vol. 6, no. 5, pp. 106–108, 1999.

[14] Jose J. Lopez and A. Gonzalez, "Experimental evaluation of cross-talk cancellation regarding loudspeakers' angle of listening," *IEEE Signal Processing Letters*, vol. 8, no. 1, pp. 13–15, Jan. 2001.

[15] Mingsian R. Bai, Chih-Wei Tung, and Chih-Chung Lee, "Optimal design of loudspeaker arrays for robust cross-talk cancellation using the taguchi method and the genetic algorithm," *The Journal of the Acoustical Society of America*, vol. 117, no. 5, pp. 2802, 2005.

[16] Daniel Menzel, Helmut Wittek, Günther Theile, Hugo Fastl, et al., "The binaural sky: A virtual headphone for binaural room synthesis," in *International Tonmeister Symposium*, 2005.

[17] Markus Guldenschuh and Alois Sontacchi, "Transaural stereo in a beamforming approach," in *Proc. DAFx*, 2009, vol. 9, pp. 1–6.

[18] Matthew S. Polk, "Sda ^{TM}surround technology white paper," *Polk Audio, Nov*, 2005.

[19] Jean-Marc Jot, Véronique Larcher, and Olivier Warusfel, "Digital signal processing issues in the context of binaural and transaural stereophony," in *Audio Engineering Society Convention 98*. Audio Engineering Society, 1995.

[20] Markus Guldenschuh, Chris Shaw, and Alois Sontacchi, "Evaluation of a transaural beamformer," in *27th International Congress of the Aeronautical Sciences*, Nice, Sept. 2010.

[21] Gunar Schlenstedt, Fabian Brinkmann, Sönke Pelzer, and Stefan Weinzierl, "Perzeptive evaluation transauraler binauralsynthese unter berücksichtigung des wiedergaberaums," in *Fortschritte der Akustik – DAGA 2016*, Aachen, Germany, March 2016, pp. 561–564.

[22] B.D. Van Veen and K.M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[23] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, July 1996.

[24] Harry L Van Trees, *Optimum array processing: part IV of detection, estimation, and modulation*, Wiley, New York, 2002.

[25] E. Mabande, A. Schad, and W. Kellermann, "Design of robust superdirective beamformers as a convex optimization problem," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 77–80.

[26] Ralph Glasgal, "360 localization via 4. x race processing," in *Audio Engineering Society Convention 123*. 2007, Audio Engineering Society.

[27] Michael Grant and Stephen Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," `http://cvxr.com/cvx`, Mar. 2014.

[28] Tim Place and Trond Lossius, "Jamoma: A modular standard for structuring patches in max," in *Proceedings of the International Computer Music Conference*, 2006, pp. 143–146.

[29] Yesenia Lacouture Parodi and Per Rubak, "A subjective evaluation of the minimum channel separation for reproducing binaural signals over loudspeakers," *Journal of the Audio Engineering Society*, vol. 59, no. 7/8, pp. 487–497, 2011.

[30] Bernhard U. Seeber and Hugo Fastl, "Subjective selection of nonindividual head-related transfer functions," in *In Proceedings of the 2003 International Conference on Auditory Display*, 2003, pp. 1–4.

[31] Yukio Iwaya, "Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears," *Acoustical science and technology*, vol. 27, no. 6, pp. 340–343, 2006.

[32] Fabian Brinkmann, Alexander Lindau, Stefan Weinzierl, Gunnar Geissler, and Steven van de Par, "A high resolution head-related transfer function database including different orientations of head above the torso," in *Proceedings of the AIA-DAGA 2013 Conference on Acoustics*, 2013.

[33] Frederic L. Wightman and Doris J. Kistler, "Resolution of front–back ambiguity in spatial hearing by listener and source movement," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2841–2853, 1999.

[34] Arne Nykänen, Axel Zedigh, and Peter Mohlin, "Effects on localization performance from moving the sources in binaural reproductions," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*. 2013, vol. 247, pp. 4023–4031, Institute of Noise Control Engineering.